Machine Learning for Medical Decision Support and Individualized Treatment Assignment

by

Finn C. Kuusisto

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2015

Date of final oral examination: 08/14/2015

The dissertation is approved by the following members of the Final Oral Committee:

Jude Shavlik, Professor, Computer Sciences
David Page, Professor, Biostatistics and Medical Informatics
Elizabeth Burnside, Professor, Radiology
Charles Dyer, Professor, Computer Sciences
Vitor Santos Costa, Associate Professor, Computer Science (University of Porto)

To Maggie, Larry, and Elina.

ACKNOWLEDGMENTS

I am indebted to many people for their support through my graduate school career.

First, I want to thank my advisors Jude Shavlik and David Page. Their motivation, guidance, and constructive feedback has constantly pushed me to improve my research and writing. I cannot imagine where I would be without their invaluable experience and advice.

I would also like to recognize my committee members: Elizabeth Burnside, Vítor Santos Costa, and Charles Dyer. Beth has been a welcoming and patient mentor as I have ventured into clinical applications, Vítor has been a great source of inspiration and technical expertise in our collaborations, and Charles has provided me with insightful comments and perspective. I owe them all an enormous debt of gratitude.

I would like to acknowledge NLM grant R01LM010921, NIH grants R01CA165229 and R01LM011028, and NIGMS grant R01GM097618 for funding and supporting my research, and I would like to acknowledge the University of Wisconsin Computer Sciences department for funding and training as a teaching assistant in the early years of my graduate career.

I wish to thank my friends and colleagues: Eric Lantz, Aubrey Barnard, Alex Cobian, Kendrick Boyd, Jeremy Weiss, Jie Liu, Brandon Smith, Houssam Nassif, Inês Dutra, Yirong Wu, Sarah Edlund, and many others. There are countless reasons I want to thank all of these people, but I want to thank Eric and Aubrey for the many late nights and conversations that made our office a fun and stimulating environment. I want to thank Alex for being a guide through hard times, and for always hosting game night. I want to thank Kendrick, Jeremy, and Jie as wonderful collaborators, and travel partners. I want to thank Brandon for sharing my sense of humor, and for weekend brewing. I want to thank Houssam, Inês, and Yirong for inspiration, guidance, and heroic efforts working with me on the hospital

datasets. Lastly, I want to thank Sarah for her constant encouragement and optimism.

Finally, I would like to thank my family for their endless love and support. To Maggie, Larry, and Elina, thank you for always encouraging me to do what I love, trusting me to make the right decisions, and being there for me through the good times and the bad. I cannot express how grateful I am to have you in my life.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

ABSTRACT

World Health Organization estimates of health care expenditure reveal a global trend of increasing costs, and health care systems need to become more efficient at treating patients to slow this trend. Incentives are in place to develop information-based health care systems, and I claim that using machine learning tools in medicine will lead to improvements in patient care. My work demonstrates new methods to improve collaboration between machine learning experts and clinicians, and new methods for modeling individual responses to treatment.

My work in collaboration with clinical experts involves the adaptation of machine learning models to address the challenging task of identifying benign breast cancer biopsies that cannot be definitively diagnosed. I first adapt an inductive logic programming learner to prefer rules that do not misclassify malignant cases, and show promising results that both adhere to the clinical objective and provide insight into the task. I later present a framework for collaboration between clinical and machine learning experts, leveraging clinician expertise to build and refine a model that meets the conservative objective of missing no malignant cases.

My work on estimating individual responses to treatment takes lessons from the marketing domain, applying uplift modeling to two primary medical tasks. One task is to identify patients at greater risk of heart attack due to treatment with COX-2 inhibitors, and another is understanding characteristics of in situ breast cancer specific to older women. I first present a statistical relational learner that constructs Bayesian networks to maximize area under the uplift curve (AUU), and show that the learned networks capture clinically-relevant characteristics of indolent, in situ breast cancer. I next present a support vector machine for maximizing AUU and show promising results on both the COX-2 inhibitor and breast cancer tasks, as well as a synthetic marketing task. Finally, I present a collabora-

tion showing strong evidence that machine learning for individualized treatment effect estimation improves upon current methods in multiple ways. Overall, I present multiple works that demonstrate improved clinical collaboration and new methods for modeling individual responses to treatment within machine learning.

1 INTRODUCTION

Good health care is one of the most important factors that can contribute to the personal well-being of everyone in the modern world. Unfortunately, health care is also costly. The World Health Organization (WHO) estimates that total health care spending in the United States was 17.0% of gross domestic product (GDP) in 2012, the highest in the world (World Health Organization, 2015). Expenditure as percentage of GDP in the United States is also growing, with the same estimate being 13.1% in 1990. Moreover, this trend of increasing costs is not unique to the United States. The WHO estimates that France spent 11.6% of GDP on health care in 2012, up from 10.1% in 1990. Germany spent 11.3%, up from 10.4%. Canada increased spending to 10.9% from 8.7%. The United Kingdom spent 9.3%, up from 6.9%. Overall, the trend represents a global phenomenon with few exceptions.

The current trend in health care costs cannot be maintained indefinitely. Health care systems across the globe will need to become more efficient at treating patients and reducing costs. As part of a much larger process to address this problem, the United States established the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009. HITECH introduced incentives for care providers to make meaningful use of electronic health records (EHR) with the ultimate goal of creating a more information-based, higher-quality health care system (Blumenthal, 2010). Since then, care providers have collected vast quantities of structured and unstructured data, but we have only begun to capitalize on the opportunity that new data and technology present to improve patient care.

1.1 Precision Medicine

In 2015, the President of the United States announced the Precision Medicine Initiative as part of the State of the Union address. As defined by the National Research Council (National Research Council, 2011):

> "Precision medicine" refers to the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those diseases they may develop, or in their response to a specific treatment.

The idea of making medical decisions and choosing treatment strategies based on individual variations is certainly not new, but with growing collections of EHR data, dropping costs in genomic sequencing, and greater access to computational resources, there has never been a better time to start such an initiative (Collins and Varmus, 2015). The precision medicine initiative is a call to arms, and the research community of clinicians, biologists, computer scientists, and more will need to work together to realize its full potential. While the initiative itself is specific to the United States, the potential impact is global.

1.2 Machine Learning in Medicine

The incentives are in place to develop an information-based health care system, but the challenge facing researchers then is what to do with all of the data and technology becoming available. The proliferation of larger medical datasets that include more relational, temporal, unstructured, and overall more complex data from multiple sources, urges the development

of new tools to make practical use of such heterogeneous datasets. Machine learning offers new tools that can be used to improve patient care (Page, 2015).

Machine learning (Mitchell, 1997) is a subfield of artificial intelligence focused on algorithms that "learn" from data to construct models that can be used to make predictions and decisions. Most people encounter machine learning every day, perhaps without even knowing it. For example, Google uses machine learning in its search engine to predict what pages are most relevant to user search keywords. Amazon uses machine learning to decide what products to suggest to a user based on which they are most likely to buy. Machine learning has already demonstrated global impact in business with companies like Google and Amazon, but it has yet to show a similar impact in clinical practice.

1.3 Thesis Statement

This dissertation focuses on developing new techniques in machine learning to improve patient care and investigates the following statement:

> Machine learning results can be made more clinically-relevant by tailoring current approaches to meet clinical objectives through the development of new algorithms to model individual response to treatment, and by incorporating clinical expertise into model development and refinement.

Many machine learning approaches are designed to optimize performance measures that are broadly applicable. In order to translate machine learning models to useful clinical applications, we need to reconsider and modify standard measures of performance to meet clinical objectives. In my work, I demonstrate new methods to improve collaboration between machine learning experts and clinicians, leveraging clinician expertise to

develop models that more effectively address real clinical objectives. I also develop new methods for modeling individual responses to treatment, building off of machine learning approaches originally developed in the marketing domain (Radcliffe and Surry, 1999).

1.4 Dissertation Organization

My dissertation is organized as follows.

Chapter 1 introduces the big picture problem and opportunity for applying machine learning in medicine. It is this chapter.

Chapter 2 introduces the basic machine learning background required to understand later chapters.

Chapter 3 introduces the basic medical background required to understand how machine learning can be applied effectively in a clinical setting.

Chapter 4 describes the three medical tasks that serve as the primary motivating applications in this dissertation.

Chapter 5 presents a study wherein we used inductive logic programming techniques to infer rules that could be used in clinical practice to determine when patients with suspicious mammogram findings might safely avoid surgery.

Chapter 6 presents a study for which we designed a framework of collaboration between computer science and clinical experts to leverage expert knowledge to produce machine learning models that more effectively meet clinical objectives.

Chapter 7 presents a study in which we developed a novel uplift modeling algorithm [1] to build models that capture characteristics of in situ breast cancer that are specific to older patients.

Chapter 8 presents a study in which we developed another novel uplift modeling algorithm to again capture characteristics of in situ breast cancer specific to older patients, and to capture characteristics of heart attack victims who had taken drugs called COX-2 inhibitors.

Chapter 9 presents a recent study in which we argue for the use of machine learning models in clinical studies in favor of more traditional models like logistic regression.

Chapter 10 discusses a few potential directions for future work, as well progress we have already made in these directions.

Chapter 11 wraps up this dissertation by briefly summarizing contributions.

[1] Uplift modeling is an approach from marketing used to assess individual responses to marketing activity.

Some basic understanding of various algorithms, evaluation metrics, and modeling approaches is required to understand much of this dissertation. In this chapter, we briefly review some basic machine learning concepts that show up in more detail in later chapters.

2.1 Bayesian Networks

A Bayesian network is a probabilistic graphical model that represents the conditional dependencies among a set of random variables. Each node in the network represents one of the variables while, edges between nodes represent conditional dependencies between the respective variables. The edges are directed and, given an edge that points from some node A to some other node B, A is referred to as a "parent" of B, and B is considered to be conditionally dependent on A.

Each node has an associated conditional probability table (CPT), or other conditional probability estimate, that represents the distribution over the values of the node variable, conditioned on the values of its "parents." To understand how this works, consider Figure 2.1. If we want to compute the joint probability of the four variables, $p(A, B, C, D)$, we can use the chain rule of probability to calculate it in terms of conditional probabilities.

$$P(A, B, C, D) = P(D|C, B, A)P(C|B, A)P(B|A)P(A)$$

Given the network in Figure 2.1, we know that some of the variables are conditionally independent of one another, given their parents. For example, we can see the D is conditionally independent of A, given B and C. This allows us to reduce the joint probability calculation.

$$P(A, B, C, D) = P(D|C, B)P(C|A)P(B|A)P(A)$$

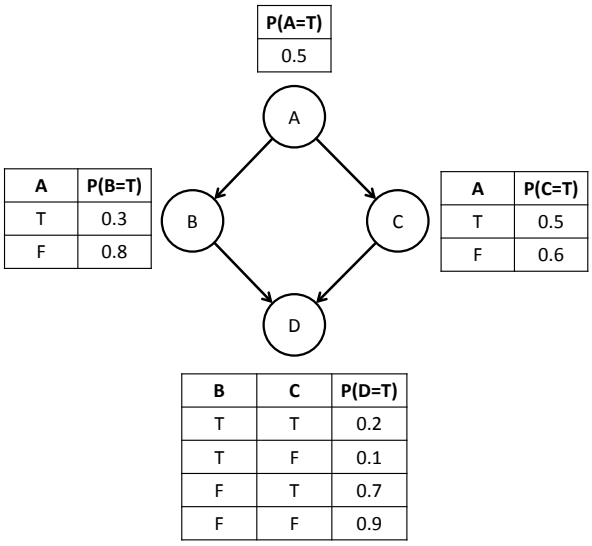


Figure 2.1: A Bayesian network.

The network is thus a compact representation of the joint probability distribution of the variables (Mitchell, 1997). Often in machine learning, a training set is used to learn the values in the CPTs given some network structure. Assuming that one of the variables is a class or outcome variable that we would like to estimate, the network can then be used to infer a distribution over the values of that class variable, given values for the other variables. There are, of course, methods for learning the structure of a network given a training dataset (Koller and Friedman, 2009), but the task is complicated, and we do not address it in depth here. Instead, we present two common approaches that make simplifying assumptions about the structure of the network.

2.1.1 Naïve Bayes

Because learning the structure of a Bayesian network is difficult, a Naïve Bayesian (NB) network makes the strong assumption that all variables are

dependent on the class variable, but are conditionally independent of one another, given the class label (see Figure 2.2a). To see how this assumption affects the model, consider Bayes' theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

In a typical binary supervised learning problem, we want to know the probability of the positive label given the features. If Y is our binary class variable and our features are X_i, we wish to compute the following:

$$P(Y_{pos}|X_1, ..., X_n)$$

Bayes' theorem tells us that we can compute that probability as:

$$\frac{P(X_1, ..., X_n|Y_{pos})P(Y_{pos})}{P(X_1, ..., X_n)}$$

The numerator can be expanded using the chain rule and the denominator can be rewritten as a marginal over the positive and negative labels (i.e. the possible values for Y):

$$\frac{P(X_1|Y_{pos})P(X_2|Y_{pos}, X_1)...P(X_n|Y_{pos}, X_1, ..., X_{n-1})P(Y_{pos})}{\sum^{Y} P(X_1|Y')P(X_2|Y', X_1)...P(X_n|Y', X_1, ..., X_{n-1})P(Y')}$$

Applying the assumption that all features are conditionally independent given the class label simplifies all of the conditional probabilities to depend only on the class label:

$$\frac{P(X_1|Y_{pos})P(X_2|Y_{pos})...P(X_n|Y_{pos})P(Y_{pos})}{\sum^{Y} P(X_1|Y')P(X_2|Y')...P(X_n|Y')P(Y')}$$

This makes computing the label distribution for any example easy because it can simply be computed from the relative frequencies of its variable settings in the dataset. Despite this strong, simplifying assumption,

NB has a proven history of good performance (Mitchell, 1997; Domingos and Pazzani, 1997).

2.1.2 Tree-Augmented Naïve Bayes

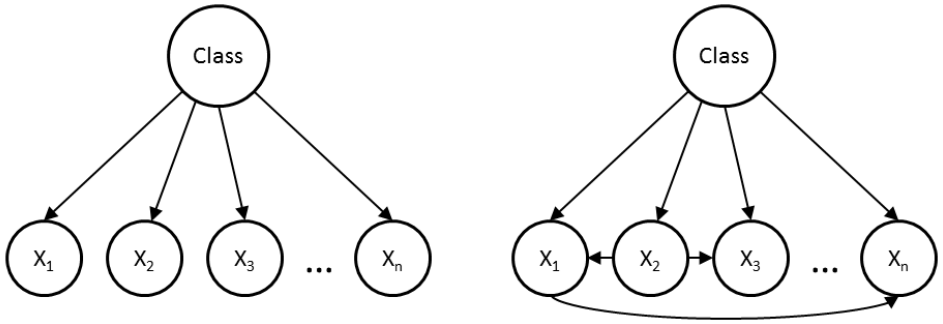
Tree-Augmented Naïve Bayes (TAN) is a modification of Naïve Bayes with the strong independence assumption relaxed (Friedman et al., 1997). Specifically, all of the non-class variables are still dependent on the class variable, but may also be dependent on one other non-class variable (see Figure 2.2b). This relaxation is accomplished by constructing a maximum-weight spanning-tree amongst all of the non-class variables. The weights of the edges used to construct the tree are the conditional mutual information of the two connected variables, conditioned on the class variable.

By building the maximum-weight spanning-tree on the graph of conditional mutual information between non-class variables, and by allowing non-class variables to be dependent on at most one other variable, TAN accomplishes two important properties.

1. It produces the maximum likelihood tree, given the data.
2. The tree can be learned in polynomial time.

2.2 Artificial Neural Networks

Artificial neural networks (ANN) are models used in machine learning that are inspired by the networks of neurons that make up a brain (Mitchell, 1997). They are typically represented by a network of nodes, organized into layers connected by edges. Each node in the network acts as a "neuron" that produces some output value based on the weighted sum of its inputs and an activation function. Each edge then acts as a weighted input to a node. For example, in Figure 2.3 the node labeled h_1 has four incoming



(a) Naïve Bayes model. (b) Tree-Augmented Naïve Bayes model.

Figure 2.2: Simplified Bayesian network models

edges from each of the nodes in the input layer. The input to h_1 is then a weighted sum of the input values.

$$In_{h_1} = x_1 w_1 + x_2 w_2 + x_3 w_3 + x_4 w_4$$

The output of h_1 is the result of an activation function on the weighted sum. For example, the logistic function is often used as an activation function, making the output of h_1:

$$Out_{h_1} = \frac{1}{1 + e^{-In_{h_1}}}$$

It becomes clear then that the outputs from the hidden layer serve as the inputs to the ouput layer, which acts identically. Note that Figure 2.3 only shows a single node in the output layer, but ANNs can be trained using an arbitrary number of outputs as needed by the task being learned.

The training process is one of learning the weights in the network such that the correct output is produced in the output layer. For one of the common ANN model types, multilayer perceptrons, this is accomplished using an algorithm called backpropagation (Mitchell, 1997).

At a high level, backpropagation works by first feeding the input from

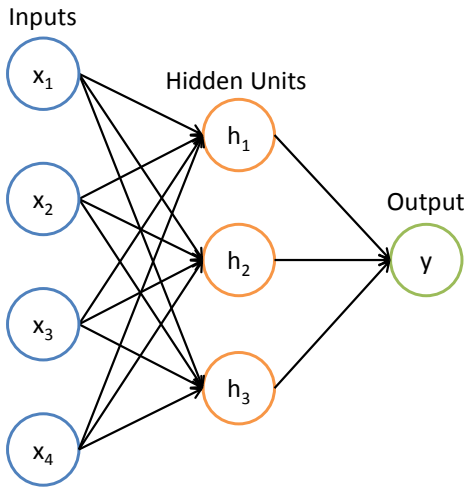


Figure 2.3: A simple feed-forward artificial neural network.

a training example (or examples) forward through the network to produce the final network output. The error is then computed based on the provided output for the training example. This error is propagated backward through the network, updating the weights along the way to reduce the error.

The goal of backpropagation is to find network weights that minimize the error, which is often accomplished using the gradient descent optimization method. With gradient descent, each weight in the network is updated according to the partial derivative of the error with respect to the weight. In order to accomplish this, the error function and activation function both need to be differentiable. Intuitively, gradient descent dictates that output error can be represented as a surface defined by input weights. For any particular point on the error surface, the slope can be computed and the weights can be updated to move downward on the error surface. This process is repeated until some minimum error is found. To avoid overfitting [1] to the training data, the process may be stopped early when

[1] A model is overfit when it demonstrates better accuracy on the training examples than on unseen examples, often because the model is overly complex.

error fails to reduce on a tuning set, rather than the training set.

2.3 Genetic Algorithms

Genetic algorithms are search algorithms that are modeled after natural selection (Mitchell, 1997). These algorithms are used to search through a space of candidate models to find one that performs the best according to some criteria. What defines the best model is a pre-specified numerical "fitness" function that can be used to rank models within the search space. Fitness functions vary from problem to problem, but one simple example might be the number of training examples a model correctly classifies in a training dataset. Models that correctly classify many examples in this scenario would then have greater fitness than other models that do not.

Implementations vary, but Algorithm 2.1 details a general genetic algorithm structure. First, an initial population of models is generated and their fitnesses calculated. This population is used to start an iterative process similar to natural selection. On each iteration, a top portion of the population is chosen to produce offspring through crossover and mutation operations somewhat analogous to real genetic processes. This new generation of offspring is then evaluated and replaces or is included in the current population. The iterative process continues like this until some number of epochs have passed or some stopping criterion is met.

A fundamental requirement for genetic algorithms is designing a genetic representation for the models being trained. One simple possibility is to represent models as a vector of weights that can be used to produce a linear combination of training example inputs. The crossover of model pairs can then be as simple as producing a new vector of weights of the same length as the parents, but with each weight being chosen randomly from one of the parents. This is very similar, in fact, to the approach we take for our experiments in Section 10.2.

Algorithm 2.1 Simple Genetic Algorithm

```
Pop ← GetRandModels();                          ▷ Init model population
Evaluate(Pop);                                  ▷ Compute initial fitness
for 1 to MaxEpochs do
    Top ← SelectBestK(Pop);                     ▷ Select models for crossover
    Pop′ ← Crossover(Top);                      ▷ Crossover pairs of models
    Mutate(Pop′);                               ▷ Small amount of random mutation
    Evaluate(Pop′);                             ▷ Compute new fitness
    Pop ← Pop′;
    if StopCriterionMet(Pop) then               ▷ May stop early
        break;
    end if
end for
M ← SelectBest1(Pop)                            ▷ Select final model
return M;
```

The possibility of overfitting the training set is a concern when using genetic algorithms for supervised machine learning. For example, the population may become crowded with similar individuals relatively with high fitness (Mitchell, 1997). Without diversity in the population, the algorithm may then stagnate and make no further progress toward other solutions with better fitness. This may be combated by changing the selection process for crossover, or by introducing new random models into the population occasionally.

2.4 Support Vector Machines

Traditional support vector machines (SVM) are two-class supervised learning models that are often used for classification (Cortes and Vapnik, 1995). Given a set of training data, an SVM constructs a maximum-margin hyperplane separating the two classes in the dataset. There may be many possible hyperplanes that separate the two classes, but the maximum-margin hyperplane is chosen to reduce overfitting. This hyperplane can

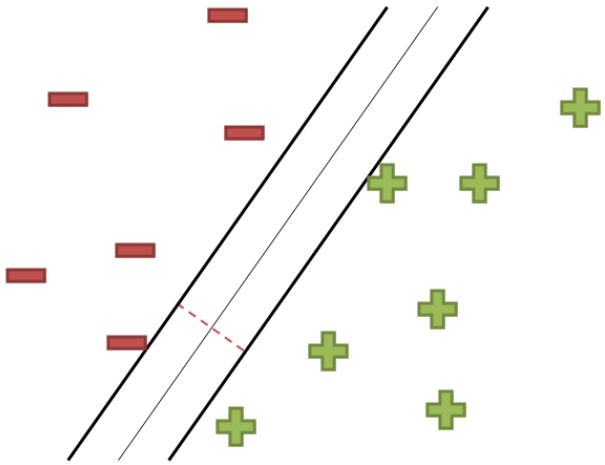


Figure 2.4: A simple SVM. The light line is the inferred decision boundary, and the dashed line shows the margin, where no examples lie.

then subsequently be used to classify new data, based on the side of the hyperplane that the new data point lies. Conveniently, if the training data are not linearly separable in the original feature space, the problem can be mapped to a higher-dimensional space using a kernel function, effectively allowing for a non-linear hyperplane in the original feature space.

As a brief introduction, consider the standard definition of the maximum margin classifier (Vapnik, 1998). This classifier minimizes:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{2.2}$$

Subject to $\xi_i \geqslant 0$. The formulation tries to minimize the two-norm of the weight vector, and hence maximize the margin, while softly allowing some errors, ξ_i, whose cost depends on the tunable parameter C. Errors here are distances from the decision boundary of examples that lie on the wrong side of the boundary.

2.5 Inductive Logic Programming

ILP is a machine learning approach that induces a set of rules (a theory) in first-order logic that predict some outcome based on a given dataset. It does this by searching through a rule search space built from a collection of background facts and rules. Rules within the search space are scored and are added to the current theory if they meet specified acceptance criteria. This search process continues until some stopping criteria are met. There are many different scoring functions that may be used to score candidate rules. Often, the scoring function is some form of coverage, whereby the learner attempts to induce rules that entail many positive examples and few (ideally zero) negative examples. Just as there are many possible scoring functions, there are a variety of search algorithms that may be used, such as breadth-first or depth-first search (Lavrac and Dzeroski, 1994).

See Figure 2.5 for a simple example where the goal is to induce the definition of the *uncle* relationship. In this example, relationships between individuals are defined by predicates. For example $\mathtt{Parent}(\mathtt{A}, \mathtt{B})$ can be read as "A is the parent of B". In this problem, the goal is to induce a general definition of $\mathtt{Uncle}(\mathtt{A}, \mathtt{B})$ using the facts provided, along with positive and negative examples of the uncle relationship.

ILP may be preferable over standard classification algorithms for a number of reasons. One important advantage is that ILP is capable of working on multi-relational data (De Raedt, 2008), which is prevalent in domains where the datasets come directly from database systems. This can reduce the need to collapse (i.e. propositionalize) a rich, relational dataset into a standard format using summarization or other techniques that may lose information (De Raedt, 2008). Another advantage is that the rules in a theory are expressed in an if-then fashion, making them easy to understand and reason about. This can be particularly important when the goal is knowledge discovery.

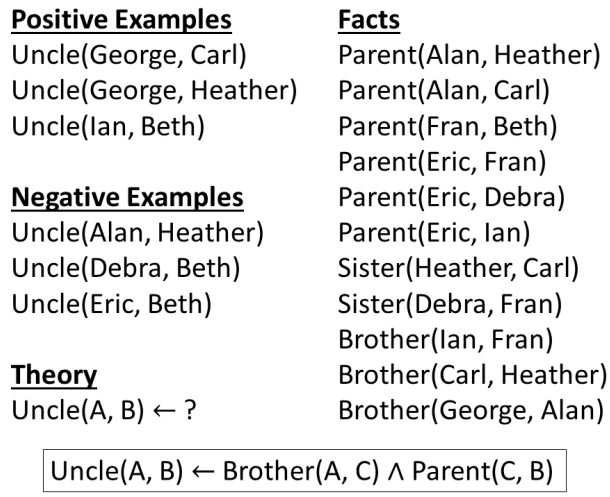


Figure 2.5: A simple ILP example showing facts and examples that may be used to learn a logical definition of the uncle relationship.

2.6 Receiver Operating Characteristic

Equally important to the algorithms that are used to train models from data, are the methods used to evaluate the trained models. The *receiver operating characteristic* (ROC) curve is a plot that can be used to understand the predictive performance of a binary classifier (Hastie et al., 2009). Specifically, the ROC curve illustrates the true positive rate (or recall) a classifier achieves versus false positive rate when varying the classifier's discrimination threshold. Each point on an ROC curve then represents a threshold for classification and defines the proportion of positive examples that the classifier correctly labels as positive at the threshold, versus the proportion of negative examples it incorrectly labels as positive.

Figure 2.6 shows two examples of ROC curves based on two separate rankings of five positive and five negative examples. As an intuitive simplification, one can imagine constructing an ROC curve by stepping through a classifier's ranking of positive and negative examples from those labeled most likely to be positive, to those least likely to be positive. Each step

that has a positive example moves the curve up, and each step that has a negative example moves the curve to the right. A perfect ranking of all positive examples followed by all negative examples then achieves the optimal ROC curve that moves from the bottom left, straight up to the top left, and then straight to the top right.

While each point within an ROC curve has important implications for understanding a classifier's performance, the area under the ROC (AUC) curve is often used as a performance metric in place of evaluating at a fixed threshold of discrimination (Hanley and McNeil, 1982). A perfect ranking has an AUC of 1.0, and a random ranking has an expected area of 0.5.

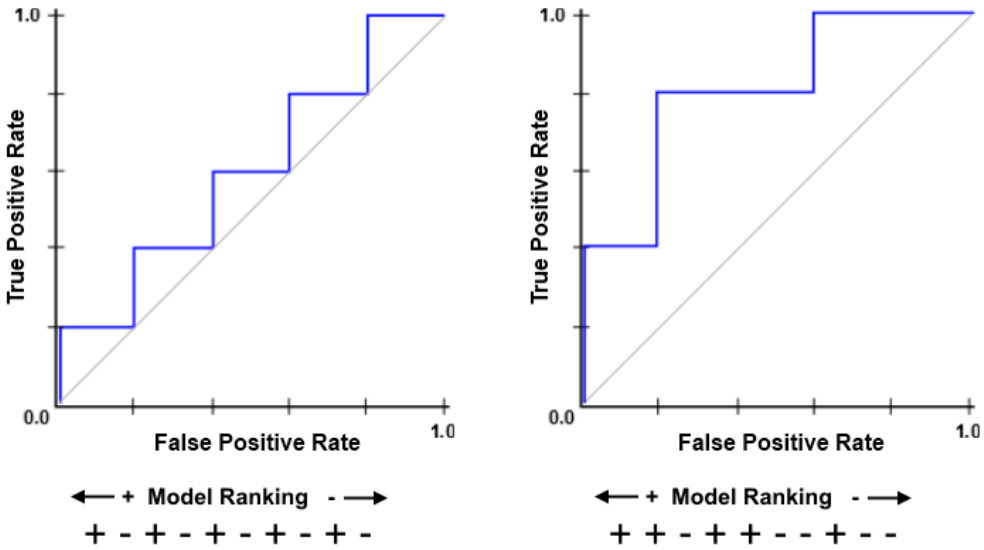


Figure 2.6: Two different ROC curves based on the ranking of five positive and five negative examples. Below each curve is the ranking that defines the curve, where the left-most example is labeled most positive.

2.7 Uplift Modeling

Differential prediction is motivated by studies where one submits two different subgroups from some population to stimuli. The goal is to gain insight on the different reactions by producing, or simply identifying, a classifier that demonstrates significantly better predictive performance on one subgroup (the *target* subgroup) over another (the *control* subgroup). Seminal work in sociology and psychology used regression to study the factors accounting for differences in the academic performance of students from different backgrounds (Cleary, 1968; Linn, 1978; Young, 2001). *Uplift modeling* is a popular technique in marketing studies. It measures the impact of a campaign by comparing the purchases made by a subgroup that was targeted by some marketing activity versus a control subgroup (Lo, 2002; Hansotia and Rukstales, 2002; Radcliffe, 2007).

In marketing, customers can be broken into four categories (Radcliffe and Simpson, 2008):

Persuadables Customers who respond positively (e.g. buy a product) when targeted by marketing activity.

Sure Things Customers who respond positively regardless of being targeted.

Lost Causes Customers who do not respond (e.g. not buy a product) regardless of being targeted or not.

Sleeping Dogs Customers who do not respond as a result of being targeted.

Thus, targeting *Persuadables* increases the value produced by the marketing activity, targeting *Sleeping Dogs* decreases it, and targeting customers in either of the other groups has no effect, but is a waste of money. Ideally then, a marketing team would only target the *Persuadables* and

avoid targeting *Sleeping Dogs* whenever possible. Unfortunately, the group to which a particular individual belongs is unknown and is not readily observable. An individual cannot be both targeted and not targeted to determine their response to marketing activity directly. Only the customer response and whether they were in the target or control group can be observed experimentally (see Table 2.1).

Table 2.1: Customer groups and their expected responses based on targeting. Only the shaded region can be observed experimentally.

Target		Control	
Response	No Response	Response	No Response
Persuadables, Sure Things	Sleeping Dogs, Lost Causes	Sleeping Dogs, Sure Things	Persuadables, Lost Causes

In this scenario, since we cannot observe customer groups beforehand, standard classifiers appear less than ideal. For example, training a standard classifier to predict response, ignoring the differences in response between the target and control subgroups is likely to result in a classifier that identifies *Persuadables*, *Sure Things*, and *Sleeping Dogs* because they represent the responders when the target and control subgroups are combined. Recall, however, that targeting *Sure Things* is a waste of money, and targeting *Sleeping Dogs* is harmful. Even training on just the target subgroup ("response modeling") is likely to produce a classifier that identifies both *Persuadables* and *Sure Things* (Lo, 2002).

The point of uplift modeling is then to quantify the difference between the target and control subgroups, producing a classifier that maximizes predictive performance on the target subgroup over the control subgroup (i.e. maximize uplift). The idea is that such a classifier characterizes properties that are specific to the target subgroup, thereby making it effective at identifying *Persuadables*. That is, such a classifier will produce a larger output for customers who are more likely to respond positively

as a direct result of targeting, and a smaller output for those who are unaffected or are more likely to respond negatively. The classifier could then be used in subsequent marketing campaigns to select who should be targeted and who should not. To validate that uplift modeling can, in fact, produce models that separate the *Persuadables* from other customer groups, we ran experiments using a simulated marketing campaign (see Sections 4.4 and 8.5.1).

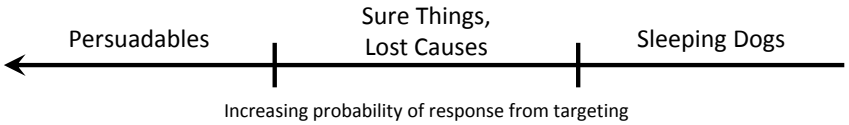


Figure 2.7: Ideal ranking of uplift model.

Seminal work includes Radcliffe and Surry's true response modeling (1999), Lo's true lift model (2002), and Hansotia and Rukstales' incremental value modeling (2002). As an example, Hansotia and Rukstales construct a regression and a decision tree, or CHART, model to identify customers for whom direct marketing has sufficiently large impact. The splitting criterion is obtained by computing the difference between the estimated probability increase for the attribute on the target set and the estimated probability increase on the control set.

In some applications, especially medical decision support systems, gaining insight into the underlying classification logic can be as important as system performance. Developments include tree-based approaches to uplift modeling (Radcliffe and Surry, 2011; Rzepakowski and Jaroszewicz, 2012), although ease-of-interpretation was not an objective in their motivating applications. Wanting to maximize rule interpretability, Nassif et al. (2012a) opted for ILP-based rule learning instead of decision-trees because the latter is a special case of the former (Blockeel and De Raedt, 1998). To the best of our knowledge, the first application of uplift modeling in medical domains is due to Jaśkowski and Jaroszewicz (2012), who adapt

standard classifiers by using a simple class variable transformation. Their transformation avoids using two models by assuming that both sets have the same size and combining the examples into a single set.

2.7.1 Uplift and Lift

The common evaluation metric used in uplift modeling is the uplift measure. *Lift*, like ROC, is a measure that can be used to evaluate the predictive performance of a binary classifier (Tufféry, 2011). Let P be the number of positive examples and N the number of negative examples in a given dataset D. Lift represents the number of true positives detected by a model amongst the top-ranked fraction ρ. Varying $\rho \in [0, 1]$ produces a lift curve. The area under the lift curve (AUL) for a given model and data becomes:

$$AUL = \int Lift(D, \rho) d\rho \approx \frac{1}{2} \sum_{k=1}^{P+N} (\rho_{k+1} - \rho_k)(Lift(D, \rho_{k+1}) + Lift(D, \rho_k)) \tag{2.3}$$

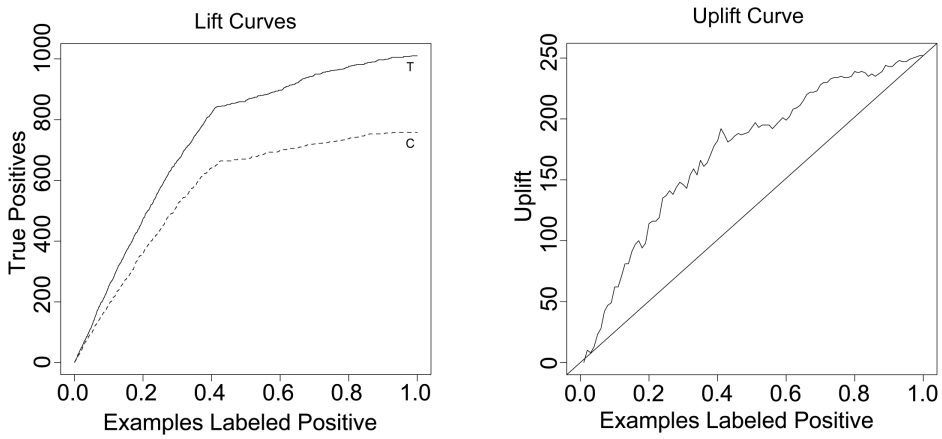
The *uplift curve* compares the difference between the model M over two groups, target T and controls C (Rzepakowski and Jaroszewicz, 2010). Recall that the uplift modeling desires to produce a classifier that maximizes the predictive performance on the target subgroup of the control subgroup. The uplift curve quantifies that difference. It is defined by:

$$Uplift(M_T, M_C, \rho) = Lift_{M_T}(T, \rho) - Lift_{M_C}(C, \rho). \tag{2.4}$$

Since each point in the uplift curve is a function of a single value for ρ, the area under the uplift curve (AUU) is the difference between the areas under the lift curves of the two models (see Figure 2.8 for an example).

$$AUU = AUL_T - AUL_C \tag{2.5}$$

Uplift modeling is effectively a differential prediction approach aimed



(a) Two example lift curves. (b) The corresponding uplift curve.

Figure 2.8: A simple example of two lift curves for target and control subgroups, and a corresponding uplift curve.

at maximizing uplift (Hansotia and Rukstales, 2002; Lo, 2002; Radcliffe and Surry, 1999).

2.7.2 Relation of Lift to ROC

In order to obtain more insight into the uplift measure it is beneficial to compare lift curves with ROC curves. Recall that AUL is the area under the lift curve, and AUC is the area under the ROC curve. There is a strong connection between the lift curve and the ROC curve. If we let $\pi = \frac{P}{P+N}$ be the prior probability for the positive class or *skew*, then:

$$AUL = P \times (\frac{\pi}{2} + (1 - \pi)\ AUC) \quad \text{(Tufféry, 2011, p. 549).} \tag{2.6}$$

In uplift modeling, the aim is to optimize for uplift over two sets, that is, obtaining new classifiers such that $AUU^* > AUU$, where $AUU =$

$AUL_T - AUL_C$. The equation $AUU^* > AUU$ can be expanded into:

$$AUL^*_T - AUL^*_C > AUL_T - AUL_C. \tag{2.7}$$

Further expanding and simplifying:

$$P_T(\frac{\pi_T}{2} + (1-\pi_T)AUC^*_T) - P_C(\frac{\pi_C}{2} - (1-\pi_C)AUC^*_C) >$$
$$P_T(\frac{\pi_T}{2} + (1-\pi_T)AUC_T) - P_C(\frac{\pi_C}{2} - (1-\pi_C)AUC_C)$$
$$P_T(1-\pi_T)AUC^*_T - P_C(1-\pi_C)AUC^*_C > P_T(1-\pi_T)AUC_T - P_C(1-\pi_C)AUC_C$$
$$P_T(1-\pi_T)(AUC^*_T - AUC_T) > P_T(1-\pi_T)(AUC^*_C - AUC_C)$$

and finally

$$\frac{AUC^*_T - AUC_T}{AUC^*_C - AUC_C} > \frac{P_C}{P_T}\frac{1-\pi_C}{1-\pi_T}. \tag{2.8}$$

In a balanced dataset, $\pi_C = \pi_T = \frac{1}{2}$ and $P_C = P_T$, so that $\frac{1-\pi_C}{1-\pi_T} = 1$. In fact, if the target and control groups have the same skew, maximizing the difference in AUL implies maximizing the difference in AUC as well. If the target and control groups do not have the same skew, we can not make this conclusion. In general, the conclusion is that the two tests are related, but that uplift is sensitive to variations of dataset size and skew. In other words, uplift is more sensitive to variations in coverage when the two groups have different size.

3 MEDICAL BACKGROUND

We next discuss some basic insights into how some clinical practice works and how it might be improved. In particular, we discuss how tools may be employed to assist in clinical decision-making, how treatment decisions are currently made, how modeling might improve clinicians' ability to choose treatment.

3.1 Clinical Decision Support

Clinicians may have to deal with many different patients and may also have a great deal of information about patients available to them. It is the clinician's job to use that information to correctly diagnose patients, and to correctly treat patients. In order to know how to use that information, however, clinicians must obtain and maintain a vast pool of knowledge about diseases, treatments, expected outcomes, and more. Substantial research in clinical trials and systematic reviews help to guide clinical practice, but the task is monumental, and critically important for providing good patient care. Furthermore, the proliferation of electronic health records (EHR) presents both a benefit and a burden to clinicians. The benefit is that information can be recorded in more structured formats for later use, but the burden is that clinicians have yet another skill set that must to be learned and maintained. It is perhaps unsurprising then that medical care is variable and often suboptimal across health care systems (Roshanov et al., 2013). Clinical decision support systems, systems that assist clinicians in diagnosing and appropriately treating patients, have a great opportunity to improve upon current patient care. In fact, research has already shown the potential of such systems to assist with problems in clinical practice, increase clinician adherence to protocol, and improve health care systems overall (Roshanov et al., 2013; Kawamoto et al., 2005;

Chaudhry et al., 2006; Ash et al., 2012). These tools may come in different forms, but we focus on the growing field of machine learning in clinical decision support.

3.2 Clinical Trials

One of the fundamental challenges faced in public health and patient care is estimating the risk of disease that can be attributed to various exposures or treatments. Researchers run clinical trials to study these exposures or treatments (Friedman et al., 2010), with randomized controlled trials (RCT) being considered the gold standard for estimating average treatment effects (ATE). In an RCT, study patients are randomized to different treatment arms (e.g. a treatment and control arm), after which the rate or probability of some outcome is measured (see Figure 3.1). Randomization is important as it balances confounding variables, leading to measures of treatment effect that are free of bias. The difference in outcome rates between treatment and control determines the average treatment effect, as shown in Equation 3.1. Treatment can demonstrate either a positive or a negative ATE, with success being determined by the desirability of the measured outcome. The treatment arm with the best success rate is selected as the preferred treatment. However, while the ATE is indicative of the true treatment effect, we can expect a diversity of effects in individuals. This makes ATE estimates less applicable for individual patients. Furthermore, the ATE is population-distribution dependent, so it inherently lacks generalizability to alternative test distributions. It would be far more useful to estimate the individualized treatment effects, which provides the effect per individual instead of at the population level.

$$\mathrm{ATE} = \mathrm{P}(Y = \mathtt{true} | \mathrm{Treat} = \mathtt{true}) - \mathrm{P}(Y = \mathtt{true} | \mathrm{Treat} = \mathtt{false}) \quad (3.1)$$

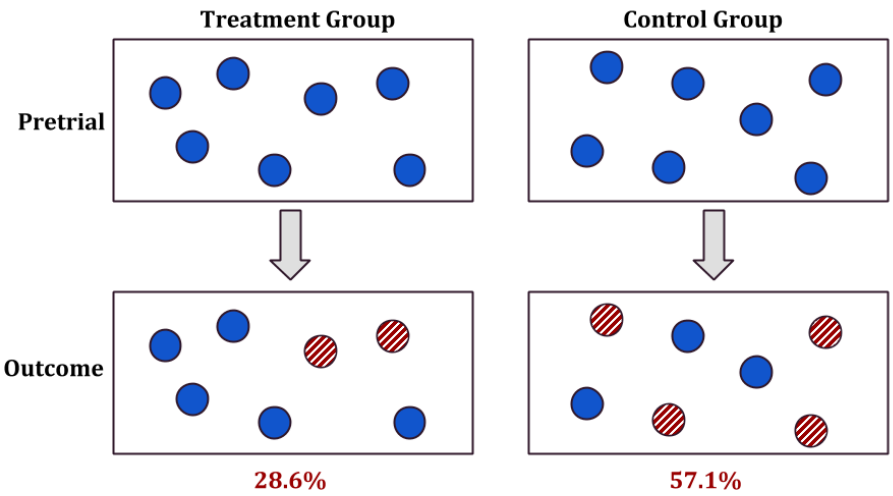


Figure 3.1: The observed average treatment effect (ATE) of a randomized controlled trial. Striping here indicates the occurrence of some outcome of interest (e.g. heart attack), while the others do not experience the outcome of interest. The control group exhibits a 57.1% rate of the measured outcome, and the treatment group shows a 28.6% rate. The ATE is then a reduction in the rate of the outcome of interest.

3.3 Individualized Treatment Effects

Modeling individualized treatment effects (ITE) considers individuals separately based on the features or characteristics associated with that individual. That is, the ITE acknowledges the fact that individuals vary in important ways that may affect how the treatment or exposure affects disease outcome (see Figure 3.2). Equation 3.2 shows the difference between the ITE and the ATE, where the ITE does not marginalize over individual features, X. With an ITE model, information about a future individual can then be leveraged to determine the optimal treatment choice for that individual.

$$\text{ITE} = P(Y = true|Treat = true, X) - P(Y = true|Treat = false, X) \quad (3.2)$$

Estimation of the ITE has many important clinical applications. For example, medications almost certainly have different effects in different individuals. Hormone replacement therapy treatment effect findings in RCTs and observational studies were of opposite sign for coronary heart disease, and advocacy of their use was rescinded when the RCT findings were released (Manson et al., 2013). Yet, estrogen therapy is still the first line treatment among women experiencing hot flushes. This raises the question of whether ITE modeling can help determine subsets of patients who are still likely to receive benefit. Similarly, many drugs are taken off the market due to excess harm from adverse drug effects. Accurate ITE estimation could bring such drugs safely back to market for select subpopulations.

The limitations of applying population-average effect estimation to individuals have already been acknowledged (Kent and Hayward, 2007; Rothwell, 1995), and work on modeling the ITE has already begun (Qian and Murphy, 2011). Modeling the ITE, however, is a more challenging task than the ATE, as treatment effects simply cannot be observed at an individual level. A researcher cannot both treat and not treat an individual and then measure the difference in outcomes. Once a treatment option is assigned, the counterfactual outcome, the outcome of the treatment not given, cannot be observed. Furthermore, unlike ATE estimation, gathering sufficient data to adequately estimate the counterfactual ITE outcome is challenging for even modest numbers of individual features, as the state space grows exponentially. Thus, modeling approaches to estimate the counterfactual outcome become necessary.

Weiss et al. (2015) suggest leveraging more recent developments in machine learning to model the conditional probability distribution (CPD)

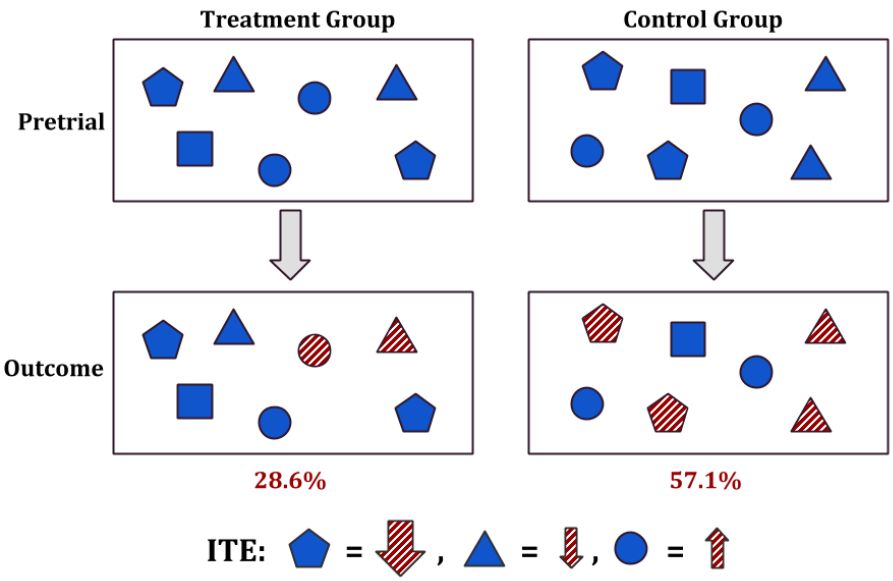


Figure 3.2: The individualized treatment effect (ITE) of a randomized controlled trial. Striping here indicates the occurrence of some outcome of interest (e.g. heart attack), while the others do not experience the outcome of interest. The average treatment effect is still a reduction in the rate of the measured outcome overall, but the ITE shows a decreased rate for some and an increased rate for others. The pentagon subgroup saw a large reduction in occurrence, while the triangles and circles saw a small reduction and increase respectively.

to estimate counterfactual outcomes, rather than using more traditional techniques like logistic regression. In particular, they suggest using AdaBoost (Freund and Schapire, 1995) because it has consistency [1] results and is a non-parametric [2] learning algorithm. The details of this study are explained in more detail in Chapter 9.

Of note is the similarity of the ITE estimation problem and that of

[1] As a consistent learner is given more data examples, its estimates converge toward the true distribution of the data.

[2] Parametric models make assumptions about the distribution from which data are drawn, whereas non-parametric models do not.

uplift modeling (see Section 2.7). Ideally, in marketing, we would be able to estimate the change in probability that a person will buy a product after having been targeted by some marketing activity versus having not been targeted. Once again, this difference is unobservable because once an individual has been targeted with the marketing activity, the counterfactual cannot be observed. The analogous equation then would be that found in Equation 3.3.

$$\begin{aligned} \mathit{Marketing\ ITE} = \mathrm{P}(\mathrm{Buy} = \mathtt{true} | \mathrm{Target} = \mathtt{true}, \mathrm{X}) \\ - \mathrm{P}(\mathrm{Buy} = \mathtt{true} | \mathrm{Target} = \mathtt{false}, \mathrm{X}) \end{aligned} \tag{3.3}$$

Persuadables are individuals for which the marketing ITE is positive. That is, their probability of buying after having been targeted with marketing activity is greater than their probability of buying if they had not been targeted. Similarly, *Sleeping Dogs* are individuals for which the marketing ITE is negative. That is, their probability of buying after having been targeted with marketing activity is less than their probability of buying otherwise. *Sure Things* and *Lost causes* have a marketing ITE of 0. Their probability of buying is independent of whether or not they are targeted with marketing activity. As described in section 2.7 though, the marketing domain tries to address this ITE estimation by optimizing for the uplift metric. Instead of modeling the CPD, the goal is to build a model that captures information specific to *Persuadables* by separating the treatment and control groups. Optimizing for the uplift metric then can be seen as an alternative approach to estimating the ITE.

4 APPLICATIONS AND DATASETS

There are certainly many potential applications of machine learning in medicine, but most of our work has focused on three particular tasks, two related to breast cancer, and one to adverse drug events. In one task, the goal is to determine appropriate treatment for patients who have received breast biopsies with a non-definitive [1] diagnosis. In another, the goal is to determine factors that differentiate the more indolent in situ breast cancer of older patients from that of younger patients when predicting the progression of breast cancer. In our third task, the goal is to identify patients who are most susceptible to the adverse reactions from taking COX-2 inhibitors.

We also refer to three other applications and datasets to a lesser extent in some chapters. One is dataset generated by simulating a marketing campaign on a simulated customer population. Another is a synthetic dataset that represents the use of statins to reduce the occurrence of myocardial infarction (heart attack). Finally, the last is a real dataset from a randomized controlled trial investigating the effect of using D-penicillamine to treat primary biliary cirrhosis [2]. This chapter describes all of these applications in more detail.

4.1 Upgrade Prediction

When a patient presents with a suspicious breast lesion, a diagnostic mammogram and possibly ultrasound are performed to further define the abnormality. If the finding remains suspicious, a core needle biopsy (CNB) is often recommended (Bevers et al., 2009). In this procedure, a

[1] A biopsy is non-definitive if radiologists and pathologists cannot together make a conclusive benign or malignant diagnosis.

[2] Primary biliary cirrhosis is an autoimmune disease of the liver.

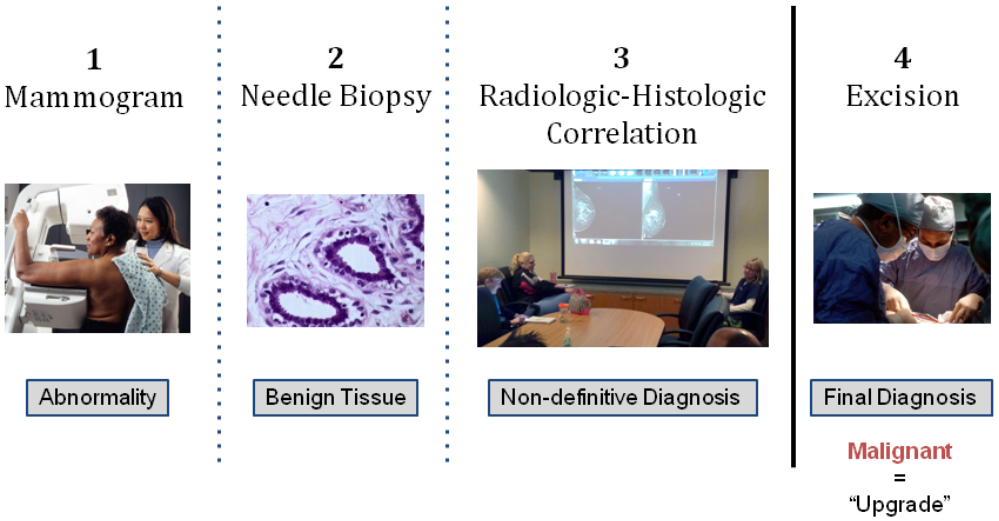
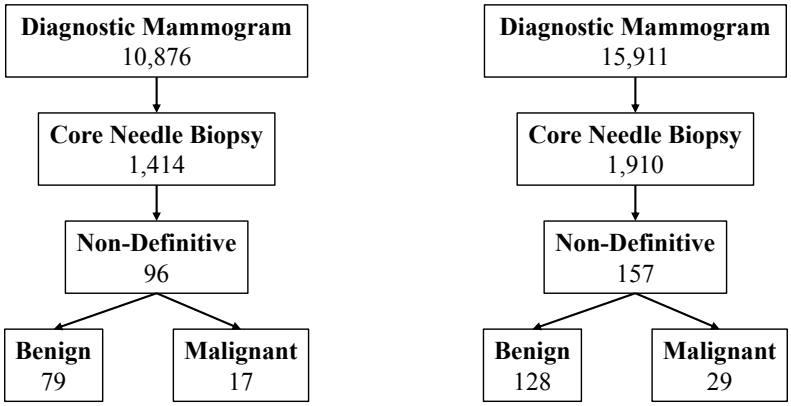


Figure 4.1: The clinical process of a non-definitive biopsy. In Step 1, a woman presents with suspicious imaging, and a needle biopsy is recommended. In Step 2, the pathologist gives the biopsy a benign diagnosis. In Step 3, radiologists and pathologists determine that no definitive diagnosis can be made, and surgery is recommended. Finally, in Step 4, surgery is performed and a final, definitive diagnosis of the abnormality is made. Image sources: 1) NIH - `http://wikimedia.org/wiki/File:Woman_receives_mammogram.jpg` 2) Itayba - `http://wikimedia.org/wiki/File:Normal.jpg` 3) UW Hospital and Clinics 4) NIH - `http://wikimedia.org/wiki/File:Surgical_breast_biopsy.jpg`

needle is inserted into the breast under imaging guidance to remove small samples ("cores") of the targeted breast abnormality. Subsequently a correlation between the histology results and the imaging features (Radiologic-Histologic correlation) is performed to ensure adequate sampling of these lesions and avoid cancers being missed (Liberman, 2000). The majority of breast biopsies will yield definitive results (Bruening et al., 2010). However, in 5% to 15% of cases, the results are non-definitive (Liberman, 2000), and surgical excisional biopsy is recommended to determine the final pathology and rule out the presence of malignancy (see Figure 4.1). If a malignancy is subsequently confirmed, the case is "upgraded" from



(a) Recruitment through 2009. (b) Recruitment through 2011.

Figure 4.2: The case-inclusion process for non-definitive biopsies from the UW Hospital and Clinics. We included consecutive core needle biopsy cases recommended from a diagnostic mammogram, and considered those given a non-definitive diagnosis. There were 96 non-definitive biopsies from January 1, 2006 through December 31, 2009 (4.2a). There were 157 non-definitive biopsies from January 1, 2006 through December 31, 2011 (4.2b).

non-definitive to malignant (approximately 10-20%). In the US, women over the age of 20 have an annual breast biopsy utilization rate of 62.6 per 10,000, translating to over 700,000 women undergoing breast core biopsy in 2010 (CDC, 1998). Knowing this, approximately 35,000 to 105,000 of these women likely underwent excision, an additional and more invasive procedure. Ultimately, a majority of these women receive a benign diagnosis.

In the mid-1990s, the American College of Radiology developed the mammography lexicon, Breast Imaging Reporting and Data System (BI-RADS), to standardize mammogram feature distinctions and the terminology used to describe them (BIR, 2003), and studies show that BI-RADS descriptors are predictive of malignancy (Liberman et al., 1998), specific histology (Burnside et al., 2004), and prognostic significance (Tabar et al.,

Table 4.1: The upgrade prediction dataset from the UW Hospital and Clinics. Counts of cases from January 1, 2006 through December 31, 2009 are found in 4.1a, and those through December 31, 2011 are found in 4.1b

(a) January 1, 2006 through December 31, 2009.

Features	Total Cases	Benign Cases	Malignant Cases
70	96	79	17

(b) January 1, 2006 through December 31, 2011.

Features	Total Cases	Benign Cases	Malignant Cases
70	157	128	29

2004). Given that 1) a complex combination of variables predicts upgrade, 2) reliable data including accurate outcomes via cancer registries are available, and 3) accurate prediction of upgrade would substantially improve management, this domain is ripe for decision support that would have a substantial impact on patient care.

The dataset we use for this task is collected from the University of Wisconsin clinical practice. We are continually growing the dataset through the prospective collection of new cases and retrospective collection of new features. Our dataset contains information about demographic risk factors (e.g. age, personal history of breast cancer), BI-RADS descriptors of abnormalities in mammograms, description of pathologies, and technical information about the biopsies as shown in Figure 4.3. Our recruitment of cases from January 1, 2006 to December 31, 2011 includes a population of patients that underwent 1,910 consecutive CNB, as a result of a diagnostic mammogram. Clinicians prospectively gave a total of 157 of those biopsies a non-definitive diagnosis, of which 128 (81.5%) were found to be benign and 29 (18.5%) were found to be malignant (see Table 4.1). Some of our earlier work included a subset of cases from January 1, 2006 to December 31, 2009. This subset includes 96 non-definitive biopsies, of which 79 (82.3%) were found to be benign and 17 (17.7%) were found to

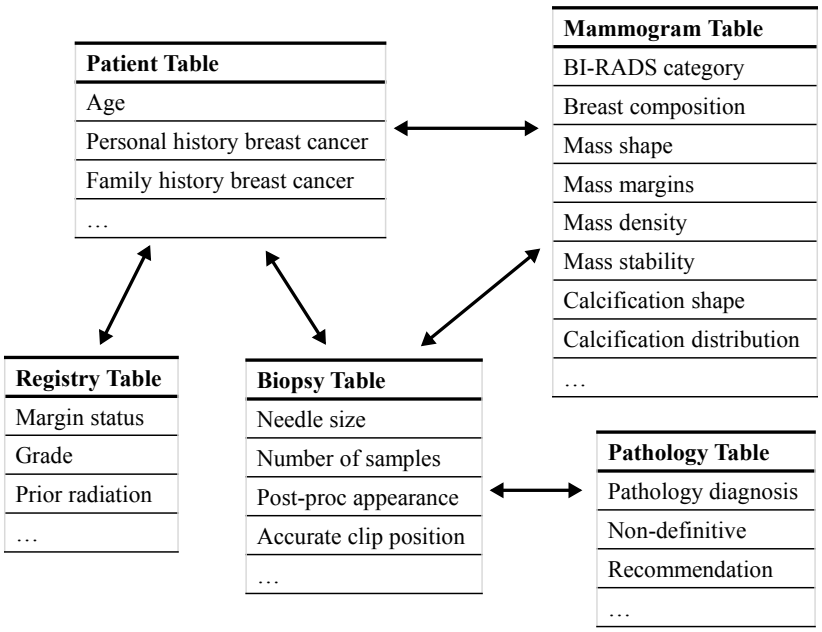


Figure 4.3: The breast-imaging database from which all of our examples are collected. Each box is a table in our database, and arrows represent relations between the different tables.

be malignant. Figure 4.2 shows the recruitment process in more detail for both time periods.

4.2 COX-2 Inhibitors

COX-2 inhibitors are a family of non-steroidal anti-inflammatory drugs (NSAIDs) used to treat inflammation and pain by directly targeting the COX-2 enzyme, without affecting the COX-1 enzyme. This is a desirable property as it significantly reduces the occurrence of various adverse gastrointestinal effects common to other NSAIDs (Russell, 2001). As such, COX-2 inhibitors, specifically Vioxx, Bextra, and Celebrex, enjoyed widespread acceptance in the medical community. Unfortunately, additional patient data later showed that the use of COX-2 inhibitors also

came with a significant increase in the rate of myocardial infarction (MI), or "heart attack" (Kearney et al., 2006). As a result, Vioxx and Bextra were pulled from the market, and Celebrex added a warning to the label. Physicians must be much more careful when prescribing these drugs. In particular, physicians want to avoid prescribing COX-2 inhibitors to patients who may be more susceptible to the adverse effects that they entail. For this problem, predicting the individualized treatment risk of MI given treatment with COX-2 inhibitors, versus no treatment, is the appropriate task to identify the at-risk patients.

Recall the similarity of the uplift modeling goal to individualized treatment effect estimation (see Section 3.3). An individual cannot both take the drug and not take the drug to determine its effect. Only the MI outcome and whether or not the individual took the drug can be observed experimentally. We propose that training a classifier to identify individuals for which taking a COX-2 inhibitor increases their risk of MI is analogous to identifying *Persuadables*.

Furthermore, the U.S. Food and Drug Administration (FDA) has recently issued a new warning about all non-aspiring NSAIDs adding cardiovascular risk (FDA, 2015). In the near future, it may be desirable to extend analysis from COX-2 inhibitors to all non-aspirin NSAIDs. In some ways though, this is more challenging since many NSAIDs are over-the-counter and records will be incomplete. We do not address this new warning in our work, but it is important to note as it broadens the potential scope of this task.

For this task, we use a dataset collected at Marshfield Clinic, which has been previously used in Davis et al. (2013). The dataset consists of information from multiple database tables: lab test results (e.g., cholesterol levels), medications taken (both prescription and non-prescription), disease diagnoses, and observations (e.g., height, weight and blood pressure). Patients are separated into two equally-sized subgroups: patients

who have been prescribed COX-2 inhibitors and those who have not. The group prescribed COX-2 inhibitors has 184 patients who had MI, and 1,776 who did not. The subgroup not prescribed COX-2 inhibitors has the same number of patients for each outcome (see Table 4.2).

Table 4.2: Composition of the COX-2 dataset from Marshfield Clinic. There are 12,496 features for each example in this dataset. The group prescribed COX-2 inhibitors has 184 patients who had MI, and 1,776 who did not. The subgroup not prescribed COX-2 inhibitors has equal numbers.

	COX-2 Inhibitors		No COX-2 Inhibitors	
Features	MI	No MI	MI	No MI
12,496	184	1,776	184	1,776

4.3 Invasive vs. In Situ Breast Cancer Prediction

Breast cancer is the most common cancer among women (American Cancer Society, 2009b) and has two basic states: an earlier *in situ* state where cancer cells are still localized, and a subsequent *invasive* state where cancer cells infiltrate surrounding tissue (see Figure 4.4). Nearly all in situ cases can be cured (American Cancer Society, 2009a), thus current practice is to treat in situ occurrences in order to avoid progression into invasive tumors (American Cancer Society, 2009b). Treatment, surgery sometimes followed by radiation therapy, is costly and may produce undesirable side-effects. Moreover, an in situ tumor may never progress to invasive state in the patient's lifetime, increasing the possibility that treatment may not have been necessary. In fact, younger women tend to have more aggressive cancers that rapidly proliferate, whereas older women tend to have more indolent cancers (Fowble et al., 1994; Jayasinghe et al., 2005). Because of this, younger women with in situ cancer should be treated due to a

greater potential time-span for progression. Likewise, it makes sense to treat older women who have in situ cancer that is similar in characteristics to in situ cancer in younger women since the more aggressive nature of cancer in younger patients may be related to those features. However, older women with in situ cancer that is significantly different from that of younger women may be less likely to experience rapid proliferation, making them good candidates for "watchful waiting" instead of treatment.

The motivating problem at hand can readily be cast as an uplift modeling problem (see Section 2.7). Like the hidden customer types in uplift modeling, which type of cancer (indolent or aggressive) a patient has is not directly observable and it is unreasonable to not treat patients in an attempt to determine which have less aggressive varieties. We propose that training a classifier to identify in situ cancers with features specific to older patients, and thus less aggressive varieties of cancer, is also analogous to identifying *Persuadables*. By maximizing the in situ cases' uplift, we are identifying the older in situ cases that are most different from younger in situ cases, and thus are the best candidates for watchful waiting.

(a) In Situ (b) Invasive

Figure 4.4: Examples of in situ and invasive cancer tissue.

The dataset we use for this task comes from the University of California San Francisco Medical Center. It consists of two cohorts: patients younger

than 50 years old form the younger cohort, while patients aged 65 and above form the older cohort. The older cohort has 132 in situ and 401 invasive cases, while the younger one has 110 in situ and 264 invasive (see Table 4.3). Each case consists of 20 features that describe the mammogram, and 35 relational features that connect a mammogram with related mammograms, discovered at the same or in prior visits. This is the same dataset used by Nassif et al. (2010) and Nassif et al. (2012a).

Table 4.3: Composition of the breast cancer dataset from UCSF. There are 55 features in the dataset. The older cohort has 132 in situ cases and 401 invasive, while the younger cohort has 110 in situ cases and 264 invasive.

	Older		Younger	
Features	In Situ	Invasive	In Situ	Invasive
55	132	401	110	264

4.4 Simulated Marketing Activity

Recall that the goal of uplift modeling is to be able to predict when a person is more likely to buy a product after having been targeted by marketing activity versus having not been targeted. Maximizing the uplift curve makes some intuitive sense, but the fact that customer groups cannot be directly observed makes it difficult to understand if it really does help to produce classifiers that can specifically identify *Persuadables*. To confirm that maximizing uplift identifies *Persuadables*, we generated a synthetic population of customers and simulated marketing activity to produce a dataset [3] for which we knew the ground truth customer groups. We present results on this synthetic dataset using algorithms presented in Chapter 8.

[3] Available at: `http://ftp.cs.wisc.edu/machine-learning/shavlik-group/kuusisto.ecml14.simcustomerdata.zip`

To generate the customer population, we first generated a random Bayesian network with 20 nodes and 30 edges (see Algorithm 4.1). We then randomly selected one node with four possible values to be the customer group feature. Next, we drew 10,000 samples from this network. This left us with a population of customers for which one feature defined the group they belonged to and the rest represented observable features.

Algorithm 4.1 Marketing Campaign Simulation

```
BN ← GenBayesNet();                          ▷ Random Bayesian network
MarkCustomerNode(BN);  ▷ Select (four-value) customer type node
Pop ← SampleCustomers(BN);       ▷ Sample a customer population
for C ∈ Pop do
    if RandomTarget(C) then                  ▷ Choose to target or not
        MarkTargetResponse(C);
    else
        MarkControlResponse(C);
    end if
end for
```

We then subjected this population to a simulated marketing activity. We randomly selected roughly 50% of the entire population to be part of the target subgroup. Next, we produced a response for each customer based on their customer group and whether or not they were chosen to be targeted. For this demonstration, we determined each response based on the strongest stereotypical interpretation of each customer group. That is, *Persuadables* always responded when targeted and never responded when not. *Sleeping Dogs* never responded when targeted and always responded when not. *Sure Things* and *Lost Causes* always and never responded respectively.

Table 4.4: Composition of synthetic customer population after simulated marketing activity. There are 20 features for each customer, including the hidden customer type.

	Target		Control	
	Response	No Response	Response	No Response
Persuadable	1,219	0	0	1,252
Sure Thing	1,221	0	1,226	0
Lost Cause	0	1,256	0	1,298
Sleeping Dog	0	1,241	1,287	0
Total	2,440	2,497	2,513	2,550

4.5 Statins and Myocardial Infarction (Synthetic)

Cardiovascular diseases (CVD) are diseases that affect the heart or blood vessels, and are the leading cause of death globally (Mendis et al., 2011). Myocardial infarction (MI), or heart attack, is one such disease and affects one million people each year in the United States (NIH, 2013). High cholesterol levels have been associated with CVD, and evidence suggests that statins, drugs used to lower cholesterol, are effective for treatment of early stage CVD and for those with elevated risk of CVD (Taylor et al., 2013). Nevertheless, statins still carry risk of side-effects (Naci et al., 2013), and being able to identify individual responses to treatment would be valuable. We desire to build machine learning models that can estimate these individual responses and discuss how they compare with more traditional methods in Chapter 9. In order to do so, however, we need to know ground truth effects, which are not available in real data, so we rely on a synthetic model instead.

We define a synthetic model of MI with thirteen binary variables: age, gender, smoking status, HDL level, LDL level, diabetes, family history of cardiovascular disease (CVD), blood pressure, history of angina, history

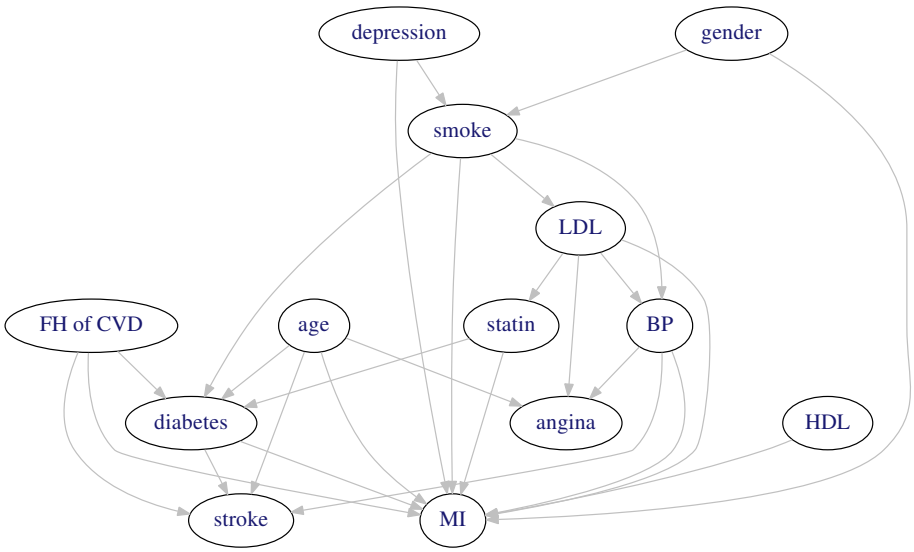


Figure 4.5: Causal Bayesian network for myocardial infarction (MI) and related variables used for synthetic data in our experiments.

Table 4.5: Marginals for each variable in the synthetic model for observational data (Obs.) and data from the RCT version randomizing on statin use. Reported values are the probability of a "yes."

Variable	Obs.	RCT	Variable	Obs.	RCT
Age (older)	55	55	Blood pressure	30	30
Smoke	28	28	History of angina	35	35
Gender (male)	48	48	History of stroke	5	6
HDL	27	27	History of depression	27	27
LDL	39	39	Statin use	25	50
Diabetes	39	42	MI	8	9
Family history of CVD	27	27			

of stroke, history of depression, statin use, and MI. The joint probability distribution is defined by the causal Bayesian network in Figure 4.5 with hand-crafted conditional probability distributions for each variable informed by medical expertise in Table 4.5. This model was defined and used in Weiss (2014).

In addition to needing the ground truth of individualized treatment effects, we desire to simulate both an observational study and a randomized controlled trial (RCT). This synthetic model allows us to do so. For observational data, we sample directly from this Bayesian network, and interventions can be simulated by removing incoming edges to the intervention variable and specifying the Bernoulli distribution parameter. We simulate data from a randomized controlled trial of statin use by removing the edge from LDL to statin and using a conditional probability distribution for statins with equal probability of "yes" and "no".

4.6 D-Penicillamine for Primary Biliary Cirrhosis

Primary biliary cirrhosis (PBC) is an autoimmune disease of the liver and is characterized by progressive destruction of small bile ducts of the liver (Hirschfield and Gershwin, 2013). D-Penicillamine was once recommended for treatment of PBC for its copper-chelating and immunological effects (Epstein et al., 1981). This is no longer the case (Gong et al., 2004), but we desired to validate our claims in Chapter 9 on a real dataset. Fortunately, there is readily available trial data for the treatment of PBC from the Mayo Clinic (Therneau and Grambsch, 2000). The trial covers a ten-year period and randomized patients across treatment with a placebo versus treatment with D-penicillamine. The data set includes 16 variables, including demographic information like age and sex, as well as various lab tests such as serum albumin, serum cholesterol, and triglycerides (see Table 4.6).

For this RCT dataset we wish to understand the effect of D-penicillamine use on three-year survival. For the three-year survival period, we censored the dataset to 288 patients, with 146 in the treatment group and 142 in the placebo group. At the end of three years, the treatment group experienced

Table 4.6: Statistics for the primary biliary cirrhosis (PBC) dataset censored to a three-year survival period. The top half of the table gives counts for the boolean and categorical features, and the bottom half gives statistics for the numerical features.

Variable	Count (n=288)	%
Sex female	255	88.5
Ascites	24	8.3
Hepatomegaly	148	51.4
Spider angiomas	84	29.2
Edema		
none	241	83.7
untreated or successful	27	9.4
present despite therapy	20	6.9
Histologic stage		
1	16	5.6
2	60	20.8
3	112	38.9
4	100	34.7
Variable	**Mean**	**Std. Dev.**
Age in years	50.6	10.5
Serum bilirunbin (mg/dL)	3.3	4.7
Serum cholesterol (mg/dL)	364.4	224.8
Serum albumin (g/dL)	3.5	0.4
Urine copper (µg/day)	96.1	84.9
Alkaline phosphate (U/L)	2021.1	2213.2
Aspartate aminotransferase (U/L)	122.2	58.1
Triglycerides (mg/dL)	124.4	65.4
Platelet count	259.1	96.4
Prothrombin time	10.8	1

Table 4.7: Three-year survival for the PBC dataset.

Treatment		Control	
Survived	Died	Survived	Died
119	27	110	32

27 (18.5%) deaths out of 146, whereas the placebo group experienced 32

(22.5%) out of 142. The average treatment effect then is a 4 percentage point reduction in death over three years.

5 HIGH-PRECISION RULES FOR NON-DEFINITIVE BREAST BIOPSY

We first present some of our work on the upgrade prediction task (see Section 4.1). Here we attempt to reduce the number of benign cases that go on to excisional biopsy using learned rules. Recall that learned rules have the advantage of being interpretable (see Section 2.5), though this can come at the cost of accuracy because predicted outcomes are strictly binary if-then statements. That is, the inferred rules do not allow for finer granularity in decision making. To account for this challenge, we here modify rule scoring to enforce careful selection of rules that meet the clinical objective (Kuusisto et al., 2013).

5.1 Introduction

Recall from Section 4.1 that when a screening mammogram presents a suspicious finding, a follow-up diagnostic mammogram is performed to further define the abnormality. If the finding remains suspicious, a core needle biopsy (CNB) may be recommended (Bevers et al., 2009). Pathologic review of biopsy is often definitive, but some are not, and surgical excision biopsy is recommended to determine the final pathology. A majority of these women who go on to surgery will receive a benign diagnosis.

In the mid-1990s, the American College of Radiology developed the mammography lexicon, Breast Imaging Reporting and Data System (BI-RADS), to standardize mammogram feature distinctions and the terminology used to describe them (BIR, 2003). Studies show that BI-RADS descriptors are predictive of malignancy (Liberman et al., 1998; Moskowitz, 1983; Swets et al., 1991), specific histology (Burnside et al., 2004; Nassif et al., 2010), and prognostic significance (Thurfjell et al., 2002; Tabar et al., 2004; Nakayama et al., 2004). For many reasons then, breast cancer diag-

nosis is an ideal domain to develop and test machine learning methods for risk prediction because 1) a standardized lexicon with probabilistic underpinnings has been established to summarize imaging features, 2) predictive risk factors are available within the standardized lexicon, and 3) accurate outcomes exist through cancer registries, collections of history, treatment, and diagnosis data on cancer patients.

In this study, we investigate using machine learning to examine the ability to predict benign entities in cases where CNB has produced a non-definitive diagnosis. Prior work in machine learning has demonstrated potential in predicting upgrade (Dutra et al., 2011) and predicting breast cancer in general using imaging features (Burnside et al., 2009). These prior works, however, have not clearly leveraged the potential of rich, multi-relational data from many sources. Our study considers demographic risk factors and mammographic features, not just biopsy and pathology characteristics, to estimate the risk of upgrade. These factors and features are organized in multiple tables, which makes the dataset suitable for relational learning (De Raedt, 2008). Additionally, the prior work has focused on predicting malignancy and has not focused on producing interpretable models that help to understand what makes these suspicious cases benign. We generate interpretable classifiers, based on first-order logic, that capture the correlation between features included in this study to predict when a patient should *not* undergo excision.

5.2 Background

In the past, our group developed a Bayesian network based on the ability of BI-RADS descriptors to convey the level of suspicion of mammographic abnormalities on a dataset from a different practice than we analyze in this work. Previously we found an upgrade rate of 1.1% (1 in 92 biopsies). Our expert system was able to integrate pathologic diagnoses and mam-

mographic findings to obtain the probability of upgrade, thereby enabling the identification of malignancy with 100% sensitivity while maintaining a specificity of 91% (Burnside et al., 2004).

Dutra et al. (2011) demonstrated the potential for improving performance of predicting upgrade cases when expert knowledge is provided to a machine learner prior to training. In one experiment, their results showed that they were able to correctly identify at most 60% (9/15) of their malignant cases, while saving 43% (34/79) of the benign cases from excision. In another experiment, they were able to correctly identify 53.3% (8/15) of their malignant cases while saving 83.5% (66/79) of the benign cases from excision. This is a remarkable result for the benign cases, as it demonstrates a substantial reduction in false positives, but it comes at the cost of missing half of the malignancies.

Current practice standards at our institution include a conference between breast radiologists and pathologists about every core needle biopsy performed to assess whether the biopsy is perceived to be non-definitive. Factors influencing decision-making include imaging characteristics of the original lesion, operational factors such as gauge of the needle used and the number of the samples taken, and clinical characteristics of the patient. However, given that these factors have been imprecise in accurately predicting which patients may have an associated malignancy (Fures et al., 2003; Destounis et al., 2011; Gumus et al., 2012; Rauch et al., 2012), surgical excision is performed for most patients with non-definitive results to ensure that no malignancy is missed.

5.3 Methods

Institutional review board approval was obtained prior to the commencement of this retrospective study. Written informed consent of patients was not required. We used the dataset described in Section 4.1, which includes

a population of patients that underwent 1,414 consecutive CNB, as a result of a diagnostic mammogram, from January 1, 2006 through December 31, 2009. Of these biopsies, 96 were prospectively given a non-definitive diagnosis after discussions in clinical conference meetings (see Figure 4.2). We limited our dataset to this subset. For all 96 cases, we collected information related to the pathological diagnoses, technical biopsy procedure and materials, as well as patient history, information about previous mammograms, and BI-RADS descriptors associated with the biopsied tissue. All 96 cases were women, and all underwent excision. Their mean age was 56 years (range$= 33 - 85$ years, sd$= 11.23$). We use the result of excisional biopsy (within 6 months after CNB) or a registry match (within 1 year after CNB) as a reference standard for final diagnosis. Of our 96 cases, 79 were ultimately confirmed to be benign while 17 of them (18%) were found to be malignant.

We use the inductive logic programming (ILP) system, Aleph (Srinivasan, 2007), to predict when a patient should *not* undergo excision. ILP is a machine learning approach that learns a set of rules in first-order logic that explain a given dataset (Lavrac and Dzeroski, 1994). We use ILP because it is well suited for our multi-relational dataset and because the logical rules produced can be easily interpreted by a human. We chose to make benign cases our "positive" class because we wish to find highly accurate rules that predict when this procedure is not needed. Unlike most machine learning approaches, ILP treats its positive and negative training asymmetrically, focusing on inducing rules that match many positive examples and few (ideally zero) negative examples. Readers should be aware of this wording ("positive" is benign), as it is somewhat counter-intuitive, but it is a choice motivated by the machine learning approach we employ.

We considered a small number of training parameters. We did not tune these parameters but instead selected what we and our clinical collaborators considered reasonable values to help achieve our clinical objective of

identifying benign cases without missing malignancies. Among the many parameters Aleph offers, we specified:

minpos The minimum number of positive examples that a rule is required to cover.

We chose a value of 2 for the minpos parameter, allowing any rules that correctly identify at least two benign cases in the training set. We chose 2 instead of 1 to require rules that generalize beyond a single case at minimum, while not assuming anything about how much further Aleph would be able to generalize.

noise The maximum number of negative examples that a rule is allowed to cover.

We chose a value of 0 for the noise parameter, disallowing any rules that incorrectly identify even a single malignant case as benign in the training set. We chose 0 due to the high cost of missing cancer (Petticrew et al., 2001).

evalfn The rule-cost evaluation function.

We chose to use an F_β measure (Manning et al., 2008) for the rule evaluation function because it allows us to balance the importance of true positives (TP), false positives (FP), and false negatives (FN).

$$F_\beta = \frac{(1+\beta^2) \times TP}{(1+\beta^2) \times TP + \beta^2 \times FN + FP}$$

We chose a value of 0.1 for β, effectively making precision 10 times as important as recall. This is again because, while we would certainly like to identify more benign cases, it is more important in the task we are addressing that we avoid calling malignant cases benign.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

ILP generates a theory that may consist of many different rules, where each rule is a conjunction of features that together predict the chosen positive class (i.e. benign in this task). To reduce overfitting on such a small dataset, we prune the output theory to a single rule. Our pruning process selects the rule with the best F_β score (as described above) on the training set. By pruning this way, we hope to reduce each theory to its best performing rule (see Algorithm 5.1). We ran all of our experiments using the YAP Prolog compiler (Santos Costa et al., 2012) and Condor (Thain et al., 2005), a high-throughput computing system.

To evaluate our learned rules, we use stratified 17-fold cross-validation, each fold including a single malignant case in its test set. We chose 17 folds instead of the more common ten because we only have 17 malignant examples, and choosing ten would lead to unbalanced folds. Cross-validation ensures that cases that were used to learn a rule are not used to evaluate the rule. Biopsies of the same patient were all placed in the same fold. In addition to the cross-validation results, we present all of the unique rules learned along with their individual performance on the full dataset. This allows us to demonstrate the interpretable nature of the rules, reason about their clinical significance, and discuss the degree to which they could affect patients in practice.

Algorithm 5.1 Rule Learning Procedure

for $\mathtt{Train}, \mathtt{Test} \in \mathtt{Folds}$ **do**
  $\mathtt{Theory} \leftarrow \mathtt{Aleph}(\mathtt{Train}, \mathtt{minpos} = 2, \mathtt{noise} = 0, \mathtt{evalfn} = F_\beta)$; ▷ Induce theory
  $\mathtt{Rule}^* \leftarrow \text{argmax } F_\beta(\mathtt{Theory}, \mathtt{Train})$; ▷ Select best rule
  $\mathtt{Evaluate}(\mathtt{Rule}^*, \mathtt{Test})$; ▷ Evaluate selected rule on test set
end for

5.4 Results

We first present the overall results from 17-fold cross-validation in Table 5.1. Recall that we are learning rules to predict benign cases, so true positives are cases that are correctly identified as benign, and false positives are malignant cases that are incorrectly identified as benign. Similarly, true negatives are the malignant cases that are correctly identified as malignant, and false negatives are the benign cases that were incorrectly identified as malignant. We also report precision and recall, otherwise known as positive predictive value and sensitivity respectively. Each row shows the results, for a single fold, on the examples that were held out from training in that fold. We compute summary statistics as suggested by Forman and Scholz (2010).

Each of the 17 folds produced a single theory that was then pruned to a single rule. In many of the folds, the rule produced was identical to that of another fold. What follows are the five unique rules that were produced amongst all the folds, sorted by the number of folds that produced them. We have manually translated them from first-order logic to English to make them easier to read (see Table 5.2). The performance of each unique rule on the full dataset can be found in Table 5.3, along with the number of folds in which each rule was learned.

Note that the third and fourth rules, learned in 1 fold each, are generalizations of Rule 2, each including only one of the mass margin descriptors along with the non-disappearance of the abnormality. These two rules are less precise, and are the source of the 2 false positives in our cross-validation results above.

5.5 Discussion

In this experiment, we demonstrate that ILP can derive rules that accurately predict when a woman may not require excision after a non-definitive core

Table 5.1: 17-Fold Cross Validation Results

Fold	TP	FP	FN	TN	Precision	Recall	$F_{0.1}$
1	2	0	3	1	1.0	0.40	0.99
2	1	0	3	1	1.0	0.25	0.97
3	4	1	1	0	0.8	0.80	0.80
4	2	0	3	1	1.0	0.40	0.99
5	1	0	3	1	1.0	0.25	0.97
6	4	0	1	1	1.0	0.80	1.00
7	0	0	4	1	0.0	0.00	0.00
8	2	0	3	1	1.0	0.40	0.99
9	4	0	1	1	1.0	0.80	1.00
10	2	0	3	1	1.0	0.40	0.99
11	1	0	4	1	1.0	0.20	0.96
12	1	1	4	0	0.5	0.20	0.49
13	0	0	5	1	0.0	0.00	0.00
14	0	0	5	1	0.0	0.00	0.00
15	0	0	4	1	0.0	0.00	0.00
16	0	0	4	1	0.0	0.00	0.00
17	1	0	3	1	1.0	0.25	0.97
Summary	**25**	**2**	**54**	**15**	**0.93**	**0.32**	**0.91**

breast biopsy. All five of the rules predict a substantial number of cases that are benign, and only two of the rules miss a malignancy. Multiple rules contain both imaging and clinical factors, with features included falling into three main categories: post-biopsy imaging (a standard part of the CNB process), mass margin descriptors, and patient history. All of the included features also have some clinically significant explanation as confirmed by our multidisciplinary (radiology, pathology, and surgery) team. Overall, the cross-validation results indicate that we can potentially reduce the total number of patients with non-definitive diagnosis from undergoing excision by around 28%, with confidence that 93% of those patients do not have a malignancy.

Table 5.2: The five unique learned rules that predict a non-definitive case is benign.

1 The patient did not have a previous surgery,
imaging did not present a spiculated mass margin,
and the abnormality did not disappear in post-biopsy imaging
2 Imaging did not present an indistinct mass margin,
imaging did not present a spiculated mass margin,
and the abnormality did not disappear in post-biopsy imaging
3 Imaging did not present a spiculated mass margin,
and the abnormality did not disappear in post-biopsy imaging
4 Imaging did not present an indistinct mass margin,
and the abnormality did not disappear in post-biopsy imaging
5 The patient has no personal history of breast cancer,
and the abnormality did not disappear in post-biopsy imaging

Table 5.3: Individual Rule Performance on Full Dataset (# Folds is the number of folds in which a rule was learned)

Rule	# Folds	**TP**	FP	FN	TN	Precision	Recall	$F_{0.1}$
1	10	**30**	0	49	17	1.00	0.38	0.98
2	4	**29**	0	50	17	1.00	0.37	0.98
3	1	**34**	1	45	16	0.97	0.43	0.96
4	1	**31**	1	48	16	0.97	0.39	0.95
5	1	**28**	0	51	17	1.00	0.35	0.98

When we look at the specific rules generated, additional interesting observations can be made. Importantly, the two rules that missed single malignant cases were each only learned in a single fold, whereas the strongest rule that misses no malignancies (Rule 1) was learned in ten (of 17) different folds. Similarly, the second strongest rule that misses no malignancies (Rule 2) was learned in four different folds. This lends support to the idea that these two rules capture a significant signal across the entire dataset. When choosing rules to implement clinically, clinicians

would undoubtedly prefer rules that do not miss a cancer. Our results may indicate that the combination of fold coverage and clinical judgement may be a criteria on which to select the most advantageous rules. In our project, this approach designates the first two rules as the most useful. Whether these rules will be generalizable to new data remains future work.

When considering features that relate directly to imaging, some predicates stand out. First, all of the rules require that the abnormality does *not* disappear in post-biopsy imaging. This is, however, a counter-intuitive requirement. From a clinical perspective, it makes greater intuitive sense to consider a case likely to be benign if the abnormality disappears, because it suggests that the entire abnormality was removed by the needle biopsy, and therefore adequately sampled. However, we believe that this can be explained by how patients with non-definitive biopsies are identified. All of our patients underwent a core needle biopsy to provide a definitive diagnosis of an abnormality identified in imaging. In those patients where the abnormality did not disappear on post-biopsy imaging, clinicians may be more concerned that the biopsy did not sample the correct tissue and, therefore, be more likely to call it non-definitive. This sampling error can then explain why it was included in our dataset in the first place, whereas the biopsy may have been deemed definitively benign if the tissue had been sampled sufficiently.

Mass margin, another predictive imaging descriptor, is included in Rules 1 through 4. Specifically, the rules indicate that a case is benign if imaging does not present an indistinct mass margin, or if imaging does not present a spiculated mass margin. Mass margin has been shown to be an important predictor of malignancy in prior literature (Liberman et al., 1998). Rule 2, in particular, suggests that a case is benign if imaging presents neither an indistinct or spiculated mass margin. Rules 3 and 4 suggest that a case is benign if either is not present, but this proves to be less precise. Both rules are strong predictors, but they each misclassify

one malignant case as benign.

Descriptors of patient history are included in Rules 1 and 5. In fact, the clinical literature indicates that patients with a personal history of breast cancer are at greater risk (Kurian et al., 2009; Bouchardy et al., 2011). Rule 5 explicitly suggests that a case is benign if the patient has no personal history of breast cancer, but Rule 1 suggests that a case is benign if the patient has not had prior breast surgery. Patients undergo breast surgery not only for cancer but for definitive diagnosis of lesions considered to be non-definitive (as discussed in this paper). Some entities for which patients may undergo surgery include high risk lesions such as atypical ductal hyperplasia or lobular carcinoma in situ. Although not cancers, these lesions increase an individual's personal risk of developing cancer in the future. We posit that history of prior surgery may represent a surrogate for these high risk lesions, though this must be verified in future work.

The choice to use ILP for learning was valuable not only because it allows us to leverage our mult-relational data, but also because the learned rules are interpretable in natural language. This means that clinicians may be able to consider them in practice easily and immediately to assist in decision-making when faced with a non-definitive diagnosis. The high precision at which the rules operate may help to reduce concerns clinicians may have about missing a malignancy. Interpretability also gives clinicians the ability to reason about the clinical significance of the features used in a rule, or even discover new and interesting combinations of features. Combining physician-generated rules with machine learned rules has been explored in previous work (Dutra et al., 2011) and we hope to extend this promising direction of research using both multi-relational data and a multidisciplinary physician team (rather than a single physician in one domain).

Despite a small dataset, our approach was able to infer highly accurate rules. We note that, while several of the rules are derived from imaging

features, the pathology features are poorly utilized. This is likely because, in our database, the imaging features are well populated and standardized using BI-RADS, but most of our pathology results are stored in free text. This suggests that an important goal is to improve our data collection process, which may be reflected in an increased use of pathology features in future work.

6 ADVICE-BASED LEARNING FRAMEWORK

Learning from the challenges of working with small training sets, we next present work on developing a process to best leverage clinical expertise to improve model performance (Kuusisto et al., 2015). In this work on breast cancer, we still focus on the upgrade prediction task and demonstrate how our process improves upon our previous work.

6.1 Introduction

Collaborations between medical domain experts (MDE) and computer science experts (CSE) often involve the use of machine learning to develop predictive models aimed at improving patient care. Unfortunately, standardized, complete, and sufficient training data for machine-learning algorithms is rarely available for a variety of reasons including variability of practice between physicians as well as institutions, low disease prevalence on a population level, and confidentiality issues (Cohen et al., 2014). The difficulty inherent in collecting large, high quality datasets represents a major challenge in the development of machine learned models for decision support. One of the solutions to this challenge is to incorporate the clinical experience and intuition of MDEs, that may help compensate for a lack of large training datasets (Mitchell, 1997). In fact, some successful cases of integrating expert knowledge with predictive and analytical models are available in the literature (Gibert et al., 2010; Velikova et al., 2013). As it is nearly impossible for MDEs, who are not programmers, to contribute their expertise directly to the software, we argue that there is a need for a framework that improves close collaboration between MDEs and CSEs to provide a method for shared dialog. Rather than solely providing training through a set of examples, it would be much more valuable if the MDEs could (a) explain what the machine learner is doing wrong

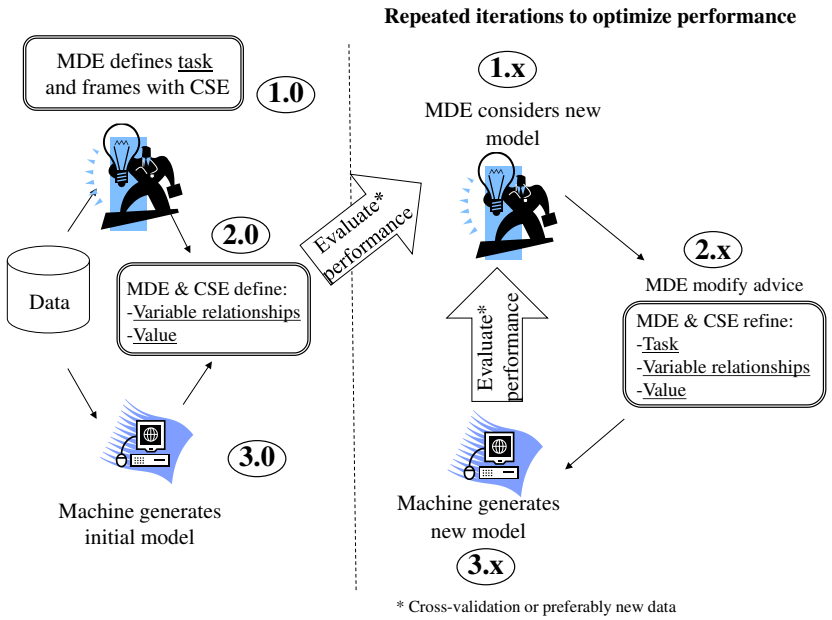


Figure 6.1: The ABLe Framework. To the left of the dotted line is the first phase of training where medical domain experts (MDE) and computer science experts (CSE) collaborate to produce an initial model. The second phase, to the right, is an iterative process in which the task, variable relationships, and parameters are refined until the model meets the clinical objective.

and (b) explain how to fix the current problem in a manner that will generalize to similar future cases. This dialog is the basic idea motivating our development of Advice-Based-Learning (ABLe). In ABLe, MDEs provide advice, and the learning algorithm is able to decide how best to absorb it, possibly rejecting the advice or refining it based on the available data. Based on continual observation of model performance, the MDEs can provide additional advice.

6.2 The ABLe Framework

Our ABLe framework (Figure 6.1) includes: (1) definitions and (2) iterative steps. The definitions are:

Task The problem and scope with quantification of appropriate predictive variables.

Variable Relationships Combinations of variables that are particularly important for the task.

Parameter Values Algorithm settings or parameters that best represent the clinical objective.

In the process of developing a decision support system, definitions will be unique to the specific clinical goal. Modeling techniques should be chosen based on both the data and the task, but this framework provides a way by which the MDEs and CSEs can interact.

Regarding the iterative steps, we follow a similar process to Gibert et al. (2010). In Steps 1 through 3, the MDEs and CSEs interact to establish an initial model. In Step 1, the MDEs (physicians in our example) define a task to address, provide the data, determine what variables will be used from the data available, and determine what is the desired outcome. At this point, the CSEs are involved in picking an appropriate machine learning algorithm based on the task, data, and the needs of the MDE.

In Step 2, the physicians and computer scientists interact to produce an initial set of variable relationships and value specifications. The variable relationships correspond with clinician intuition about predicting the chosen task based on relevant knowledge (e.g. the literature) and available data. This advice is encoded in a way that allows it to be incorporated directly into the chosen algorithm. There are multiple ways for prior knowledge to be incorporated into learning algorithms (Simard et al.,

1992; Towell and Shavlik, 1994; Lucas, 2001), and the method used for each application will depend on decisions made in Step 1. For example, physicians may choose to provide the structure for a Bayesian network based on their expert knowledge of the variable dependencies. The value specifications correspond to proper selection of algorithm parameters and other experimental settings in order to obtain clinical significance. For example, the physicians can help the computer scientists specify costs of misclassification or a weighting scheme for importance of examples.

Finally, in Step 3 the initial model is trained and produces results on an unseen set of data. Ideally, the unseen data will be truly new data, but this can also mean methods such as cross-validation or bootstrap sampling. The evaluation of the model will depend on the specific task being addressed. For example, in the upgrade prediction task, the proportion of malignant cases that are misclassified as benign is the most important factor in deciding if the model meets the clinical objective. The results must include such statistics.

Steps 1.x through 3.x show the iterative refinement process that occurs after initial model production. In Step 1.x, the MDEs consider the results produced by the model. Another interaction between MDEs and CSEs takes place, and previous definitions of the task, variable relationships, or value specifications are then modified in Step 2.x. For example, if the current task definition proves to be too challenging, the MDEs can redefine the task in the most fruitful direction. Similarly, the MDEs can modify, remove, or provide additional variable relationships or value specifications. Finally, in Step 3.x, a new model is trained and produces results on unseen data, which again leads to Step 1.x+1. This process continues until the MDEs are satisfied with the results produced by the model learned.

We next present the use of ABLe on an example task, predicting undiagnosed malignancy in the setting of a benign but non-definitive image-guided breast core biopsy. In the next two sections, we briefly review the

task and explain how we used ABLe to improve our model performance. Although we apply it to a specific example, the framework works for a wide range of medical tasks.

6.3 Task Prediction

For application of this framework, we focus on the upgrade prediction task described in Section 4.1. Recall, again, that the clinical objective in this task is to reduce the number of benign lesions that are excised as a result of a non-definitive core needle biopsy (CNB). As we determined in previous work, perhaps the greatest challenge in this particular task is the relative rarity of the event. Though the number of women affected by non-definitive CNB across the US is substantial, the availability of data to any particular institution is relatively limited, thus making the development of decision support systems more difficult.

6.4 Application of ABLe to Upgrade Prediction

In our first meeting to establish a model, the MDEs defined the task (Step 1.0) as predicting upgrade using a dataset of 157 biopsies that were prospectively given a non-definitive diagnosis at radiologic-histologic correlation conference. To incorporate physician advice about relationships between variables (Step 2.0), we opted to use a logic-based language. Our variables consist of imaging findings, demographic information, and some pathology findings. Physicians and computer scientists hand-coded expert rules expressing combinations of variables that increase or decrease risk of upgrade according to physician experience and the literature.

1. Risk of upgrade decreases if imaging features are "typically benign."
2. Risk of upgrade decreases if BI-RADS category is low (3 or 4A).

3. Risk of upgrade decreases for atypical/radial scar if the imaging finding is explained by another pathology.
4. Risk of upgrade increases if imaging features are "high probability of malignancy."
5. Risk of upgrade increases if BI-RADS category is high (4C or 5).
6. Risk of upgrade increases in dense breasts on mammography (heterogeneously or extremely dense).
7. Risk of upgrade increases if complexity/density of breast tissue is high (extremely dense).

To incorporate the rules into the model, we took a similar approach to Dutra et al. (2011), treating them as binary features in our dataset. For each example, every rule became a feature that was given a value of true if the rule applied to that example, otherwise it was given a value of false. We chose to use the Naïve Bayes algorithm primarily for its simplicity, making it more approachable for clinicians, and for its history of good performance, despite strong independence assumptions (Zhang, 2004). We also considered logistic regression as an alternative model because it is generally accepted by the clinical audience, an important detail if a model is to be put into practice. As shown by Ng and Jordan (2002) though, when working with smaller datasets, generative models (e.g. Naïve Bayes) may be preferable to discriminative models (e.g. logistic regression) because they reach their asymptotic error faster.

The initial model (Step 3.0) demonstrated no substantial ability to identify benign non-definitive cases without misclassifying malignant cases. Our MDEs considered the results produced by this initial model (Step 1.1) and formed a hypothesis about what the primary challenge was. Specifically, they surmised that the non-definitive population as a whole was too challenging to be addressable with the predictive features

available in a typical clinical dataset, and that targeting a subpopulation would be most fruitful.

Non-definitive biopsies can be broken into three subtypes (discordant, insufficient, and atypical/radial scar), and each subtype has distinct features that are likely to predict upgrade. *Discordance* means that the histologic findings do not provide an acceptable explanation for the imaging features and indicates that the targeted tissue may not have been sampled adequately. In this situation, the imaging features are likely to provide the most influential variables in predicting upgrade. The other two subtypes focus more on the histologic factors (cellular features) that raise the possibility that abnormal tissue remains in the region of the biopsy, in which case, pathology features should be more influential in predicting upgrade.

The dataset available to us for the task contains a limited set of pathology features, but we use structured reporting in our imaging practice and adhere to the BI-RADS lexicon, so our experts identified an opportunity to employ machine learning methods that capitalize on the imaging features. Thus, we chose to alter the task (Step 2.1) and focus on estimating the probability of malignancy for discordant cases specifically. Due to the alteration of the task, our physician experts also reduced the initial set of variable relationship rules (Step 2.1) to a set of four specifically related to predicting discordance (Rules 1, 2, 4, and 5 above). This also led us to begin collecting a larger set of features in our medical practice for the sake of future work on the other subpopulations. We again tested the model with cross-validation (Step 3.1).

Evaluation of the next model (Step 1.2) demonstrated a marked improvement over the initial model (see Table 6.1). When trained with the combined base feature set and the binary advice rules, the model showed improvement over models trained on either feature set alone, though at the cost of missing malignancies. We thus considered value specifications that would help the model better address the clinical objective (Step 2.2).

This led us to specify a highly skewed cost-ratio for false negatives versus false positives. We selected a skew of 50:1 based on the 50 benign cases in the discordant set to suggest that the algorithm prefer to misclassify every benign case before misclassifying a single malignancy (Step 2.2). We again tested the model with cross-validation (Step 3.2).

Review of the new model (Step 1.3) demonstrated an improvement over the previous model, but still misclassified a single malignant case. Upon inspection, this single misclassification was shown to be the result of incorrectly entered features in our reporting software. These kinds of errors should be expected in real-world data, which led us to reconsider our skewed cost-ratio (Step 2.3). Our MDEs suggested an even more conservative cost-ratio of 150:1 based not just on counts in the dataset, but on recent work on utility analysis in mammography (Abbey et al., 2013) (Step 3.3). This led us to our final model, which demonstrated our best performance.

6.5 Methods

We used the dataset described in Section 4.1, which includes a population of patients that underwent 1,910 consecutive CNB, as a result of a diagnostic mammogram, from January 1, 2006 to December 31, 2011. Clinicians prospectively gave a total of 157 biopsies a non-definitive diagnosis at radiologic-histologic correlation conference, and 60 of these were categorized as discordant. Recall that we have chosen to focus on the discordant cases. The mean age of these patients was 55.2 years (range$= 25 - 83$ years, sd$= 12.2$), all 60 cases were women, and all underwent excision. As a reference standard for final diagnosis, we use the result of excisional biopsy (within 6 months after CNB) or a registry match (within 1 year after CNB). Fifty were confirmed to be benign while 10 (16.7%) were found to be malignant. A diagram of our case inclusion process can be seen in

Figure 6.2, with the original numbers from the base dataset on the left and the discordant subset on the right.

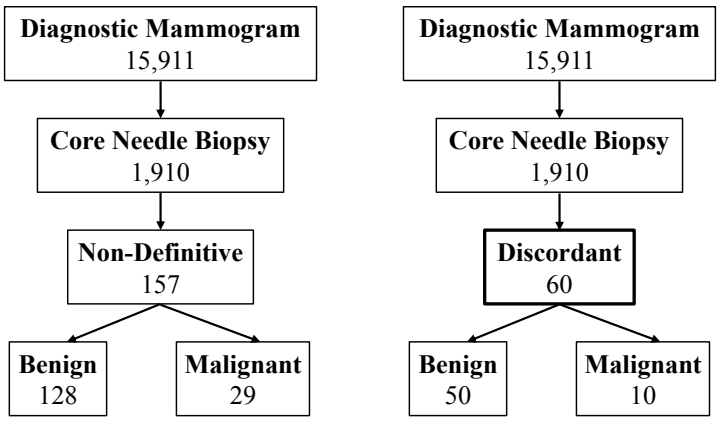


Figure 6.2: Case inclusion diagrams for the entire non-definitive set (left), and for our subset of interest, the discordant cases (right).

Radiologists described and recorded all mammographic findings using BI-RADS terms by the interpreting radiologist at the time of mammography interpretation using structured reporting software (PenRad®, Minnetonka, MN), which is routinely used in the University of Wisconsin clinical practice. We derive mammography features and demographic risk factors from the diagnostic mammogram that precedes the biopsy and has an abnormal BI-RADS assessment category.

We use 10-fold stratified cross-validation for evaluation, and the Weka (Hall et al., 2009) software package (version 3.7) to train a Naïve Bayes (NB) model for each fold. Note that NB is known to often accurately predict the most probable label even though its predicted probabilities are not well calibrated (Zadrozny and Elkan, 2001). We show results with a FN:FP cost-ratio of 1:1 and a suggested cost ratio of 150:1 drawn from the literature (Abbey et al., 2013). We also show results when the expert advice rules are either included or excluded from the dataset at training time to better assess the importance of the variable relationship advice.

Table 6.1: 10-fold cross-validated performance of Naïve Bayes classifiers with FN:FP cost-ratio of 1:1 and our final model with cost-ratio 150:1 at 2% threshold of excision.

		FN:FP cost-ratio 1:1			FN:FP cost-ratio 150:1		
Parameter	Baseline	Data	Rules	Both	Data	Rules	Both
Biopsy	60	28	42	30	55	55	48
No Biopsy	0	32	18	30	5	5	12
Malignant Excisions	10	7	9	7	10	10	10
Benign Excisions	50	21	33	23	45	45	38
PPV (%)	16.7	25.0	21.4	23.3	18.2	18.2	20.8
Specificity (%)	0.0	58.0	34.0	54.0	10.0	10.0	24.0
Spec. $>$ 0.0 p-Value	-	0.000$*$	0.003$*$	0.000$*$	0.026	0.026	0.004$*$

We compare our trained models to the baseline of current standard practice. For comparison, we consider a 2% threshold of predicted malignancy that would hypothetically be used to decide if the patient should go on to excisional biopsy. This threshold has been used previously and is clinically reasonable (Burnside et al., 2009). The baseline current practice is to excise all discordant cases, so there is no distinction of treatment at any threshold. For each model, we compare the number of malignant cases that would be excised, the number of benign cases that would be excised, the positive predictive value (PPV) of malignancy, and specificity. We use a one-sided, one-sample t-test at the 99% confidence level to compare the specificity of the NB classifier to the baseline specificity of 0.0 (i.e. excision of all cases).

6.6 Results

The results produced by our final model are in the rightmost block of Table 6.1. This shows the performance of our NB classifier trained on our discordant cases, with a FN:FP cost-ratio of 150:1, and a threshold for excision with model output greater than or equal 0.02. The first column of results shows the baseline performance for comparison, while the subse-

quent blocks show results when trained on just the data, just the binary advice features, and the two combined, at different cost-ratios. The left block of Table 6.1 shows results when the cost-ratio is set to 1:1. We do not present results of training on all non-definitive biopsies, but initial experiments showed no significant ability to reduce benign excisions at an output threshold of 0.02.

6.7 Discussion

We present a framework called ABLe for incorporating expert clinical knowledge into machine learning models for decision support. The framework consists of three different categories of advice that are used to iteratively refine model development. We describe ABLe in detail and illustrate its application to the upgrade prediction task. For this task, we train Naïve Bayes models to estimate the probability of upgrade following a discordant core needle biopsy. Note that this differs from work from Chapter 5 in that we no longer infer logical rules, and instead favor a probabilistic model. This makes the model less interpretable, but it allows for the possibility of shared decision making between clinician and model. The clinician can assess their own personal prediction and weigh it with that of the model.

We train our models using our base dataset, using just the binary rule features, and the two combined. Additionally, we train our models with costs skewed such that misclassifying benign cases is far preferable to misclassifying even a single malignancy. Our results suggest that, by refining our model with the ABLe process, we can significantly reduce the number of truly benign discordant cases that go on to excision without missing a single malignancy. We find these results and the incorporation of expert knowledge very promising, and our future goals are to collect more data to improve performance and to further validate our methods.

7 STATISTICAL RELATIONAL UPLIFT MODELING

We next take a step back from the upgrade prediction task and see how we might be able to leverage uplift modeling to improve treatment assignment in our other breast cancer application (Nassif et al., 2013a).

7.1 Introduction

Recall the breast cancer application introduced in Section 4.3. Breast cancer has two basic states: an earlier *in situ* state where cancer cells are still confined to where they developed, and a subsequent *invasive* state where cancer cells infiltrate surrounding tissue. Since nearly all in situ cases can be cured (American Cancer Society, 2009a), current practice is to treat in situ occurrences in order to avoid progression into invasive tumors (American Cancer Society, 2009b). Nevertheless, the time required for an in situ tumor to reach invasive state may be sufficiently long for an older woman to die of other causes, raising the possibility that treatment may not have been necessary.

Cancer occurrence and diagnosis are determined through biopsy, a costly, invasive, and potentially painful procedure. Treatment, which includes surgery sometimes followed by radiation therapy, is also costly and may induce undesirable side-effects. Hence there is a need for pre-biopsy methods that can accurately identify patient subgroups that would benefit most from treatment, and especially, those who do not need treatment. For the latter, the risk of progression would be low enough to employ watchful waiting (mammographic evaluation at short term intervals) rather than biopsy (Schnitt, 2010).

The literature confirms that the pre-biopsy mammographic appearance as described by radiologists can predict breast cancer (Tabar et al., 2004; Thurfjell et al., 2002). Furthermore, based on age, different pre-biopsy

mammographic features can be used to classify cancer state (Nassif et al., 2010).

As described in Section 4.3, younger women tend to have more aggressive cancers that rapidly proliferate, while older women tend to have more indolent cancers (Fowble et al., 1994; Jayasinghe et al., 2005). We assume that younger in situ patients should always be treated, due to the longer potential time-span for cancer progression. We further assume that older in situ patients whose mammography features are similar to in situ in younger patients should also be treated, because the more aggressive nature of cancer in younger patients may be conditioned on those features. On the other hand, older in situ patients whose mammography features are significantly different from features observed in younger in situ patients are less likely to experience rapid proliferation, and can thus be recommended for watchful waiting.

The motivating problem at hand can readily be cast as an uplift modeling problem (see Section 2.7 and Table 7.1). By maximizing the in situ cases' uplift, we are identifying the older in situ cases that are most different from younger in situ cases, and thus are the best candidates for watchful waiting. Exactly like a marketing campaign would want to target consumers who are the most prone to respond, we want to target the ones that differ the most from the control group.

Table 7.1: Casting mammography analysis in uplift modeling terms.

Intervention	Target Group	Control Group	Positive Class	Negative Class
Time	Older cohort	Younger cohort	In Situ	Invasive

In recent work, Nassif et al. (2012a) inferred older-specific differentially-predictive in situ mammography rules. They used Inductive Logic Programming (ILP) (Lavrac and Dzeroski, 1994), but defined a differential-

prediction-sensitive clause evaluation function that compares performance over age-subgroups during search-space exploration and rule construction. To assess the resulting theory (final set of rules), they constructed a TAN classifier (Friedman et al., 1997) using the learned rules and assigned a probability to each example. They finally used the generated probabilities to construct the uplift curve to assess the validity of their model.

The ILP-based differential prediction model (Nassif et al., 2012a) had several shortcomings. First, this algorithm used a differential scoring function based on m-estimates (Mitchell, 1997) during clause construction, and then evaluated the resulting theory using the area under the uplift curve. This may result in sub-optimal performance, since rules with a high differential m-estimate score may not generate high uplift curves. Second, it decoupled clause construction and probability estimation: after rules are learned, a TAN model is built to compute example probabilities. Coupling these two processes together may generate a different theory with a lower ILP-score, but with a more accurate probability assignment. Finally, rules were added to the theory independently of each other, resulting in redundancies. Having the addition of newer rules be conditioned on the prior theory rules is likely to improve the quality and coverage of the theory.

In this work, we present a novel uplift modeling Statistical Relational Learning (SRL) algorithm that addresses the above shortcomings. Our method, Score As You Lift (SAYL), uses the area under the uplift curve score during clause construction and final theory evaluation, integrates rule learning and probability assignment, and conditions the addition of new theory rules to existing ones. This work makes two main contributions. First, we present the first multi-relational uplift modeling system, and introduce, implement and evaluate a novel method to guide search in an SRL framework. Second, we compare our algorithm to previous approaches, and demonstrate that the system can indeed obtain differential

rules of interest to an expert on real data, while significantly improving the data uplift.

7.2 Background

In order to compare SAYL to prior work we first introduce the state of the art for differential prediction in ILP. We also introduce the SRL algorithm upon which SAYL is based.

7.2.1 Differential Prediction ILP

The earliest work (Nassif et al., 2012b) in differential relational learning included a model-filtering method, whereby a standard ILP algorithm is used to first generate rules trained on the target stratum of the dataset. The rules are then subsequently filtered, removing those rules that are also highly predictive in another stratum of the dataset. See Figure 7.1 for a diagram of the filtration process.

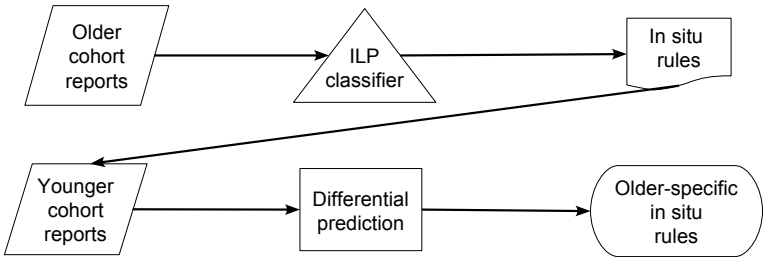


Figure 7.1: Diagram of the model-filtering approach (Nassif et al., 2012b).

Subsequent work (Nassif et al., 2012a) accomplished differential prediction by modifying the rule-scoring function used in a standard ILP algorithm, where the score function is designed to be positively correlated to the performance of a rule over the target stratum and negatively correlated to the performance of a rule over the other stratum. This score

function then guided the ILP learner to select rules that were differentially predictive during theory construction, rather than having to select differential rules as a post-process. Integrating the differential rule selection process into the theory construction process proved to be much more effective than the model filtering approach. See Figure 7.2 for a visual representation of the approach.

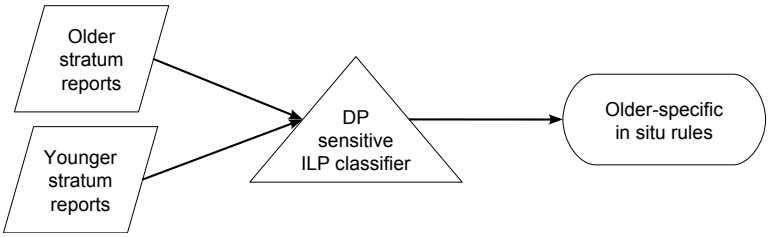


Figure 7.2: The differential prediction (DP) score function approach (Nassif et al., 2012a).

7.2.2 The Score as You Use (SAYU) Algorithm

Score As You Use (SAYU) (Davis et al., 2005) is a Statistical Relational Learner (Getoor and Taskar, 2007) that integrates search for relational rules and classification. It starts from the well known observation that a clause or rule r can be mapped to a binary attribute b, by having $b(e) = 1$ for an example e if the rule r matches e, and $b(e) = 0$ otherwise.

This makes it possible to construct classifiers by using rules as attributes, an approach known as *propositionalization* (Železný and Lavrac, 2006). One limitation, though, is that often the propositional learner has to consider a very large number of possible rules. Moreover, these rules tend to be highly correlated, making it particularly hard to select a subset of rules that can be used to construct a good classifier.

SAYU addresses this problem by evaluating the contribution of rules to a classifier as soon as the rule is generated. Thus, SAYU generates rules

using a traditional ILP algorithm, such as Aleph (Srinivasan, 2007), but instead of scoring the rules individually, as Aleph does, every rule SAYU generates is immediately used to construct a statistical classifier. If this new classifier improves performance over the current set of rules, the rule is added as an extra attribute.

Algorithm 7.1 SAYU

```
Rs ← {}; M_0 ← InitClassifier(Rs)        ▷ Initialize theory and classifiers
while DoSearch() do
    e+ ← RandomSeed();                          ▷ Select seed example
    ⊥_e+ ← saturate(e);                         ▷ Construct bottom clause
    while c ← reduce(⊥_e+) do                   ▷ Loop through clause space
        M ← LearnClassifier(Rs ∪ {c}); ▷ Train classifier with theory
                                                   and new clause
        if Better(M, M_0) then      ▷ If new clause improves performance
            Rs ← Rs ∪ {c}; M_0 ← M;      ▷ Then add new clause to theory
            break
        end if
    end while
end while
```

Algorithm 7.1 shows SAYU in more detail. SAYU maintains a current set of clauses, Rs, and a current reference classifier, M_0. SAYU extends the Aleph (Srinivasan, 2007) implementation of Progol's MDIE algorithm (Muggleton, 1995). Thus, it starts search by randomly selecting a positive example as seed, e^+, generating the corresponding bottom clause[1], $\bot_{e^+}$, and then generating clauses that subsume $\bot_{e^+}$. For every new such clause c, it constructs a classifier M and compares M with the current M_0. If better, it accepts c by adding it to Rs and making M the default classifier. SAYU can terminate search when all examples have been tried without adding new clauses. In practice, termination is often controlled by a time limit.

[1]The bottom clause refers to the most specific hypothesis covering a particular example.

Quite often, most of the execution time will be spent learning classifiers. Therefore, it is important that the classifier can be learned in a reasonable time. Further, the classifier should cope well with many related attributes. We use the TAN classifier (Friedman et al., 1997) because it is computationally inexpensive, making it acceptable for SAYU, and because it compensates well for highly dependent attributes.

Comparing two classifiers is not trivial. SAYU reserves a tuning set for this task: if the classifier M has a better score on both the initial training and tuning sets, the new rule is accepted. The scoring function depends on the problem at hand. Most often SAYU has been used in skewed domains, where the area under the precision-recall curve is regarded as a good measure (Boyd et al., 2012), but the algorithm allows for any metric.

The original SAYU algorithm accepts a logical clause as soon as it improves the network. It may be the case that a later clause would be even better. Unfortunately, SAYU will switch seeds after selecting a clause, so the better clause may be ignored. One solution is to make SAYU less greedy by *exploring* the search space for each seed, up to some limit on the number of clauses, before accepting a clause. We call this version of SAYU *exploration SAYU*: we will refer to it as *e-SAYU*, and to the original algorithm as *greedy SAYU*, or *g-SAYU*.

Algorithm 7.2 details e-SAYU. It differs from g-SAYU in that it keeps track, for each seed, of the current best classifier M_{e^+} and best clause c_{e^+}. At the end, if a clause c_{e^+} was found, we commit to that clause and update the classifier.

7.3 SAYL: Integrating SAYU and Uplift Modeling

SAYL is a Statistical Relational Learner based on SAYU that integrates search for relational rules and *uplift modeling*. Similar to SAYU, every valid

Algorithm 7.2 e-SAYU

```
Rs ← {}; M_0 ← InitClassifier(Rs)
while DoSearch() do
    e^+ ← RandomSeed();
    ⊥_{e^+} ← saturate(e^+);
    c_{e^+} ← ⊤; M_{e^+} ← M_0;
    while c ← reduce(⊥_{e^+}) do
        M ← LearnClassifier(Rs ∪ {c});
        if Better(M, M_e) then    ▷ If new clause improves performance
            c_{e^+} ← c; M_{e^+} ← M;          ▷ Mark new clause as best
        end if
    end while
    if c_{e^+} ≠ ⊤ then                          ▷ If clause was found
        Rs ← Rs ∪ {c_{e^+}}; M_0 ← M_{e^+};      ▷ Add new clause to theory
    end if
end while
```

rule generated is used for classifier construction via propositionalization, but instead of constructing a single classifier, SAYL constructs two classifiers; one for each of the target and control groups. Both classifiers use the same set of attributes, but are trained only on examples from their respective groups. If a rule improves the area under the uplift curve (AUU) by threshold θ, the rule is added to the attribute set. Otherwise, SAYL continues the search.

The SAYL algorithm is shown as Algorithm 7.3. Like SAYU, SAYL maintains separate training and tuning example sets, accepting rules only when the classifiers produce a better score on both sets. This requirement is often extended with a specified threshold of improvement θ, or a minimal rule coverage requirement $minpos$. Additionally, SAYL also has a greedy (g-SAYL) and exploratory (e-SAYL) versions that operate in the same fashion as they do for SAYU.

The key difference between SAYL and SAYU, then, is that SAYL maintains a distinction between the groups of interest by using two separate

Algorithm 7.3 SAYL

```
Rs ← {}; M_0^s, M_0^c ← InitClassifiers(Rs)
while DoSearch() do
    e_s^+ ← RandomSeed();
    ⊥_{e_s^+} ← saturate(e);
    while c ← reduce(⊥_{e_s^+}) do
        M^s, M^c ← LearnClassifiers(Rs ∪ {c});    ▷ Learn two classifiers
        if Better(M^s, M^c, M_0^s, M_0^c) then ▷ If new clause improves uplift
            Rs ← Rs ∪ {c}; M_0^s, M_0^c ← M^s, M^c;          ▷ Add to theory
            break
        end if
    end while
end while
```

classifiers. This is what allows SAYL to demonstrate differential performance as opposed to standard metrics, such as the area under a precision-recall curve. To compute AUU, SAYL simply computes the area under the lift curve (AUL) for each of the groups using the two classifiers and returns the difference.

SAYL and SAYU also differ in selecting a seed example to saturate. Instead of selecting from the entire set of positive examples, SAYL only selects seed examples from the positive examples in the target group. This is not necessary, but makes intuitive sense as clauses produced from examples in the target set are more likely to produce greater lift on the target set in the first place.

7.4 Methods and Results

We apply SAYL to the breast cancer data used in Nassif et al. (2012a) and described in Section 4.3. The data consists of two cohorts: patients younger than 50 years form the *younger* cohort, while patients aged 65 and above form the *older* cohort. The older cohort has 132 in situ and 401 invasive

cases, while the younger one has 110 in situ and 264 invasive (see Table 4.3).

We use 10-fold cross-validation, making sure all records pertaining to the same patient are in the same fold. We run SAYL with a time limit of one hour per fold. We run folds in parallel. For each cross-validated run, we use four training, five tuning and one testing folds. For each fold, we used the best combination of parameters according to a nine-fold internal cross-validation using four training, four tuning and one testing folds. We try both e-SAYL and g-SAYL search modes, vary the minimum number $minpos$ of positive examples that a rule is required to cover between 7 and 13 (respectively 5% and 10% of older in situ examples), and set the threshold θ to add a clause to the theory if its addition improves the AUU to 1%, 5% and 10%. We concatenate the results of each testing set to generate the final uplift curve.

Table 7.2: 10-fold cross-validated SAYL performance. AUL is the area under the lift curve and AUU is the area under the uplift curve. Rule number averaged over the 10 folds of theories. For comparison, we include results of Differential Prediction Search (DPS) and Model Filtering (MF) methods (Nassif et al., 2012a). We compute the p-value comparing each method to DPS, * indicating significance.

Algorithm	AUU	Older AUL	Younger AUL	Rules Avg #	DPS p-value	
SAYL	58.10	97.24	39.15	9.3	0.002	*
DPS	27.83	101.01	73.17	37.1	-	
MF	20.90	100.89	80.99	19.9	0.0039	*
Baseline	11.00	66.00	55.00	-	0.0020	*

Table 7.2 compares SAYL with the previously published Differential Prediction Search (DPS) and Model Filtering (MF) ILP methods (Nassif et al., 2012a), both of which had $minpos = 13$ (10% of older in situ). A baseline random classifier achieves an AUU of 11. We use the Mann-Whitney

test at the 95% confidence level to compare two sets of experiments. We show the p-value of the 10-fold AUU paired Mann-Whitney of each method as compared to DPS, DPS being the state-of-the-art in relational differential prediction. We also plot the uplift curves in Figure 7.3.

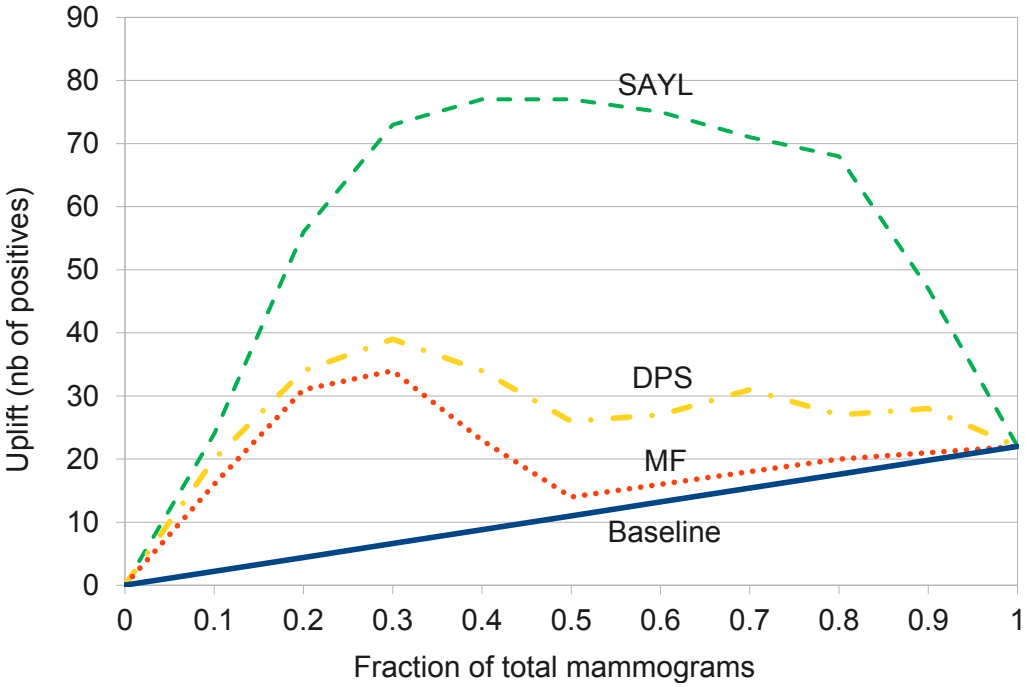


Figure 7.3: Uplift curves for the ILP-based methods (Differential Prediction Search (DPS) and Model Filtering (MF), both with $minpos = 13$ (Nassif et al., 2012a)), a baseline random classifier, and SAYL with cross-validated parameters. Uplift curves start at 0 and end at 22, the difference between older (132) and younger (110) total in situ cases. The higher the curve, the better the uplift.

SAYL 10-fold cross-validation chose g-SAYL in 9 folds and e-SAYL in 1, while $minpos$ was 13 (10% of older in situ) in 5 folds, and 7 (5%) in the remaining 5 folds. Parameter θ was selected to be 1% in 4 folds, 5% in 3 folds, and 10% in the remaining 3 folds. Table 7.3 shows how sensitive SAYL is to those different parameters.

Table 7.3: 10-fold cross-validated SAYL performance under various parameters. Parameter $minpos$ is the minimum number of positive examples that a rule is required to cover. Parameter θ is the AUU improvement threshold for adding a rule to the theory. We also include results of SAYL using cross-validated parameters and Differential Prediction Search (DPS). We compute the p-value comparing each method to DPS, * indicating significance. The maximum values for each column are in bold.

$minpos$	θ (%)	search mode	AUU	Older AUL	Younger AUL	Rules Avg #	DPS p-value	
13	1	g-SAYL	63.29	96.79	33.50	16.4	0.002	*
13	1	e-SAYL	43.51	83.82	40.31	2.0	0.049	*
13	5	g-SAYL	58.06	96.14	38.07	5.9	0.002	*
13	5	e-SAYL	53.37	85.66	32.29	1.8	0.027	*
13	10	g-SAYL	61.68	96.26	34.58	3.6	0.002	*
13	10	e-SAYL	65.36	90.50	25.14	1.1	0.002	*
7	1	g-SAYL	**65.48**	98.82	33.34	18.3	0.002	*
7	1	e-SAYL	25.50	74.39	48.90	3.0	0.695	
7	5	g-SAYL	58.91	96.67	37.76	5.8	0.002	*
7	5	e-SAYL	32.71	79.52	46.81	2.5	0.557	
7	10	g-SAYL	61.98	96.87	34.89	3.6	0.002	*
7	10	e-SAYL	52.35	83.64	31.29	1.6	0.002	*
-	-	SAYL	58.10	97.24	39.15	9.3	0.002	*
13	-	DPS	27.83	**101.01**	**73.17**	**37.1**	-	

7.5 Discussion

We now discuss both the overall model performance as well as the interpretation of the resulting theories.

7.5.1 Model Performance

SAYL significantly outperforms DPS (Table 7.2, Figure 7.3), while ILP-based runs have the highest older and younger AUL (Tables 7.2, 7.3). This is because ILP methods use different metrics during clause construction

and theory evaluation, and decouple clause construction from probability estimation. SAYL builds models that are slightly less predictive of in situ vs. invasive over the younger subset, as measured by the slightly lower older AUL, but on the other hand it effectively maximizes uplift. In fact, increasing lift on one subset will most often increase lift on the other subset, since both sets share similar properties. SAYL avoids this pitfall by selecting rules that generate a high differential lift, ignoring rules with good target lift that are equally good on the controls. These results confirm the limitations of a pure ILP approach, demonstrating significantly higher uplift using SAYL.

The e-SAYL approach explores a larger search space for a given seed before selecting a rule to add to the theory. This results in smaller theories than greedy g-SAYL. Increasing θ, the AUU improvement threshold for adding a rule to the theory, also results in smaller theories, as expected. Ranging $minpos$ between 7 and 13 does not seem to have a sizable effect on rule number.

The g-SAYL approach shows performance that remains constant across all parameters, its AUU varying between 58.06 and 65.48. At the same time, its theory size ranges from 3.6 to 18.3. This indicates that the number of rules is not correlated with AUU. Another indication comes from e-SAYL, whose theory size changes little (1.1 – 3.0), while its performance tends to increase with increasing $minpos$ and θ. Its AUU jumps from the lowest score of 25.50, where it is significantly worse than g-SAYL, to nearly the highest score of 65.36. In fact, g-SAYL outperforms e-SAYL on all runs except $minpos = 13$ and $\theta = 10\%$.

The e-SAYL approach is more prone to over fitting, since it explores a larger search space and is thus more likely to find rules tailored to the training set with a poor generalization. By increasing $minpos$ and θ, we are restricting potential candidate rules to the more robust ones, which decreases the chances of converging to a local minima and overfitting. This

explains why e-SAYL had the worst performances with lowest $minpos$ and θ values, and why it achieved the second highest score of all runs at the highest $minpos$ and θ values. These limited results seem to suggest using e-SAYL with $minpos$ and θ equal to 10%.

7.5.2 Model Interpretation

SAYL returns two TAN Bayes-net models, one for the older and one for the younger, with first-order logic rules as the nodes. Each model includes the classifier node, presented top-most in Figures 7.4 and 7.5, and the same rules. All rules depend directly on the classifier and have at least one other parent. Although both models have the same rules as nodes, TAN learns the structure of each model on its corresponding data subset separately, resulting in different networks. SAYL identifies the features that best differentiate amongst target and control positive examples, while TAN uses these features to create the best classifier over each set.

To generate the final model and inspect the resulting rules, we run SAYL once with five folds for training and five for tuning. As an example, Figures 7.4 and 7.5, respectively, show the older and younger cases TAN models of g-SAYL with $minpos = 13$ and $\theta = 5\%$. The older cohort graph shows that the increase in the combined BI-RADS score is a key differential attribute. The BI-RADS score is a number that summarizes the examining radiologist's opinion and findings concerning the mammogram (BIR, 2003)

We then can see two sub-graphs: the left-hand side sub-graph focuses on the patient's history (prior biopsy, surgery and family history), whereas the right-hand side sub-graph focuses on the examined breast (BI-RADS score, mass size). In contrast, the younger cohort graph is very different: the graph has a shorter depth, and the combined BI-RADS increase node is linked to different nodes.

As the number of rules increases, it becomes harder for humans to interpret the cohort models, let alone their uplift interaction. In ILP-based

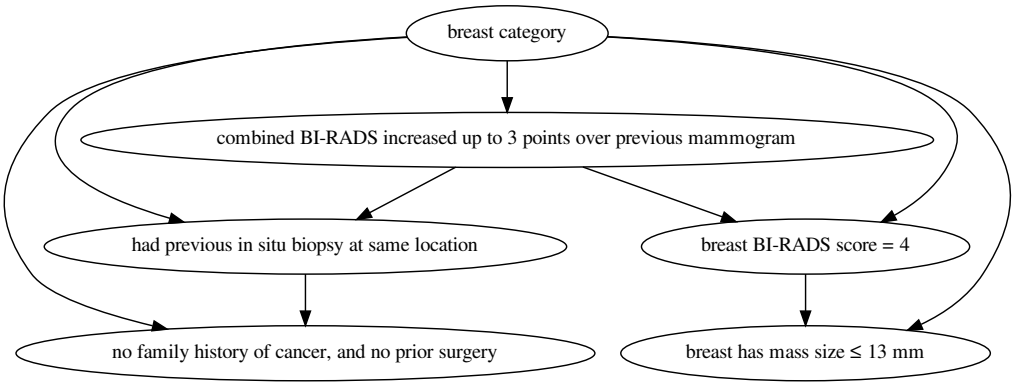


Figure 7.4: TAN model constructed by SAYL over the older cases: the topmost node is the classifier node, and the other nodes represent rules inserted as attributes to the classifier. Edges represent the main dependencies inferred by the model.

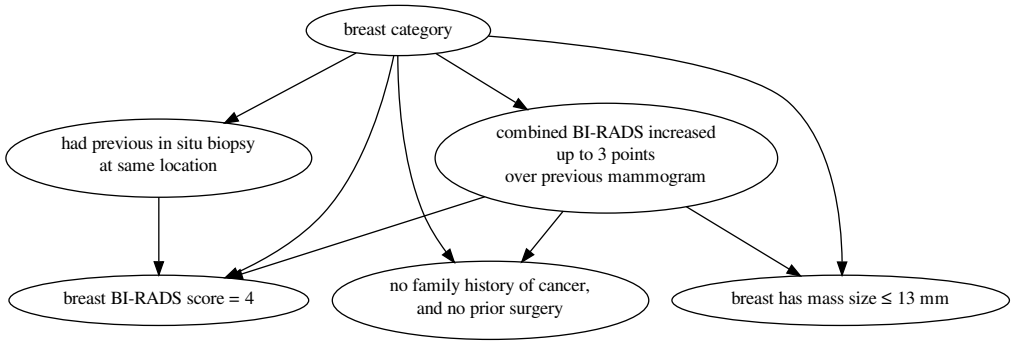


Figure 7.5: TAN model constructed by SAYL over the younger cases. Notice that is has the same nodes but with a different structure than that of the older model shown in Figure 7.4.

differential prediction methods (Nassif et al., 2012a), theory rules are independent and each rule is an older in situ differential rule. In SAYL, theory rules are dependent on each other, whereas a rule can be modulating another rule in the TAN graph. This is advantageous because such modulated rule combinations cannot be expressed in an ILP theory, and therefore might not be learnable. On the other hand, SAYL individual

rules are not required to be specific to older patients with in situ breast cancer. A SAYL rule can predict invasive, or be younger specific, as long as the resulting model is uplifting older in situ. Which decreases clinical rule interpretability.

The average number of rules returned by SAYL is lower than ILP-based methods (Table 7.2). SAYL effectively removes redundant rules by conditioning the addition of a new rule on previous ones. We also note that SAYL, like SAYU, tends to select short rules (Davis et al., 2005). DPS produced five themes amongst its older in situ rules with a significantly better precision and recall: calcification, prior in situ biopsy, BI-RADS score increase, screening visit, and low breast density (Nassif et al., 2012a).

For SAYL runs returning small theories, the resulting rules tend to be differential and fall within the above five themes. For example, g-SAYL with $minpos = 13$ and $\theta = 10\%$ returns three rules:

1. Current study combined BI-RADS increased up to three points over previous mammogram.

2. Had previous in situ biopsy at same location.

3. Breast BI-RADS score $= 4$.

These rules cover two of the five DPS themes, namely prior in situ biopsy and BI-RADS score increase.

As the number of SAYL returned rules increases, rule interactions become more complex, individual rules tend not to remain older in situ differential, and rules are no longer confined to the above themes. In the Figures 7.4 and 7.5 example, we recover the prior in situ biopsy and BI-RADS score increase themes, but we also have non-thematic rules like "no family history of cancer, and no prior surgery". In the two runs returning the largest theories, g-SAYL with $\theta = 1\%$ and $minpos = 7$ and 13, we recover four of the themes, only missing calcification. Note that, as the

graph size increases, medical interpretation of the rules becomes more difficult, as well as identifying novel differential themes, since rules are conditioned on each other.

Although the SAYL rules may not be differential when viewed individually, the SAYL final model is differential, significantly outperforming DPS in AUU. DPS, on the other hand, is optimized for mining differential rules, but performs poorly as a differential classifier. SAYL returns a TAN Bayes net whose nodes are logical rules, a model that is human interpretable and that offers insight into the underlying differential process. Greedy g-SAYL's performance depended little on the parameters, while exploratory e-SAYL's performance increased when requiring more robust rules.

8 SUPPORT VECTOR MACHINES FOR UPLIFT MODELING

Building on our experience with creating models to maximize uplift, we present and evaluate another approach to do so. We introduce uplift maximization to one of the more popular machine learning approaches, support vector machines (SVM). This work was published in Kuusisto et al. (2014).

8.1 Introduction

Section 2.7 introduced the concepts of differential prediction and uplift modeling. Differential prediction has broad and important applications across multiple domains, but as specific motivating applications, we consider two medical tasks here. One is the task described in Section 4.3 in which we want to identify older patients with breast cancer who are good candidates for "watchful waiting" as opposed to treatment. The other task, described in Section 4.2, is one in which we want to identify patients who are most susceptible to adverse effects of COX-2 inhibitors, and thus not prescribe such drugs for these patients.

In the adverse drug effects task, due to individual variance in response to drugs, there will be some people at increased risk of MI as a result of taking the drug, some who are at increased risk of MI regardless of treatment, some who are at decreased risk regardless, and perhaps even some who are at decreased risk as a result of taking the drug. Just like in the marketing task, which group an individual belongs to cannot be directly observed. We propose that training a classifier to identify individuals for whom taking a COX-2 inhibitor increases their risk of MI is analogous to identifying *Persuadables*.

In the breast cancer task, we know that younger patients often have aggressive cancers while older patients have both aggressive and indolent

cancers. Again, like the uplift modeling task, which type of cancer a patient has is not directly observable. We propose that training a classifier to identify in situ cancers with features specific to older patients, and thus likely less aggressive varieties of cancer, is also analogous to identifying *Persuadables*.

The adverse drug event task alone is of major worldwide significance, and the significance of the breast cancer task cannot be overstated. Finding a model that is predictive of an adverse event for people on a drug versus not could help in isolating the key causal relationship of the drug to the event, and using machine learning to uncover causal relationships from observational data is a big topic in current research. Similarly, finding a model that can identify patients with breast cancer that may not be threatening enough in their lifetime to require treatment could greatly reduce overtreatment and costs in healthcare as a whole.

Several classification and regression algorithms have been proposed and evaluated according to the uplift measure (Radcliffe and Surry, 2011; Rzepakowski and Jaroszewicz, 2012; Nassif et al., 2012a; Jaśkowski and Jaroszewicz, 2012; Zaniewicz and Jaroszewicz, 2013). These models were designed to improve the uplift curve, but do not directly optimize it. We show that indeed it is possible to directly optimize uplift using a support vector machine (SVM), and we propose and evaluate our SVM^{Upl} model, which does so. This model is constructed by applying Joachims' work on the optimization of multivariate measures (Joachims, 2005) with SVMs. We evaluate multiple models on our motivating applications and SVM^{Upl} shows the best performance in differential prediction in most cases.

8.2 Uplift-Agnostic Models

Recall from Section 2.7 that uplift modeling is an approach used in marketing to identify the hidden *Persuadable* customer types, separating cus-

tomers into target and control subgroups. For notational convenience, we refer to the target and control subgroups as A and B respectively.

Section 2.7.1 explained that the *lift* curve reports the total percentage of examples that a classifier must label as positive (x-axis) in order to obtain a certain recall (y-axis), expressed as a count of true positives instead of a rate. As usual, we can compute the corresponding area under the lift curve (AUL).

Section 2.7.1 also explained that uplift is the difference in lift produced by a classifier between subgroups A and B, at a particular threshold percentage of all examples. We can compute the area under the uplift curve (AUU) by subtracting their respective AULs, where higher AUU indicates an overall stronger differentiation of subgroup A from B:

$$\mathrm{AUU} = \mathrm{AUL}_A - \mathrm{AUL}_B \tag{8.1}$$

In the following subsections, we first introduce a number of possible baseline modeling approaches to which we compare our $\mathrm{SVM}^{\mathrm{Upl}}$ approach. These baselines are all SVM implementations that can be used to address the differential prediction task, yet they do not directly optimize the uplift measure at training time.

8.2.1 Standard SVM

We start with the standard SVM approach as a first baseline of comparison to $\mathrm{SVM}^{\mathrm{Upl}}$. As described in Section 2.4, support vector machines attempt to find a maximum-margin separating plane between positive and negative examples. The standard soft-margin definition (Vapnik, 1998) minimizes:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{8.2}$$

Subject to $\xi_i \geqslant 0$. The formulation tries to minimize the two-norm of the weight vector, $\mathbf{w}$, and hence maximize the margin, while softly allowing some errors, ξ_i, whose cost depends on the tunable parameter C. Errors here are distances from the decision boundary of examples that lie on the wrong side of the boundary.

For the sake of comparison with SVM^{Upl}, we evaluate the ability of a standard linear SVM model to produce uplift in our applications of interest (see Algorithm 8.1).

Algorithm 8.1 Standard SVM Experiment

```
Input :Train, Test                          ▷ Given a train and test set
C ← CVSelectC(Train);                       ▷ Use cross-validation to select C
                                              producing best uplift
M ← TrainSVM(Train, C);                     ▷ Train standard SVM
EvaluateUplift(M, Test);                    ▷ Measure uplift on test set
```

8.2.2 Target-Only SVM

Our second baseline is a standard SVM trained on only the target subgroup of the training set (see Algorithm 8.2). This is the marketing "response modeling" approach described in Section 2.7. In marketing, the idea is that a model trained on only the target subgroup will be able to predict customers that will respond after being targeted with marketing activity. While this may be true, recall that this approach thus does not distinguish between *Persuadables* and *Sure Things*.

In our applications, this means only training on the data for the older subgroup of breast cancer patients, or the subgroup of MI patients who have been prescribed COX-2 inhibitors.

8.2.3 Flipped-Label SVM

Jaśkowski and Jaroszewicz (2012) propose a general method for adapting

Algorithm 8.2 Target-Only SVM Experiment

Input :$Train, Test$ ▷ Given a train and test set
$Train_A, Train_B \leftarrow GetT\&C(Train)$; ▷ Split training set by target and control groups
$C \leftarrow CVSelectC(Train_A)$; ▷ Use cross-validation to select C producing best uplift
$M \leftarrow TrainSVM(Train_A, C)$; ▷ Train standard SVM
$EvaluateUplift(M, Test)$; ▷ Measure uplift on test set

standard models to increase uplift, which we will use as our third baseline for comparison. This is accomplished by flipping the classification labels in the control subgroup during training. In this way, the classifier is trained to correctly predict the positive class on the target subgroup, A, whereas it is trained to predict the negative class in control subgroup, B (see Algorithm 8.3). The resulting classifier should then perform better on subgroup A than subgroup B, thus increasing uplift.

Algorithm 8.3 Flipped-Label SVM Experiment

Input :$Train, Test$ ▷ Given a train and test set
$Train_A, Train_B \leftarrow GetT\&C(Train)$; ▷ Split training set by target and control groups
$Train'_B \leftarrow SwapPosNegLabels(Train_B)$; ▷ Flip the positive and negative labels on control group
$Train \leftarrow Train_A \cup Train'_B$; ▷ Combine back into one set
$C \leftarrow CVSelectC(Train)$; ▷ Use cross-validation to select C producing best uplift
$M \leftarrow TrainSVM(Train, C)$; ▷ Train standard SVM
$EvaluateUplift(M, Test)$; ▷ Measure uplift on test set

8.2.4 Two-Cost SVM

Our fourth and final baseline approach is to simply treat the errors on the target and control subgroups differently (see Algorithm 8.4). Specifically, we propose the following adaptation of the standard minimization

problem:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C_A \sum_{i=1}^{|A|} \xi_i + C_B \sum_{j=1}^{|B|} \xi_j \tag{8.3}$$

subject to $\xi_i \geqslant 0, \xi_j \geqslant 0, C_A \geqslant 0$ and $C_B \geqslant 0$. We also assume $C_A \geqslant C_B$, penalizing errors on the target subgroup more than those on the control subgroup. Similar to the flipped-label model, the resulting classifier should then perform better on subgroup A than subgroup B, thus increasing uplift.

Algorithm 8.4 Two-Cost SVM Experiment

```
Input :Train, Test                                  ▷ Given a train and test set
Train_A, Train_B ← GetT&C(Train);                   ▷ Split training set by target and
                                                      control groups
C_A, C_B ← CVSelectC(Train_A, Train_B);             ▷ Use cross-validation to select C_A
                                                      and C_B producing best uplift
M ← TrainSVM(Train_A, Train_B, C_A, C_B);           ▷ Train standard SVM with separate
                                                      costs for target and control
EvaluateUplift(M, Test);                            ▷ Measure uplift on test set
```

8.3 Multivariate Performance Measures

Next, before we can define our SVM$^{\text{Upl}}$ approach, we briefly review Joachims' SVM$^{\text{perf}}$ approach (Joachims, 2005) to maximize area under the ROC curve (AUC) (Joachims et al., 2009). Note that we use an $(\mathbf{x}, y)$ feature vector and label pair notation to represent examples throughout. Let tuples $\bar{\mathbf{x}} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ and $\bar{y} = (y_1, ..., y_n)$. Also, let tuple $\bar{y}' = (y'_1, ..., y'_n)$ be a predicted assignment over the n examples, and let $\bar{\mathcal{Y}}$ be the set of all possible assignments. This approach proposes that we want to find the hypothesis that maximizes some objective function over the training data:

$$\underset{\bar{y}' \in \bar{\mathcal{Y}}}{\operatorname{argmax}}\, f(\bar{\mathbf{x}}, \bar{y}) + \Delta(\bar{y}', \bar{y})$$

where $\Delta(\bar{y}', \bar{y})$ is a problem-specific loss function, and $f(\bar{x}, \bar{y})$ is a score function.

We first define the loss function for AUC, Δ_{AUC}. This loss formulation applies to the AUC when AUC is defined as:

$$1 - \frac{\text{BadPairs}}{N \times P}$$

where N is the number of negative examples, P is the number of positive examples, and *BadPairs* is the number of pairs (i, j) such that $y_i = 1, y_j = -1$, and $y'_i < y'_j$. That is, *BadPairs* is the number of pairs of positive and negative examples, such that the positive example has a predicted score lower than that of the negative example. Joachims addresses the optimization problem in terms of pairs y'_{ij}, where y'_{ij} is 1 if $y'_i > y'_j$, and -1 otherwise. The loss function for AUC is simply the number of swapped pairs:

$$\Delta_{AUC}(\bar{y}', \bar{y}) = \sum_{i=1}^{P} \sum_{j=1}^{N} \frac{1}{2}(1 - y'_{ij}) \tag{8.4}$$

The score function, $f(\bar{x}, \bar{y})$, is a product of a weight vector, $\mathbf{w}$, and a function, Ψ, of input features and predicted outputs:

$$\mathbf{w}^T \Psi(\bar{x}, \bar{y}') = \frac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{N} y'_{ij}(\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{x}_j) \tag{8.5}$$

This score function then increases as pairs of positive and negative examples are pushed farther apart on correct sides of the decision boundary, and the loss function penalizes for the number of swapped pairs, thus optimizing for AUC. Joachims' algorithm solves for this problem.

8.4 Maximizing Uplift

Our goal is to find the parameters $\mathbf{w}$ that are optimal for maximizing AUU. Similar to AUC (Joachims, 2005; Zhang et al., 2009; Narasimhan and Agarwal, 2013), AUL depends on the rankings between pairs of examples. Recall from Section 2.7.1 the relationship between lift and ROC. As shown in Tufféry (2011), if we define the positive skew of a dataset as $\pi = \frac{P}{P+N}$, the AUL is related to the AUC by:

$$AUL = P\left(\frac{\pi}{2} + (1-\pi)AUC\right) \tag{8.6}$$

Expanding Equation 8.1 with Equation 8.6:

$$AUU = P_A\left(\frac{\pi_A}{2} + (1-\pi_A)AUC_A\right) - P_B\left(\frac{\pi_B}{2} + (1-\pi_B)AUC_B\right) \tag{8.7}$$

where P_A, P_B, π_A, and π_B are properties of the two subgroups, and thus independent of the classifier. Removing constant terms we see that maximizing uplift is equivalent to:

$$\begin{aligned} max(AUU) &\equiv max(P_A(1-\pi_A)AUC_A - P_B(1-\pi_B)AUC_B) \\ &\propto max\left(AUC_A - \frac{P_B(1-\pi_B)}{P_A(1-\pi_A)}AUC_B\right) \end{aligned} \tag{8.8}$$

Defining $\lambda = \frac{P_B(1-\pi_B)}{P_A(1-\pi_A)}$ we have:

$$max(AUU) \equiv max(AUC_A - \lambda AUC_B) \tag{8.9}$$

Therefore, maximizing AUU is equivalent to maximizing a weighted difference between two AUCs.

Equation (8.9) suggests that we can use the AUC-maximizing SVM formulation to optimize AUU. First, we make AUU maximization into the maximization of a sum by switching positive and negative labels in

subgroup B (call this AUC_B^-):

$$\begin{aligned} max(AUU) &\equiv max(AUC_A - \lambda(1 - AUC_B^-)) \\ &\equiv max(AUC_A + \lambda AUC_B^-) \end{aligned} \tag{8.10}$$

We can now encode our problem using Joachims' formulation of the AUC. In this case, we simply have two AUCs. One, as before, is obtained from the y_{ij} where the (i, j) pairs range over the target subgroup A. The second corresponds to pairs y_{kl} where the (k, l) pairs range over B. On switching the labels, we must consider y_{lk} where k ranges over the positives in B, and l over the negatives in B.

After switching labels, we can translate Equation 8.4 to obtain our new loss Δ_{Auu} as the weighted sum of two AUC losses:

$$\Delta_{Auu}(\bar{y}', \bar{y}) = \sum_{i=1}^{P_A} \sum_{j=1}^{N_A} \frac{1}{2}(1 - y'_{ij}) + \lambda \sum_{k=1}^{P_B} \sum_{j=1}^{N_B} \frac{1}{2}(1 - y'_{lk}) \tag{8.11}$$

From Equation 8.5 we construct a corresponding score function:

$$\mathbf{w}^T \Psi(\bar{x}, \bar{y}') = \frac{1}{2} \sum_{i=1}^{P_A} \sum_{j=1}^{N_A} y'_{ij}(\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{x}_j) + \lambda \frac{1}{2} \sum_{k=1}^{P_B} \sum_{l=1}^{N_B} y'_{lk}(\mathbf{w}^T \mathbf{x}_l - \mathbf{w}^T \mathbf{x}_k) \tag{8.12}$$

The optimization is now simply a function of two independent sets of examples for the same weight vector $\mathbf{w}$, and we can plug these new functions into SVM^{perf}. See Algorithm 8.5 for the experimental process for SVM^{Upl}.

Algorithm 8.5 SVM$^{\text{Upl}}$ Experiment

Input :$Train, Test$ ▷ Given a train and test set
$Train_A, Train_B \leftarrow GetT\&C(Train)$; ▷ Split training set by target and control groups
$C \leftarrow CVSelectC(Train_A, Train_B)$; ▷ Use cross-validation to select C producing best uplift
$M \leftarrow TrainSVM^{\text{Upl}}(Train_A, Train_B, C)$; ▷ Train SVM$^{\text{Upl}}$
$EvaluateUplift(M, Test)$; ▷ Measure uplift on test set

8.5 Methods

We implemented our SVM$^{\text{Upl}}$ method [1] using the SVM$^{\text{perf}}$ codebase, version 3.00[2]. We implemented the two-cost model using the LIBSVM codebase (Chang and Lin, 2011), version 3.17[3]. All other uplift-agnostic approaches were run using LIBSVM, but required no changes to the code.

8.5.1 Simulated Customer Experiments

As described in Section 2.7 the goal of uplift modeling is to identify the hidden *Persuadable* customer group. The fact that customer groups cannot be directly observed, however, makes it difficult to understand if maximizing area under the uplift curve really does help to produce classifiers that can specifically identify *Persuadables*. To confirm this assumption, we first present an experiment on the synthetic dataset described in Section 4.4. For this dataset, we generate a synthetic customer population and subject them to a simulated marketing campaign, where customers are made to respond stereotypically for their customer group. The dataset is composed of 10,000 customers (see Table 4.4 for more detail).

[1] The code can be found at: `http://ftp.cs.wisc.edu/machine-learning/shavlik-group/kuusisto.ecml14.svmuplcode.zip`

[2] `http://www.cs.cornell.edu/people/tj/svm_light/svm_perf.html`

[3] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

We removed the customer group feature from the training set and trained three different classifiers to demonstrate performance. First, we trained a standard SVM classifier on the entire dataset. Next, we trained a target-only SVM on just the target subgroup. Finally, we trained SVM^{Upl}. We evaluated the results using 10-fold cross-validation and used internal cross-validation to select parameters.

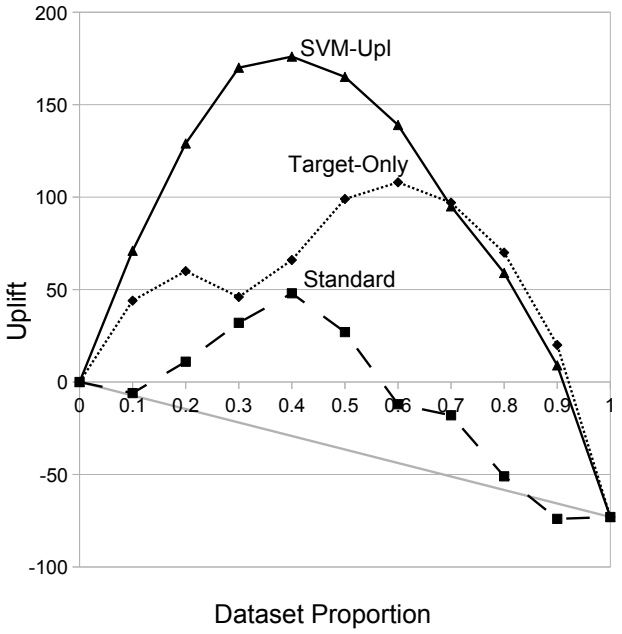


Figure 8.1: Uplift curves (higher is better) for three different classifiers on the simulated customer dataset. SVM-Upl is the uplift maximizing support vector machine presented in Chapter 8. Target-Only is a support vector machine trained only on the targeted subgroup of the population. Standard is a support vector machine trained on both the target and control subgroups with no distinction made between them.

Figure 8.1 shows the uplift curves on the synthetic customer dataset. As expected, the SVM designed to maximize uplift produces the highest uplift curve, while the standard SVM trained on the entire dataset produces the lowest. More importantly, Figure 8.2 shows ROC curves on the synthetic

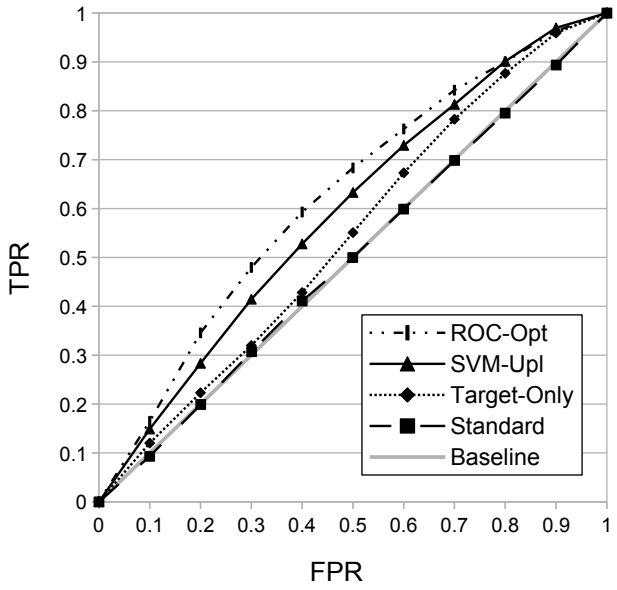


Figure 8.2: ROC curves (higher is better) for three different classifiers on the simulated customer dataset when the hidden *Persuadable* customer group is treated as the positive class. Note that the ROC-Opt curve is for an SVM trained to maximize AUC when trained with the ground truth *Persuadable* labels, representing an empirical optimum ROC curve.

customer dataset when the *Persuadable* customers are considered to be the positive class. Recall that this feature was unobserved at training time, but identifying *Persuadables* is the real goal in the marketing domain. The ROC-Opt model shown in this second figure is an SVM trained to optimize the AUC when given the true *Persuadable* label. This model then represents an empirical optimum ROC curve that can be achieved when given the ground truth, which is not otherwise available. As hoped, the SVM^{Upl} has the highest ROC curve and is quite close to the empirical optimum curve, whereas the standard SVM trained on the entire dataset hovers around the diagonal.

8.5.2 Medical Application Experiments

Recall that our motivating applications are to produce a classifier to predict in situ breast cancer specific to older patients, and produce a classifier to predict myocardial infarction (MI) specific to patients who took COX-2 inhibitors. We apply all of the proposed approaches to the breast cancer data described in Section 4.3 and the COX-2 data described in Section 4.2 (see Tables 4.3 and 4.2 for more detail).

The breast cancer data consists of two cohorts: patients younger than 50 years old form the *younger* cohort, while patients aged 65 and above form the *older* cohort. The older cohort has 132 in situ and 401 invasive cases, while the younger one has 110 in situ and 264 invasive.

The COX-2 dataset consists of patients separated into two equally-sized subgroups: patients who have been prescribed COX-2 inhibitors and those who have not. The group prescribed COX-2 inhibitors has 184 patients who had MI, and 1776 who did not. The subgroup not prescribed COX-2 inhibitors has the same number of patients for each outcome.

Table 8.1: 10-fold cross-validated performance for all proposed approaches on the breast cancer dataset (* indicates significance).

Model	Older AUL	Younger AUL	AUU	Per-fold AUU μ	Per-fold AUU σ	SVM^{Upl} p-value	
SVM^{Upl}	64.26	45.05	19.21	1.93	0.78	-	
Two-Cost	74.30	60.76	13.54	1.45	1.18	0.432	
Older-Only	67.70	61.85	5.85	1.03	1.15	0.037	*
Standard	75.35	64.34	11.01	1.26	0.38	0.049	*
Flipped	53.90	49.08	4.82	0.77	0.58	0.020	*
Baseline	66.00	55.00	11.00	1.10	0.21	0.004	*

We use 10-fold cross-validation for evaluation. Cost parameters were selected for each fold using 9-fold internal cross-validation. For all approaches, the cost parameter was selected from {10.0, 1.0, 0.1, 0.01, 0.001, 0.0001, 0.00001}. For the two-cost model, C_A and C_B were selected from

Table 8.2: 10-fold cross-validated performance for all proposed approaches on the MI dataset (* indicates significance).

Model	COX-2 AUL	No COX-2 AUL	AUU	Per-fold AUU μ	Per-fold AUU σ	SVM^{Upl} p-value	
SVM^{Upl}	123.38	72.70	50.68	5.07	2.04	-	
Two-Cost	126.23	106.25	19.99	2.43	1.54	0.004	*
COX-2-Only	151.50	137.70	13.80	1.18	1.52	0.002	*
Standard	147.69	146.49	1.20	-0.16	1.25	0.002	*
Flipped	102.15	73.63	28.52	2.97	1.35	0.037	*
Baseline	0.00	0.00	0.00	0.00	0.00	0.002	*

all combinations of values from the set such that $C_A > C_B$. We plot the final uplift curves for each approach along with the uplift for a baseline random classifier in Figures 8.3 and 8.4.

Tables 8.1 and 8.2 compare SVM^{Upl} with every other approach proposed as well as a fixed baseline random classifier. We use the Mann-Whitney test at the 95% confidence level to compare approaches based on per-fold AUU. We show the per-fold mean, standard deviation, and p-value of the 10-fold AUU paired Mann-Whitney of each method as compared to SVM^{Upl} (* indicates significance).

8.6 Discussion

We now discuss the overall model performance for both medical applications, as well as interpretation of the breast cancer model.

8.6.1 Model Performance

The results on the breast cancer dataset in Table 8.1 show that SVM^{Upl} produces significantly greater uplift than all proposed approaches, except for the two-cost model. This exception may be a result of the higher variance of the model on this particular dataset. The results on the MI

dataset in Table 8.2 show that SVM^{Upl} produces the greatest uplift in all cases.

Figure 8.3 shows SVM^{Upl} with an uplift curve that dominates the rest of the approaches until around the 0.7 threshold on the breast cancer dataset. Most other approaches produce curves that sit around or below the baseline.

Figure 8.4 tells a similar story, with SVM^{Upl} dominating all other methods across the entire space on the MI dataset. In this dataset, however, only the standard SVM approach consistently performs below the baseline, whereas all other methods appear to produce at least modest uplift.

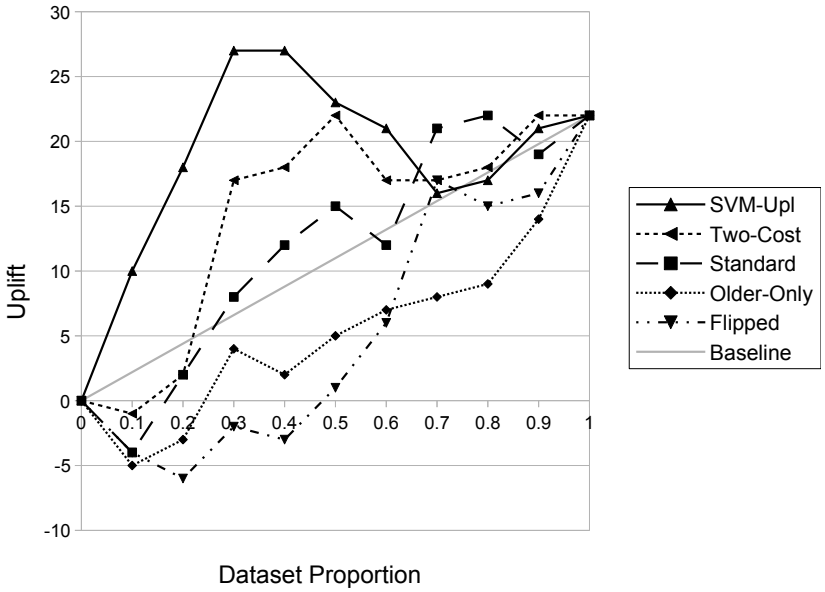


Figure 8.3: Uplift curves (higher is better) for all approaches on the breast cancer dataset.

8.6.2 Model Interpretation

While the main goal of SVM^{Upl} is to develop a model that directly maximizes uplift, it is also desirable to be able to interpret learned models,

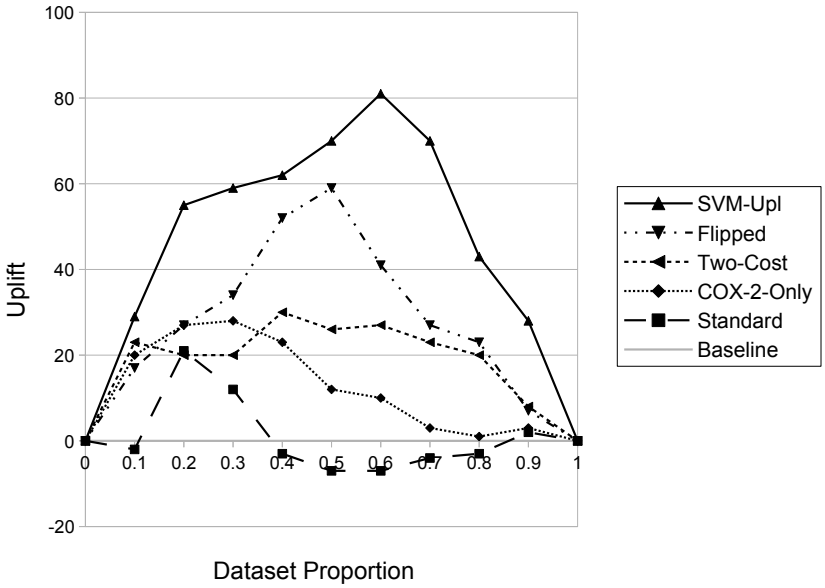


Figure 8.4: Uplift curves (higher is better) for all approaches on the MI dataset. Note that the baseline uplift lies on the x-axis due to the equal number of patients with MI in each subgroup.

especially from a clinical standpoint. SVMs do not lend themselves easily to interpretation, but one method of SVM feature selection that can be used for interpreting feature importance is recursive feature elimination (RFE) (Guyon et al., 2002). RFE allows for interpretation of the model by ranking individual features by way of learned model coefficients.

Briefly, RFE operates by training an initial model with all of the features included in the dataset (see Algorithm 8.6). Once the initial model has been trained, the feature with the smallest magnitude coefficient is identified and removed. The removed feature is assumed to be the least important feature, as it affects the model output the least (assuming feature values have been normalized). Furthermore, the sign of the coefficient can be interpreted as determining the directionality of the feature contribution. After this least important feature is removed from the dataset, a new model is trained. The coefficient with the smallest magnitude in this model is

the second least important feature. The process continues in this way, eliminating features one at a time.

Algorithm 8.6 Recursive Feature Elimination

```
R ← {};                                          ▷ Init rank list
T ← Train;                                       ▷ Init data with full training set
M ← TrainSVM(T);                                 ▷ Train initial SVM
while T ≠ ∅ do
    f ← MinMagnitudeCoeff(M);                    ▷ Select least important feature
    T ← T \ f;                                   ▷ Remove feature
    R ← R ∪ f;                                   ▷ Add feature to ranking
    M ← TrainSVM(T);                             ▷ Train new SVM
end while
return R;                                        ▷ Return feature ranking
```

We ran RFE experiments with SVM^{Upl} on our breast cancer dataset and presented the top ranked features to our expert radiologist. The top five features correlated with in situ breast cancer specific to older women can be found in Table 8.3. The table includes positive/negative correlation, as well as assessments made by an expert radiologist, where 10 indicates a clinically interesting feature and 1 indicates a clinically uninteresting feature.

As the table shows, SVM^{Upl} was able to pick up on clinically-relevant features. Remember, the features are not just associated with in situ cancer in general, but are relevant for understanding in situ breast cancer specific to older patients. We find these results promising as they demonstrate even greater potential of uplift modeling for the understanding of disease and knowledge discovery.

In summary, we have introduced a support vector model directed toward uplift modeling. The SVM^{Upl} approach optimizes uplift by relying on the relationship between AUL and AUC, and on multivariate function optimization used in prior work to optimize AUC. Our results suggest that SVM^{Upl} does indeed achieve better uplift in unseen data than the other

Table 8.3: The five most important features to predict older-specific in situ breast cancer as determined by recursive feature elimination. This table also includes the positive/negative correlation directionality, as well as a radiologist assessment of relevance (10 = clinically interesting, 1 = clinically counter-intuitive).

Rank	Feature	Older In Situ Correlation	Radiologist Assessment
1	Linear Calc. Distribution Present	Positive	10
2	Spiculated Mass Margin Present	Negative	10
3	Palpable Lump Present	Positive	3
4	Irregular Mass Shape Present	Negative	9-10
5	No Family History	Negative	8

approaches, and we have demonstrated that interpretation of the model coefficients demonstrates clinically-relevant feature rankings driven by uplift modeling.

9 EXPERIMENTS IN INDIVIDUALIZED TREATMENT ASSIGNMENT

In this chapter, we present work that investigates the use of machine learning to estimate individualized treatement effects (ITE) as an alternative to traditional statistical approaches. We do not present new algorithms in this work, but instead simply argue that machine learning offers tools to help improve patient care. This is joint work with Jeremy Weiss, and much of it has already been included in his dissertation (Weiss, 2014). Since his dissertation was published, we have performed additional experiments with real data and a successful peer review (Weiss et al., 2015).

9.1 Introduction

Recall from Chapter 3 that estimation of the risk of a disease attributable to an exposure or treatment is one of the fundamental tasks in epidemiology and is typically determined using randomized controlled trials (RCTs). The *average* treatment effect (ATE)–the primary outcome of an RCT–is the average difference between treatment arms in the probability of the outcome, which is then used to recommend future treatments for individual patients. While ATEs are indicative of true treatment effects even in the presence of confounders, they have limited applicability for individual patients because we do not expect the same treatment effect in every person and diversity of effects goes beyond a population's nonuniform prior risk. Furthermore, the ATE is population-distribution dependent, so it inherently lacks generalizability to alternative test distributions. Therefore, we consider modeling the *individualized* treatment effect (ITE).

Currently, non-randomized epidemiological studies adopt classical statistical procedures, such as logistic regression (LR), in seeking to improve patient outcomes. However, machine learning has developed many alter-

native models for conditional probability distributions (CPDs). Advances in machine learning should be leveraged in estimation of the treatment effect–a crucial epidemiological outcome of interest. Our work proposes the use of non-parametric algorithms possessing consistency results in place of logistic regression because of their theoretical ability to accurately recover the CPDs. Parametric models make assumptions about the distribution from which data are drawn, whereas non-parametric models do not, and consistent learners' estimates converge toward the true distribution of the data as they are given more examples.

In this chapter, we show the value of ITE over ATE as well as the use of conditional probability models over logistic regression, using both synthetic and real data. We demonstrate our ability to recover the true ITE in synthetic data, and we show the generalizability of the conditional probability model to alternative population distributions of increasing Kullback-Leibler (KL) divergences. We also show that a conditional probability model learned with a consistent, non-parametric algorithm achieves a lower mean squared error (MSE) estimate of the ITE than logistic regression. Furthermore, we show that the conditional probability model produces a better estimate of the ITE than logistic regression on a real RCT dataset of D-penicillamine use for primary biliary cirrhosis. Additionally, we show that learning from propensity-score matched (PSM) examples and stable inverse probability of treatment-weighted ((s)IPT-W) examples do not improve over unweighted examples for making ITE prediction when only observational data is available. Thus, by casting treatment effect estimation in a machine learning framework, we introduce ways machine learning can be used to develop improved, personalized-risk estimates and treatment recommendations.

9.2 Background

Randomized controlled trials randomize patients to different treatment arms and measure the rate or probability of an outcome. The treatment arm with the highest success rate determines the preferred treatment. Randomization is crucial to balance confounders, which are covariates that lead to the outcome and are associated with the treatment (e.g. smoking is a confounder for the effect of alcohol consumption on outcomes like lung cancer). Randomization also balances confounders not measured in the study, so the conclusion is free of confounding bias in expectation.

In general, one cannot know what will happen to a specific patient under each treatment arm. The treatment that is given elicits the "true" outcome, and the treatment(s) not given elicits the "counterfactual" outcome. The counterfactual outcome is impossible to measure, because a patient cannot be both given and not given treatment, but with randomization and the assumption that patients are drawn from an underlying population distribution, the expected outcome of patients assigned to a treatment arm is the same as the expected outcome of patients with the same treatment, true or counterfactual. Thus, RCTs provide a recommendation about the treatment effect for every treatment arm in the study for every patient.

The RCT, however, is impractical or infeasible for many exposure-outcome pairs. For example, randomization to a harmful treatment, such as smoking, is unethical. In such cases, observational studies are used to derive risk attribution statements. These include studies that use known-confounder-modeling (Prentice, 1976), propensity scoring (Austin, 2011; Rosenbaum and Rubin, 1983), inverse probability of treatment-weighting (Robins et al., 2000), and doubly-robust estimators (Bang and Robins, 2005). The two main ideas in these methods are to (1) adjust for confounders by modeling them, and (2) manipulate the population distribution so that the treatment is independent of confounders given the outcome. These meth-

ods rely upon modeling, but cannot do so effectively if they are missing important contributors to their model: the unobserved confounders. Thus, one key assumption in all of these methods is that there are no unobserved confounders (NUCA), which is difficult to determine in practice. Also, in most of these approaches, a model is assumed for the CPD of the outcome given the exposure and covariate. In these cases, the counterfactual outcomes, which are never observed, are assumed to follow the model CPD.

A second assumption made in clinical studies is the exclusion of intermediate variables–covariates that are on the causal pathway from the treatment to the disease. If included, the treatment effect is underestimated because the effect can be "explained away" by the intermediate variable. The exclusion of intermediate variables decreases the richness of the model, as the intermediate variable may also modify the treatment effect, and analyses that acknowledge and integrate this information exist (Robins, 1989).

The ITE provides the effect per individual instead of a population-level effect, and information about future individuals can be leveraged in determining optimal treatment choices. Unlike in ATE estimation though, acquiring sufficient counts to estimate the counterfactual ITE outcome is unachievable for any moderate-sized feature vector because the number of possible feature states is exponentially large. Therefore, a modeling approach to estimate the counterfactual outcome becomes necessary. These can be the same CPD models used in pseudo-randomized ATE estimation, e.g. logistic regression, but in Section 9.3 we will discuss two reasons to adopt other machine learning models: non-uniform treatment recommendations and non-parametric consistency.

Modeling of the individualized treatment has been implemented in several studies. Qian and Murphy (2011) develop the framework of reward modeling and using model predictions to estimate individualized

treatment rules (ITRs). Our work is related but instead makes statements about the utility of the ITE, the generalizability of the ITE, and the preference for using unweighted observational data for ITE estimation, all with simulations to illustrate these advantages. Our simulations based on synthetic data have access to a ground truth ITE, which we use to assess our ITE estimations.

However, it is possible to assess the benefit of ITE without access to ground truth. Vickers et al. (2007) provides an unbiased method to estimate the advantages of using the ITE recommendation over the ATE recommendation using existing RCT data. They show that by counting outcomes in the subset of patients where ITE- and ATE- treatment recommendations agree, the expected difference in treatment recommendations can be estimated. Our experiments include such analyses to show that the ITE-recommendation can be estimated without access to the counterfactual outcomes.

The methods we adopt do not directly optimize the individualized treatment recommendation. Instead, we model the conditional probability distribution, and then the differences in probability are determined using the estimates for the treatment effect of the true and counterfactual treatments. Zhao et al. (2012) develop a method to directly optimize for the ITR under a surrogate loss function from RCT data. While this method produces individualized recommendations, we believe a model should also provide treatment effect estimates under each treatment arm, because the treatment effect itself is critical information clinically.

9.3 Methods

We first define ITE modeling formally. Let the ITE for an outcome $y \in \{0, 1\}$ of a patient with features v given treatment $u \in \{0, 1\}$ be the difference in estimates: $p(y = 1|u = 1, v) - p(y = 1|u = 0, v)$. The key assumption made

in these modeling approaches is that both potential outcomes–the true outcomes y_{true} and the counterfactual outcomes y_{cf}–come from the CPD model, that is, $p(y_{\text{cf}}|u, v) = p(y_{\text{true}}|u, v) = p(y|u, v)$ for all u and v. The interpretation of the ITE is only causal if the no unmeasured confounders assumption (NUCA) is made; otherwise, it is just a statement about the difference in outcome probability given a new patient described by (u, v).

If we have a correctly specified model and NUCA holds, for any new patient, we have their ITE that guides our treatment choice. This statement is notably population-distribution free and thus can generalize to arbitrary population distributions of (u, v). The ATE does not have this characteristic; its calculation is dependent on the distribution of (u, v) so its application should be limited to populations with similar covariate distributions unless the treatment effect is believed to be uniform.

Recalling that the application of the RCT-recommended treatment suggests that every patient should receive that treatment, a logistic-regression-based model similarly provides a uniform decision. Its decision will be in agreement with the sign of the treatment parameter. However, in many cases, and particularly in those where the treatment effect has small magnitude but high variance, the optimal treatment decision is nonuniform. Thus, we adopt machine learning methods which can estimate the CPD while also providing nonuniform treatment choices. In particular, we use AdaBoost because it has consistency results and is a non-parametric learning algorithm (Freund and Schapire, 1996; Culp et al., 2006). In other words, AdaBoost will recover the correct CPD given enough examples (consistent), and will do so regardless of the training (u, v) distribution provided proper support (non-parametric).

With the adoption of a non-parametric learning algorithm comes the parametric/non-parametric learning trade-off. Parametric models may require smaller sample sizes to learn effectively but are not consistent if misspecified; non-parametric models may require larger sample sizes to

achieve good CPD estimates but have arbitrary joint distribution consistency results.

Propensity-score matching (PSM) and (stabilized) inverse probability-of-treatment weighting ((s)IPT-W) are methods to produce pseudo-randomized data for the estimation of the ATE (Austin, 2011; Rosenbaum and Rubin, 1983; Robins et al., 2000). With ITE as the target statistic, these methods become less desirable. In modeling the CPD, PSM and IPT-W weighting reduce the effective sample size, reducing our numbers for estimation. Thus under the modeling assumption and the goal of modeling ITE, we argue for unweighted CPD estimation. Table 9.1 compares and summarizes the advantages and disadvantages of study methods related to ATE and ITE estimation.

9.4 Experimental Approach

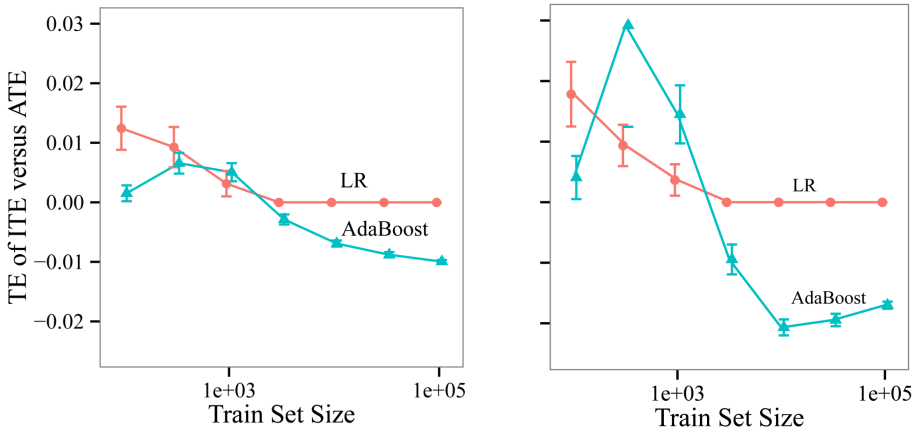
In this section, we restate the claims and reasoning in support of the individualized risk framework and provide ways to confirm them experimentally using synthetic data with access to ground truth, or using observational or RCT data.

Table 9.1: Discussed methodologies with positive and negative characteristics in green and red respectively. ATE is average treatement effect, ITE is individualized treatment effect, RCT is randomized controlled trial, PSM is propensity-score matching, (s)IPT-W is (stabilized) inverse probability-of-treatment weighting, LR is logistic regression, CPD is conditional probability distribution, and NUCA is no unmeasured confounders assumption.

Topic	Negative	Positive
ATE	applicability, generalizability	clinical trial gold standard
ITE	hard to estimate in RCT	applicability, generalizability
RCT	impractical	balanced confounders
PSM, (s)IPT-W	NUCA, decreased effective sample size	pseudo-randomized
LR	uniform treatment recommendation	log odds interpretation
CPD	potential outcomes follow model	mature machine learning

As already noted, there is a strong argument for estimating the ITE over the ATE because the ITE is applicable for patient-specific recommendations as opposed to ATE-based, population-average recommendations. With correct specification and NUCA, the ITE is also generalizable to arbitrary population distributions, though it is harder to estimate than the ATE. The value of the ITE recommendation can be compared against an alternative–for example, ATE recommendation–using the subset of randomized patients where treatment recommendations are the same (Vickers et al., 2007). With these methods, we test the hypothesis of ITE superiority and illustrate the benefits of ITE estimation on synthetic data.

We suggest that, in preference for generalizability of study outcome, the conditional probability distribution $p(y|u, v)$ should be modeled with non-parametric learning algorithms. That is, our goal should be to learn the correct $p(y|u, v)$ irrespective of the distribution $p(u, v)$ because future data distributions $p'(u, v)$ may be different. Non-parametric learning algorithms achieve independence from $p(u, v)$ in the limit of increasing data. We use a synthetic dataset to empirically characterize the recovery of the ITE varying the training set data size and compare the performance of parametric and non-parametric learners varying the similarity of train and test set population distributions. We also use a real RCT dataset to compare the treatment assignment policies of parametric and non-parametric learners in the presence of a substantial average treatment effect. Additionally, we use synthetic data to compare ITE estimation generalizability for parametric and non-parametric learners when the test set distribution $p(u, v)$ varies from the training set distribution, though the conditional probability distribution $p(y|u, v)$ remains the same. We also show experimentally that estimating $p(y|u, v)$ directly from the original data distribution outperforms analogous estimators from propensity-score-weighting and similarly to stabilized inverse probability-of-treatment weighting.



(a) Entire population

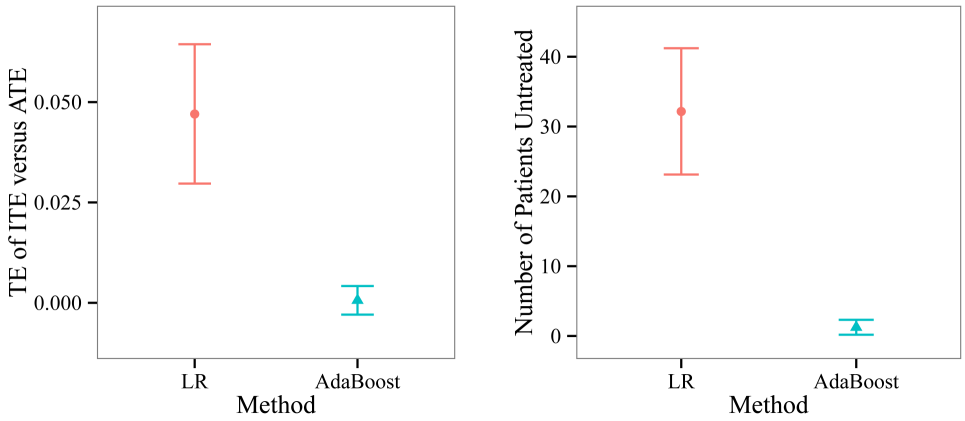
(b) Differing recommendations

Figure 9.1: Average difference in treatment effect using the ITE recommendation in place of the ATE recommendation as a function of training set size. The estimated difference in the population is shown in 9.1a; the estimated difference in the subpopulation where treatment recommendations differ is shown in 9.1b. The difference in treatment effect is estimated by the Vickers et al. (2007) method with a test set of 100,000 examples. The 95% confidence intervals are shown calculated over 100 replications with different training sets.

9.5 Methods

For our experiments using synthetic data, we use the synthetic model described in Section 4.5. Refer to Figure 4.5 and Table 4.5 to review details about the structure and conditional probability distributions of this model. Briefly, it is a synthetic model of myocardial infarction (MI) with thirteen binary variables: age, gender, smoking status, HDL level, LDL level, diabetes, family history of cardiovascular disease (CVD), blood pressure, history of angina, history of stroke, history of depression, statin use, and MI. We simulate both observational and RCT data from this model.

The question we seek to answer for our synthetic model is the effect of



(a) Entire population ITE vs ATE

(b) Patients untreated

Figure 9.2: Average difference in treatment effect using the ITE recommendation in place of the ATE recommendation on the PBC dataset. The estimated difference in the population is shown in 9.2a; the number of patients given a recommendation to not treat is shown in 9.2b. The 95% confidence intervals are shown calculated over 100 replications with different sampled training sets.

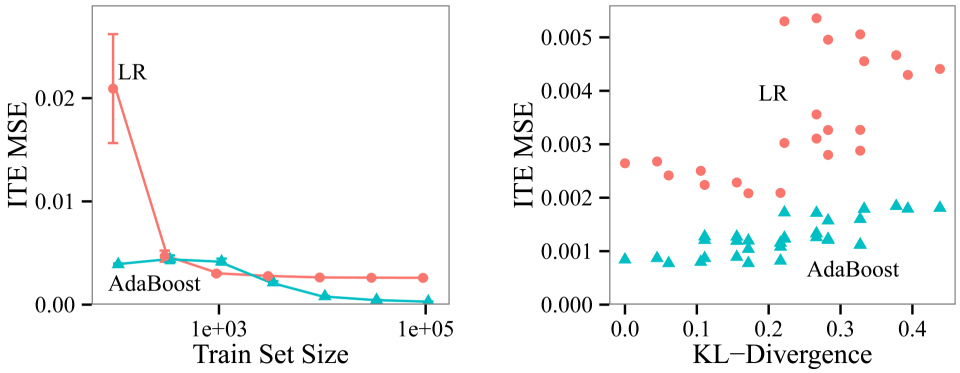
statin use on heart attack or MI. We test the per-patient recommendations for or against statin use from logistic regression and boosted trees against the ATE recommendation of always prescribing statins. Testing uses data generated from our synthetic, randomized distribution and evaluation is performed using both the RCT method of Vickers et al. (2007) and by comparing the predicted ITE to the ground truth ITE calculated exactly from the Bayesian network. Unless otherwise specified, train and test data are generated from the RCT distribution where the LDL to statin edge is removed from the network. When learning our AdaBoost and logistic regression models, we need to ensure the intermediate and confounder assumptions described in Section 9.2 are met. Since we are using synthetic data, there are no unobserved confounders outside our model. However,

diabetes is on a causal pathway from statins to MI, so we exclude it from the features available to our models.

With our synthetic model we seek to characterize estimation of the ITE for each method by looking at error modes of each model and producing learning curves for the models as a function of training set size. To test for applicability to an arbitrary test population distribution, we alter the test distributions by changing CPDs for variables with no parents in the causal graph. Finally, to evaluate ITE estimation from observational data, we use training set data from the observational Bayesian network and compare the estimation from the unweighted training set with estimation using altered datasets via propensity-score matching and stabilized inverse probability-of-treatment weighting.

To validate our claims on real data, we run experiments using the trial data for treatment of primary biliary cirrhosis (PBC) described in Section 4.6. Briefly, the dataset includes 16 variables, including demographic information like age and sex, as well as various lab tests such as serum albumin, serum cholesterol, and triglycerides (refer to Table 4.6 for more detail). The question we seek to answer for this RCT dataset is the effect of D-penicillamine use on three-year survival. For the three-year survival period, the dataset is censored to 288 patients, with 146 in the treatment group and 142 in the placebo group. At the end of three years, the treatment group experienced 27 deaths out of 146, whereas the placebo group experienced 32 out of 142 (see Table 4.7). The trial thus demonstrates an average treatment effect of around a 4 percentage point reduction in death rate over three years.

With the PBC trial data, we compare the estimation of the ITE for each method against the ATE recommendation to treat all patients. Furthermore, given the strength of the average treatment effect, we compare the number of times each method suggests that a patient receive no treatment as opposed to the ATE recommendation. We estimate the average ITE



(a) ITE learning curves (b) ITE test distribution generalization

Figure 9.3: Learning curves for logistic regression and AdaBoost showing test set ITE mean-squared error. The 95% confidence intervals are calculated from 100 replications. Test set ITE MSE is shown as a function of training set size in 9.3a and as a function of KL-divergence between the training set distribution and test set distribution in 9.3b.

versus ATE and number of untreated patients for each method by running 100 bootstrap replicates.

We use the 'ada' package implementation of AdaBoost in R to learn the boosted forest (Freund and Schapire, 1996; Culp et al., 2006). Though the consistency guarantees for AdaBoost require the number of iterations to grow linearly with the training set size (Bartlett and Traskin, 2007), we use the square root of the training set size to reduce the computational burden. Otherwise, we use the default settings from the 'ada' package. Logistic regression models are trained using the 'glm' function of R.

9.6 Results

Figure 9.1 shows the utility of adopting the ITE recommendation over the ATE recommendation on our synthetic model. We want to lower the

risk of MI, so a negative difference between ITE and ATE is desirable. From Figure 9.1a, we see that the adoption of the ITE recommendation, as calculated from the AdaBoost model, lowers the probability of MI by 0.01 on average, provided sufficient training data. That is, the number needed to treat is about 100, so treating 100 patients with the ITE-recommended treatment results in one less MI on average than the ATE-recommended treatment of giving everyone statins. Only AdaBoost is able to provide an improved recommendation because of its ability to accurately estimate the conditional probability distribution. Since there are no interaction terms in our logistic regression model, the recommendation converges to the ATE recommendation of giving everyone statins, resulting in the observed difference of 0 for larger training set sizes in Figure 9.1.

Figure 9.1b shows the estimated expected difference in probability of MI between ITE- and ATE- recommended treatments among patients where the recommendations disagree on treatment choice. We see that the ITE recommendation lowers the probability of MI in this subset by about 0.02, or a NNT of 50. The upturn for AdaBoost as the training set size approaches 100,000 is likely due to correctly identifying more patients with small benefits from not taking statins. This dilutes the ITE- and ATE-difference among those patients where the recommendations disagree, but the population-wide probability of MI, which is the primary value of interest, continues to decrease.

Figure 9.2 shows the utility of adopting the ITE recommendation over the ATE recommendation on the PBC trial data. We want to lower the rate of death over a three-year period, so a negative difference between ITE and ATE is desirable, just as it is with our synthetic model. In Figure 9.2a we see that neither the AdaBoost model nor the logistic regression model outperform the ATE recommendation to treat all patients. While we would prefer to see the ITE outperforming the ATE, this result is not altogether surprising given the effectiveness of treatment. We do, however,

see that the AdaBoost model roughly matches the ATE and outperforms the logistic regression model as we hypothesize and demonstrate with our synthetic data. In Figure 9.2b we show the average number of patients for which each model recommends no treatment. The AdaBoost model rarely recommends no treatment, showing that it has effectively learned the treat-all policy, whereas the logistic regression model frequently recommends no treatment.

Learning curves for logistic regression and AdaBoost are shown in Figure 9.3a. These curves show mean-squared error of the ITE predictions versus training set size. As we expect, the parametric logistic regression converges to a non-zero error because the model is misspecified (since the ground truth model is not log-linear in the exposure and covariates). The error of AdaBoost, however, continues to decrease towards 0 as training set size increases, showing very accurate estimation of the ITE is possible with sufficient data. AdaBoost's approach toward 0 error is in line with the non-parametric consistency results of Bartlett and Traskin (2007).

We also investigate the generalizability of ITE predictions for AdaBoost compared to logistic regression. We simulate applying results to different populations by adjusting the prevalence of the five variables in our Bayesian network (refer back to Figure 4.5) with no parents: age, gender, HDL, depression, family history of CVD. We train on our default RCT data, but test on modified RCT data with different prevalences of the aforementioned variables. Thus, we change $p(v)$, but $p(y|u,v)$ remains the same, so an accurate prediction of CPD, which is exactly $p(y|u,v)$, should handle the changing test distribution gracefully. Figure 9.3b shows that the MSE does tend to increase for both logistic regression and AdaBoost as the KL-divergence between the train and test distribution increases. However, AdaBoost degrades more slowly, demonstrating that learning a non-parametric, consistent model provides better generalization to other populations.

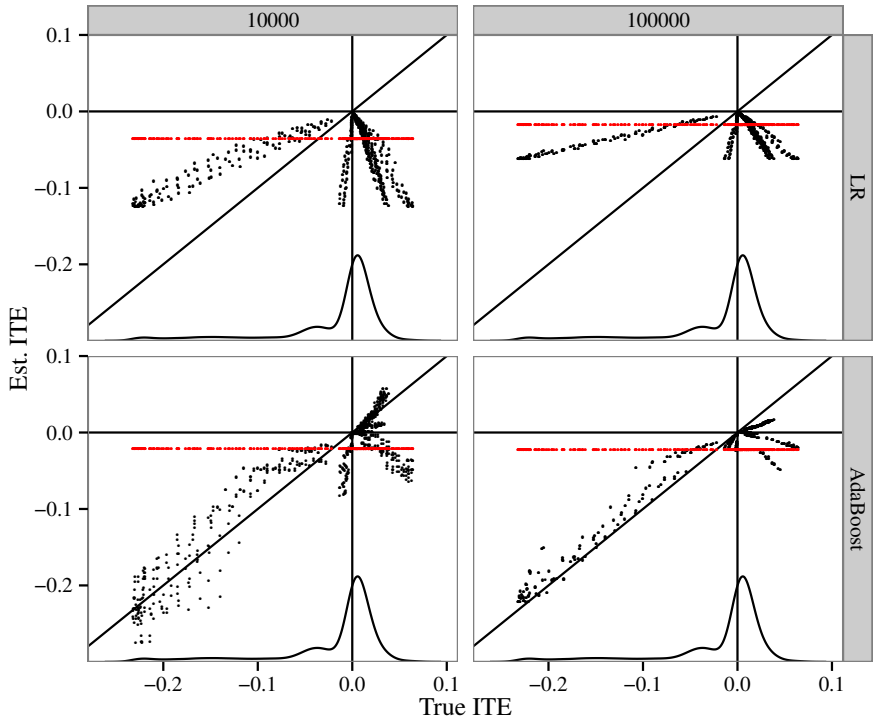


Figure 9.4: Estimated ITE (black) and ATE (horizontal red) versus the ground truth ITE for logistic regression (top) and AdaBoost (bottom) for training set sizes of 10,000 (left) and 100,000 (right). Optimal estimation (*i.e.*, mean-squared error of 0) is given by the line $y = x$. Smoothed, empirical density of the true ITE is shown at the bottom of each plot.

To further investigate the errors in ITE prediction, we show the predicted ITE versus the ground truth ITE (as calculated from our Bayesian network) in Figure 9.4. Additionally, we plot the ATE which effectively predicts an identical treatment effect for all patients. The goal is to have predictions as close as possible to the true value, *i.e.*, to have points as close to the $y = x$ line as possible. In agreement with the results of Figure 9.3a, AdaBoost makes better ITE predictions than logistic regression and improves noticeably from 10,000 to 100,000 examples. For logistic

regression (top), all ITE estimates will be above or below the line $y = 0$ because the model assumes that a single coefficient determines the direction of the effect. The four groupings of points extending down (at various angles) from the origin correspond to various settings of the variables LDL, HDL, and smoking. This suggests that including interaction terms among statins, LDL, HDL, and smoking in the logistic regression would improve its performance. AdaBoost has the capability to learn arbitrary interactions and can provide individualized recommendations, though the errors remain greater than zero as shown in the bottom of Figure 9.4. Indeed, some of the groupings of points lie off the $y = x$ line also correspond to certain patient subpopulations for which the ITE is consistently misestimated. We expect, due to the consistency of AdaBoost, that these errors will decrease as more training data is available.

The effect of different data-weighting and matching schemes is shown in Figure 9.5. The recovery of the CPD model, and thus the ITE, requires the fewest examples when leaving the examples unweighted or using stabilized inverse weighting. While propensity score matching produces worse estimates of ITE, there is no benefit for using stabilized inverse weighting over no weighting for this task. One important consideration is that our dataset includes some patients without elevated LDL who take statins, motivated by the suggestion that there could be therapeutic benefit of statins even in borderline hypercholesterolemia. However, in a dataset with fewer normal-LDL statin users, propensity-score matching and particularly stabilized inverse weighting will impair the CPD model, because it will attach large excess weight to few examples.

9.7 Discussion

In this work, we illustrated the parallels between the standard clinical study framework designed to determine the ATE and the burgeoning

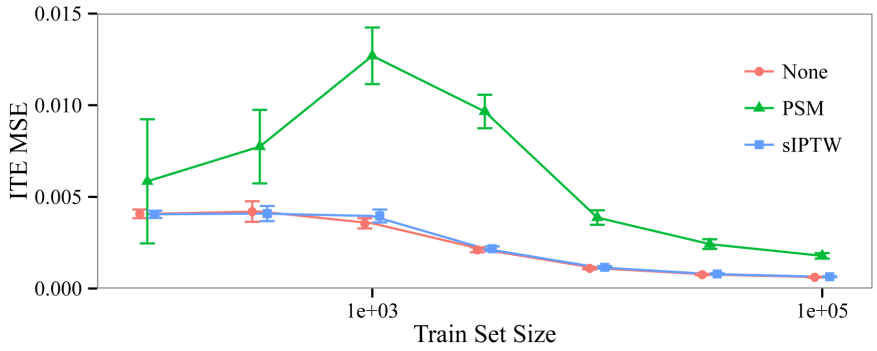


Figure 9.5: Learning curves for AdaBoost trained on observational data. Test set ITE mean-squared error as a function of training set size is shown comparing training from unweighted examples (None), propensity-score matched samples with a 1:1 ratio (PSM), and stabilized inverse probability of treatment weighted examples (sIPTW). The 95% confidence intervals are calculated over 100 replications.

clinical study framework designed to determine the ITE. We highlighted shortcomings of the ATE; first, that the ATE is an average outcome, when in practice we usually care about the ITE for future patients, and second, that the ATE is population-distribution dependent. We then discussed modeling of the ITE. Notably, logistic regression can only recommend one treatment arm if we exclude non-linear and exposure-covariate interaction terms because the coefficient for exposure is either negative or non-negative. Furthermore, unless correctly-specified, logistic regression is not a consistent learning algorithm, so we cannot always recover the true conditional probability distribution, even from large populations. Instead, we adopted another popular framework, boosted trees. We showed that the forest-based ITE outperformed the ATE on a synthetic problem of MI prediction using binary variables, and that the forest-based ITE outperformed logistic regression-based ITE on a real problem of primary biliary cirrhosis treatment with D-penicillamine. Additionally, we showed that the forest-based ITE better generalized to different test set distribu-

tions than the logistic regression-based ITE. Finally, we showed that the use of propensity-score matching and inverse-probability-of-treatment weighting failed to improve the learning of the conditional probability distribution, suggesting that unweighted samples should be used for learning a model of the ITE. Modeling of the ITE has large theoretical advantages, though robustness guarantees and validation of its performance should be established before large-scale clinical deployment. A few sources of validation include replication studies and heterogeneity of treatment effect analyses using ITE model strata.

10 OTHER EXPERIMENTS AND FUTURE WORK

In this chapter we present some ideas for future directions of investigation and preliminary experiments working toward them. Note that none of this work has been peer-reviewed and published, and some is still in active development.

10.1 Uplift Bayesian Networks

While some of our work has used Bayesian networks in uplift modeling (see Chapter 7), this work was accomplished by learning separate models for the differential strata. No method has been developed to perform single-network structure learning for uplift modeling in a way that is computationally efficient and scalable to much larger datasets. Furthermore, no method has been developed to perform single-network parameter learning to maximize uplift.

To address the challenge of structure learning, we have begun initial steps with a differential Tree-augmented Naïve Bayes (TAN) model. Recall from Section 2.1.2 that TAN relaxes the simplifying assumption made by Naïve Bayes (NB), namely conditional independence between all predictor variables given the class. TAN allows each predictor to be dependent on one other variable other than the class and does so by constructing a maximum-weight spanning tree amongst the predictor variables. The edge weights that TAN uses to construct the tree are the conditional mutual information between variables, but the weights could hypothetically be replaced with another function to suit the needs of uplift modeling.

We replace the edge weights with a differential conditional mutual information between the variables. That is, the difference in conditional mutual information between the two variables, based on the separate

treatment and control strata in the dataset.

$$I_{DIFF}(A; B|Class) = I_{treatment}(A; B|Class) - I_{control}(A; B|Class)$$

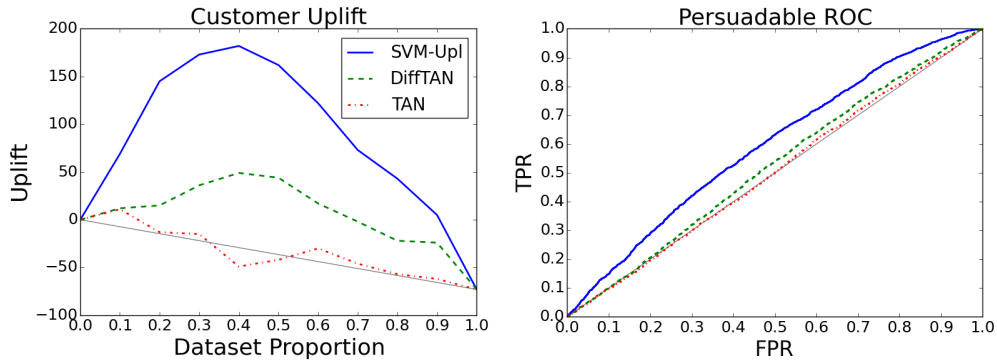
This function maintains the TAN property that the structure can be learned efficiently. It does not, however, maintain the property of producing the maximum likelihood tree, given the data, nor does it make any guarantees about maximizing uplift.

We implemented this model in Weka and have run experiments on the simulated customer dataset described in Section 4.4. For comparison, we ran three algorithms: SVM^{Upl}, TAN, and our newly define differential TAN (DiffTAN). We used 10-fold cross-validation for evaluation and use a Wilcoxon signed rank test to compare area under the uplift curve (AUU), as well as area under the ROC curve (AUC), as compared to SVM^{Upl}. Like in the customer simulation experiments before (see Section 8.5.1), the ROC curves are calculated with the hidden *Persuadable* customer type treated as the positive class. This allows us to see how well the models identify the relevant latent information. Results are shown in Table 10.1 and final curves are shown in Figure 10.1.

Table 10.1: Areas under the ROC curve (AUC) and areas under the uplift curve (AUU) for all three models. 10-fold cross-validation p-values are shown for comparison of both TAN models to SVM^{Upl}. Statistically significant differences are marked with *.

Algorithm	AUC	AUC p-value	AUU	AUU p-value
DiffTAN	0.522	0.002*	13.472	0.002*
TAN	0.503	0.002*	-33.649	0.002*
SVM^{Upl}	0.593	-	95.726	-

Table 10.1 shows DiffTAN does not perform statistically better than SVM^{Upl}. DiffTAN does appear to produce greater uplift than TAN, however, and also demonstrates a slight improvement in identifying *Persuad-*



(a) Customer uplift.

(b) Persuadable ROC.

Figure 10.1: ROC curves and Uplift curves for all three models. The ROC curves are constructed with the hidden *Persuadable* customer type considered to be the positive class. The higher the ROC curve, the better the model is at capturing relevant latent information about customer types.

ables over TAN (see Figure 10.1). Overall, DiffTAN may be useful for uplift modeling, but it does not perform as well as the state-of-the-art.

10.2 Net Benefit Maximization

Uplift modeling has been a large portion of our work so far, and it shows potential for applications in medicine. One flaw, however, is that there is still fairly limited evidence that it delivers on the promise to identify the latent factors that cause individualized response. Vickers et al. (2007) presented the *net benefit* method for evaluating individualized treatment effect (ITE) models on randomized controlled trial (RCT) datasets by evaluating on patients where the treatment recommendations were the same (see Figure 10.2). Certainly, uplift modeling approaches should be evaluated on RCT datasets in the future with this evaluation method in order to demonstrate improvement over the average treatment effect (ATE) recommendation. Having such an evaluation method though, brings another possibility to

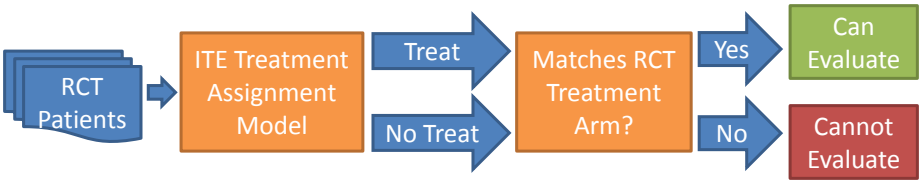


Figure 10.2: The Vickers method of evaluating ITE predictive models on randomized controlled trial (RCT) data. When an individualized treatment effect (ITE) model recommends the same treatment as was assigned in the RCT, the outcome of the recommendation is known and can be used for evaluation.

the table. If net benefit is a good measure of the clinical objective, new machine learning approaches should be developed to optimize for that metric directly.

In order to train models to improve net benefit, we might use any number of algorithms and select the best parameters for net benefit through internal cross-validation. Because we want a model that directly optimizes net benefit, we have chosen to use artificial neural networks (ANN) for their flexibility. We do not have a differentiable error function for net benefit though, so we cannot use a standard optimization approach like gradient descent. Instead, we have implemented a genetic algorithm approach (see Section 2.3) to train our ANNs. We use a simple algorithm to train a model for each fold (see Algorithm 2.1). Crossover takes two ANNs and produces a child ANN with the same structure as the parents, but each weight value is chosen randomly from one of the parents. See Figure 10.3 for an example of how weight crossover between parents works for the outgoing weights of a single node in a network. Once a final model has been selected, we evaluate the net benefit of the model on the test set.

We ran preliminary experiments using the RCT dataset for the treatment of primary biliary cirrhosis (PBC) described in Section 4.6. Recall that we desire to estimate the effect of D-penicillamine use on three-year survival. Instead of k-fold cross-validation, we used bootstrap sampling to

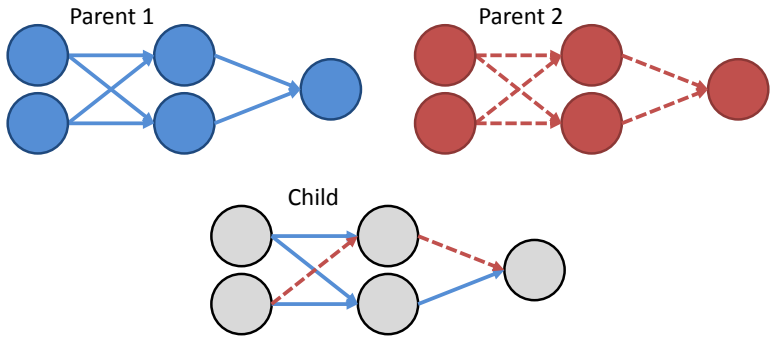


Figure 10.3: Crossover between two parents for a simple artificial neural network (ANN). The child network has the same structure, but inherits weights randomly from the parents.

Table 10.2: Individualized treatment effect (ITE) model average reduction in death rate over three years, as well as the difference from the average treatment effect (ATE) recommendation to treat all (negative is better). Neither trained model is superior to the ATE here.

Algorithm	Death Reduction	ITE vs ATE
Logistic Regression	-0.037 ± 0.061	0.008 ± 0.030
Net Benefit ANN	-0.015 ± 0.049	0.030 ± 0.044
Treat All (ATE)	-0.045 ± 0.063	-

run 100 experimental folds. We specified a maximum number of epochs of 100, a population size of 100, a selection size of 50, a tuning set of 50% of the training set, and an early stopping condition of five epochs with no improvement in net benefit. We compare our net benefit ANN with the default Scikit-learn logistic regression model (Pedregosa et al., 2011), as well as the recommendation to treat all patients. Table 10.2 shows the mean reduction in death rate and the difference from the ATE recommendation to treat all patients.

This method does demonstrate the ability to achieve a positive net benefit in our preliminary experiments, but it does not currently outperform logistic regression or the recommendation to treat all patients. It is too

early to judge this ongoing work, but the results are promising enough to warrant further investigation.

10.3 Model Calibration

Much of our work relates to close collaboration between machine learning experts and clinical experts. We argue that this practice is necessary to develop machine learning models that meet clinical objectives, and can be translated into real practice. One less-obvious problem that we have not yet addressed, however, is model calibration. Consider Chapter 6 wherein we develop a Naïve Bayes (NB) model to predict the probability of a discordant core needle biopsy being benign. This is valuable because providing physicians with a probabilistic risk of upgrade is intuitive and may help facilitate discussions between the physicians and their patients. However both physician and machine estimated probabilities can be inaccurate. Without calibrated probabilities, a clinician cannot be certain that a model prediction of 90% risk really means 90%. In particular, NB is known to accurately predict the most probable label, but its predicted probabilities are not well calibrated (Zadrozny and Elkan, 2001).

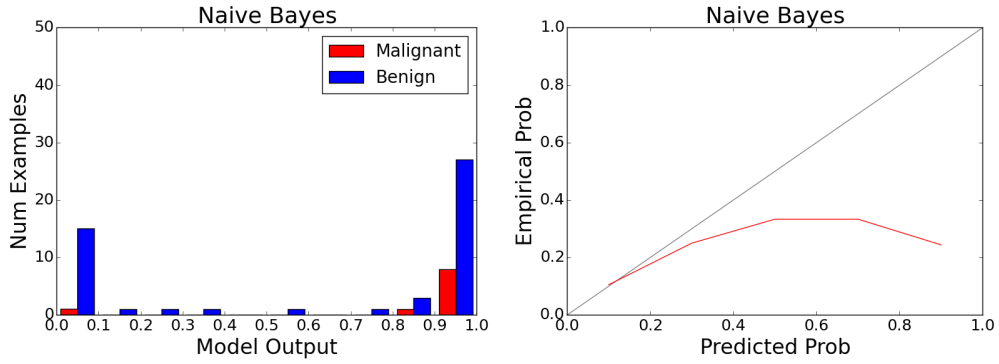
Figure 10.4 demonstrates the good discrimination and poor calibration produced by NB. Note the bimodal distribution of outputs of Figure 10.4a. All of the output "probabilities" are focused near 0.0 and 1.0. Given the marked benign skew of the dataset, a unimodal output distribution near 0.0 might be more appropriate. Note also the poor calibration in Figure 10.4b. Of the cases that the model scored in the 0.8 to 1.0 range, only 20% are positive resulting in a domed calibration curve. A well-calibrated model would exhibit a calibration curve that falls on the diagonal.

There are reliable methods for calibration already available, such as binning (Zadrozny and Elkan, 2001), but the small size of our datasets necessitates reducing the number of bins used, thereby reducing the gran-

ularity of predictions that the learned model can make. If using five bins, the calibrated model can only produce five different predicted probabilities. This is less useful in practice as it disallows finer discrimination of cases.

In a first attempt to address this issue, we have used a bootstrap sampling approach to build calibration models (see Algorithm 10.1). We used the same folds, data, and model parameters presented in Chapter 6. For each fold, we drew 100 bootstrap samples from the training set. For each bootstrap sample, we trained a model and recorded the predictions on the out-of-bag examples. We pooled all of the bootstrap predictions and built a binning calibration model with 100 bins. Next, we trained the main fold model on the training set, made predictions on the test set, and calibrated the test set predictions using the calibration model. The results are shown in Figure 10.5.

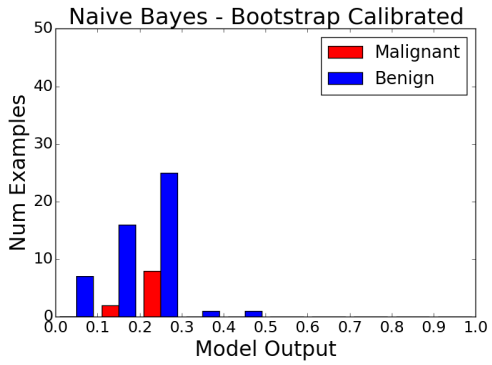
The output probabilities of the calibrated model in Figure 10.5a look somewhat more reasonable than the uncalibrated probabilities in Figure 10.4a because they are more focused near 0.0. The scaled calibration



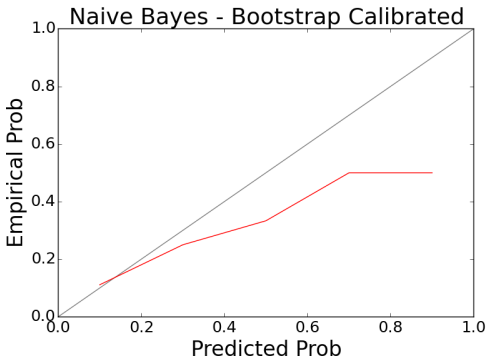
(a) Model output histogram.

(b) Model calibration curve.

Figure 10.4: Calibration figures of published Naïve Bayes model for upgrade prediction (Kuusisto et al., 2015).

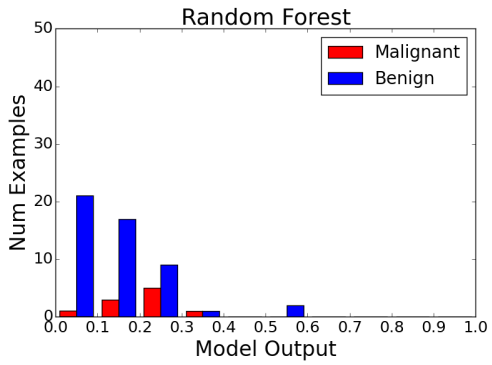


(a) Model output histogram.

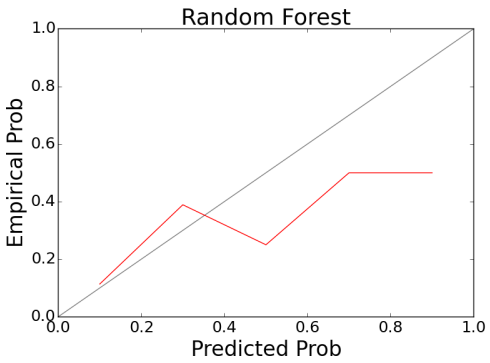


(b) Model calibration curve.

Figure 10.5: Bootstrap calibrated Naïve Bayes.



(a) Model output histogram.



(b) Model calibration curve.

Figure 10.6: Calibration figures for Random Forest on the upgrade prediction data. These are the results for a standard Random Forest implementation without any extra steps taken to calibrate the model.

curve in Figure 10.5b shows that the model is slightly improved, but still poorly calibrated. Perhaps calibration can be further improved by more sophisticated approaches, but it may also be more reasonable to investigate other models that tend to be better calibrated in first place. Random Forests are one such option. Figure 10.6 shows the analogous histogram and curves

Algorithm 10.1 Bootstrap Calibration Procedure

```
Input: N ← # bootstrap samples, B ← # calibration bins;
for Train, Test ∈ Folds do
    CalPreds ← {};                                  ▷ Init bootstrap predictions
    for TuneTrain, TuneTest ∈ BootstrapSample(Train, N) do
        TuneM ← TrainNB(TuneTrain);                 ▷ Train bootstrap model
        TuneP ← Predict(TuneM, TuneTest);           ▷ Test bootstrap model
        CalPreds ← CalPreds ∪ TuneP;                ▷ Store bootstrap predictions
    end for
    CalM ← BuildCalibModel(CalPreds);               ▷ Build calibration
                                                      model from boot-
                                                      strap predictions
    M ← TrainNB(Train);                             ▷ Train main model
    TestP ← Predict(M, Test);                       ▷ Get test predictions
    CalTestP ← Calibrate(CalM, TestP);              ▷ Calibrate test predic-
                                                      tions
    Evaluate(CalTestP);                 ▷ Evaluate calibrated test predictions
end for
```

without the use of any extra calibration methods. Notice that the Random Forest approach looks more similar to the calibrated NB model in Figure 10.5 than to the uncalibrated model in Figure 10.4.

The goals of the precision medicine initiative urge the development of new tools and technology that allow us to capitalize on the increasing collection of heterogeneous medical data. I believe that machine learning is going to play an important role in this domain going forward, but current methods cannot simply be applied as is. Machine learning researchers need to adapt to the domain. Novel models and methods need to be developed to address new, challenging objectives.

This thesis presents my work on improving the application of machine learning to precision medicine. In particular, this thesis presents my work on adapting standard machine learning methods to meet challenging clinical objectives by working in close collaboration with clinicians and leveraging their expertise. This thesis also presents my work on developing novel uplift modeling algorithms and applying them to the challenging tasks of treatment assignment and understanding of risk factors.

11.1 Contributions

In this section, I summarize my contributions to the areas of clinical decision support and applications of machine learning to medicine. While some of this work has led to further investigations (Nassif et al., 2013b; Elezaby et al., 2015; Gegios et al., 2015), this section only summarizes my contributions detailed in this dissertation.

11.1.1 Clinical Decision Support

Machine learning models will never translate to clinical use if they do not meet clinical objectives. My work in decision support has contributed

to the adaptation of machine learning models to real clinical problems through collaboration with clinical experts.

In Chapter 5 we adapted an inductive logic programming (ILP) learner to the challenging task of identifying truly benign cases amongst non-definitive breast biopsies (Kuusisto et al., 2013). Instead of relying on the standard rule evaluation functions available to us, we made use of an F_β measure. Based on our direct interactions with clinicians, we chose a β parameter to prefer rules that do not misclassify malignant cases. Through this work, we were able to infer rules that both adhered to the clinical objective and provided insight into our motivating task.

In Chapter 6 we presented a framework for collaboration between clinical and machine learning experts (ABLe) (Kuusisto et al., 2015). The framework we presented, ABLe, defined an iterative process that can be used to iteratively improve clinical machine learning models until they meet the objective of interest. Taking lessons from our previous work, we used this framework to leverage expert clinical advice to account for the challenge of limited data. We demonstrated the application of ABLe to the upgrade prediction task and showed how it can successfully be used to meet the exceedingly conservative objective of missing no malignant cases.

11.1.2 Treatment Effect Estimation and Understanding

Estimating the effects treatment and exposure is challenging. My work takes lessons from the marketing domain, and applies uplift modeling to this task. My work in uplift modeling has contributed to the development of new algorithms and the application of uplift modeling to medicine for knowledge discovery and treatment effects estimation.

In Chapter 7 we developed a novel statistical relational uplift modeling algorithm (Nassif et al., 2013a). We applied this new approach to a breast cancer task involving the differences between in situ and invasive cancers.

Knowing that older patients tend to have more indolent in situ cancers than younger patients, we mapped older and younger patients to separate target and control group concepts respectively. We then used this model to try to understand what factors contribute to the indolent nature of older in situ breast cancer, and we found that our approach discovers themes that are clinically interesting.

In Chapter 8 we developed a support vector machine (SVM) model that optimizes for area under the uplift curve (AUU) Kuusisto et al. (2014). We first used this approach to validate that building uplift models to maximize the AUU captures relevant latent information about the differences between the target and control groups. To do so, we generated a synthetic dataset of customers and simulated a marketing campaign. Because the dataset was synthetic, we knew true customer types. We showed that our algorithm is better able to identify the *Persuadable* customer group than other baseline comparisons. We also applied the model to the task of identifying patients who are susceptible to risk of heart attack from taking COX-2 inhibitors. While we do not know the true individual susceptibilities in this real dataset, we show that we are able to produce significantly greater uplift on this dataset than other methods. We also applied the model to the task of identifying the factors associated with in situ breast cancer specific to older women, just as we did in Chapter 7. In this task, we again found that our algorithm was capable of discovering clinically-relevant features.

In Chapter 9 we discussed the importance of investigating new models to estimate individualized treatment effects and make superior treatment recommendations to that of the average treatment effect (ATE) (Weiss et al., 2015). We did not develop new algorithms, but we argued for the use of machine learning methods in favor of more traditional methods like logistic regression. We supported this argument by showing how AdaBoost can improve upon ITE estimation over logistic regression on

both a synthetic dataset, as well as a real dataset.

11.2 Summary

In my thesis, I present my contributions at the intersection of machine learning and medicine. I also present suggestions for directions of future work, along with experiments detailing initial steps in those directions. My work demonstrates that close collaboration between clinicians and computer scientists is key to the success of machine learning in precision medicine. First, leveraging the expertise of clinicians can help to alleviate challenges of gathering sufficient standardized data to model important, but relatively rare diseases. Second, close collaboration is essential to develop models that actually meet clinical objectives, instead of relying on standard machine learning objectives to meet those needs. Additionally, my work demonstrates advances in modeling individualized response to treatment, using multiple potential approaches. To accomplish this, I demonstrate the potential of leveraging and expanding upon uplift modeling approaches from marketing. In general, my work has contributed to the advancement of machine learning for use in medicine.

2003. *Breast Imaging Reporting and Data System (BI-RADS™)*. American College of Radiology, Reston, VA, USA, 4th ed.

Abbey, C, M Eckstein, and J Boone. 2013. Estimating the relative utility of screening mammography. *Medical Decision Making* 33(4):510–520.

American Cancer Society. 2009a. *Breast Cancer Facts & Figures 2009-2010*. Atlanta, USA: American Cancer Society.

———. 2009b. *Cancer Facts & Figures 2009*. Atlanta, USA: American Cancer Society.

Ash, J, J McCormack, D Sittig, A Wright, C McMullen, and D Bates. 2012. Standard practices for computerized clinical decision support in community hospitals: A national survey. *Journal of the American Medical Informatics Association* 19(6):980–987.

Austin, P. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46(3):399–424.

Bang, H, and J Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4):962–973.

Bartlett, P, and M Traskin. 2007. Adaboost is consistent. *Journal of Machine Learning Research* 8:2347–2368.

Bevers, T, B Anderson, E Bonaccio, S Buys, M Daly, P Dempsey, W Farrar, I Fleming, J Garber, R Harris, et al. 2009. Breast cancer screening and diagnosis. *Journal of the National Comprehensive Cancer Network* 7(10):1060–1096.

Blockeel, H, and L De Raedt. 1998. Top-down induction of first-order logical decision trees. *Artificial Intelligence* 101:285–297.

Blumenthal, D. 2010. Launching HITECH. *New England Journal of Medicine* 362(5):382–385.

Bouchardy, C, S Benhamou, G Fioretta, H Verkooijen, P Chappuis, I Neyroud-Caspar, M Castiglione, V Vinh-Hung, G Vlastos, and E Rapiti. 2011. Risk of second breast cancer according to estrogen receptor status and family history. *Breast Cancer Research and Treatment* 127(1):233–241.

Boyd, K, J Davis, D Page, and V Santos Costa. 2012. Unachievable region in precision-recall space and its effect on empirical evaluation. In *Proceedings of the 29th International Conference on Machine Learning, ICML'12*. Edinburgh, Scotland.

Bruening, W, J Fontanarosa, K Tipton, J Treadwell, J Launders, and K Schoelles. 2010. Systematic review: Comparative effectiveness of core-needle and open surgical biopsy to diagnose breast lesions. *Annals of Internal Medicine* 152(4):238–246. Epub 2009 Dec 14.

Burnside, E, J Davis, J Chhatwal, O Alagoz, M Lindstrom, B Geller, B Littenberg, K Shaffer, C Kahn, and D Page. 2009. Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. *Radiology* 251:663–672.

Burnside, E, D Rubin, R Shachter, R Sohlich, and E Sickles. 2004. A probabilistic expert system that provides automated mammographic-histologic correlation: Initial experience. *American Journal of Roentgenology* 182:481–488.

CDC. 1998. *National Vital Statistics Report*. National Center for Health Statistics.

Chang, C, and C Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3):27:1–27:27.

Chaudhry, B, J Wang, S Wu, M Maglione, W Mojica, E Roth, S Morton, and P Shekelle. 2006. Systematic review: Impact of health information technology on quality, efficiency, and costs of medical care. *Annals of Internal Medicine* 144(10):742–752.

Cleary, T. 1968. Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement* 5(2):115–124.

Cohen, I, R Amarasingham, A Shah, B Xie, and B Lo. 2014. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs* 33(7):1139–1147.

Collins, F, and H Varmus. 2015. A new initiative on precision medicine. *New England Journal of Medicine* 372(9):793–795.

Cortes, C, and V Vapnik. 1995. Support-vector networks. *Machine learning* 20(3):273–297.

Culp, M, K Johnson, and G Michailidis. 2006. `ada`: An R package for stochastic boosting. *Journal of Statistical Software* 17(2):9.

Davis, J, E Burnside, I Dutra, D Page, and V Santos Costa. 2005. An integrated approach to learning Bayesian networks of rules. In *Proceedings of the 16th European Conference on Machine Learning*, 84–95. Porto, Portugal.

Davis, J, D Page, V Santos Costa, P Peissig, and M Caldwell. 2013. A preliminary investigation into predictive models for adverse drug events. In *Proceedings of the AAAI-13 Workshop on Expanding the Boundaries of Health Informatics Using AI*. Bellevue, WA.

De Raedt, L. 2008. *Logical and Relational Learning*. Springer.

Destounis, S, M Skolny, R Morgan, A Arieno, P Murphy, P Somerville, P Seifert, and W Young. 2011. Rates of pathological underestimation for 9 and 12 gauge breast needle core biopsies at surgical excision. *Breast Cancer* 18(1):42–50. Epub 2010 Mar 4.

Domingos, P, and M Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29(2-3):103–130.

Dutra, I, H Nassif, D Page, J Shavlik, R Strigel, Y Wu, M Elezaby, and E Burnside. 2011. Integrating machine learning and physician knowledge to improve the accuracy of breast biopsy. In *American Medical Informatics Association Symposium (AMIA)*, 349–355. Washington, DC.

Elezaby, M, F Kuusisto, J Shavlik, Y Wu, I Dutra, H Neuman, W DeMartini, and E Burnside. 2015. Core needle biopsies: A predictive model that identifies low probability ($\leqslant$2%) lesions to safely avoid surgical excision. In *Radiological Society of North America (RSNA) 101st Scientific Assembly and Annual Meeting*. Chicago, IL. Accepted for oral presentation.

Epstein, O, S Jain, R Lee, D Cook, A Boss, P Scheuer, and S Sherlock. 1981. D-penicillamine treatment improves survival in primary biliary cirrhosis. *The Lancet* 317(8233):1275–1277.

FDA. 2015. FDA drug safety communication: FDA strengthens warning that non-aspirin nonsteroidal anti-inflammatory drugs (NSAIDs) can cause heart attacks or strokes.

Forman, G, and M Scholz. 2010. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explorations Newsletter* 12(1):49–57.

Fowble, B, D Schultz, B Overmoyer, L Solin, K Fox, L Jardines, S Orel, and J Glick. 1994. The influence of young age on outcome in early stage

breast cancer. *International Journal of Radiation Oncology, Biology, Physics* 30(1):23–33.

Freund, Y, and R Schapire. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory*, 23–37. Springer.

———. 1996. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*.

Friedman, L, C Furberg, D DeMets, et al. 2010. *Fundamentals of Clinical Trials*, vol. 4. Springer.

Friedman, N, D Geiger, and M Goldszmidt. 1997. Bayesian network classifiers. *Machine Learning* 29:131–163.

Fures, R, D Bukovi, C Lez, M Zadro, N Bukovi, D Smud, and E Giudici. 2003. Large-gauge needle biopsy in diagnosing malignant breast neoplasia. *Collegium Antropologicum* 27(1):259–262.

Gegios, A, M Elezaby, W DeMartini, J Cox, C Montemayor-Garcia, H Neuman, F Kuusisto, J Hampton, and E Burnside. 2015. Differential upgrade rates for non-definitive image-guided core needle breast biopsies based on BI-RADS features. In *Radiological Society of North America (RSNA) 101st Scientific Assembly and Annual Meeting*. Chicago, IL. Accepted for poster presentation.

Getoor, L, and B Taskar, eds. 2007. *An Introduction to Statistical Relational Learning*. MIT Press.

Gibert, K, C García-Alonso, and L Salvador-Carulla. 2010. Integrating clinicians, knowledge and data: Expert-based cooperative analysis in healthcare decision support. *Health Research Policy and Systems* 8(28): 1478–4505.

Gong, Y, S Louise Klingenberg, and C Gluud. 2004. D-penicillamine for primary biliary cirrhosis. *The Cochrane Library*.

Gumus, H, M Gumus, H Devalia, P Mills, D Fish, P Jones, A Uyar, and A Sever. 2012. Causes of failure in removing calcium in microcalcification-only lesions using 11-gauge stereotactic vacuum-assisted breast biopsy. *Diagnostic and Interventional Radiology* 18(4):354–359. Epub 2012 Apr 5.

Guyon, I, J Weston, S Barnhill, and V Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1-3):389–422.

Hall, M, E Frank, G Holmes, B Pfahringer, P Reutemann, and I Witten. 2009. The WEKA Data Mining Software: An update. *SIGKDD Explorations Newsletter* 11(1):10–18.

Hanley, J, and B McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1):29–36.

Hansotia, B, and B Rukstales. 2002. Incremental value modeling. *Journal of Interactive Marketing* 16(3):35–46.

Hastie, T, R Tibshirani, and J Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer.

Hirschfield, G, and M Gershwin. 2013. The immunobiology and pathophysiology of primary biliary cirrhosis. *Annual Review of Pathology: Mechanisms of Disease* 8(1):303–330.

Jaśkowski, M, and S Jaroszewicz. 2012. Uplift modeling for clinical trial data. In *ICML 2012 Workshop on Clinical Data Analysis*. Edinburgh, Scotland.

Jayasinghe, U, R Taylor, and J Boyages. 2005. Is age at diagnosis an independent prognostic factor for survival following breast cancer? *ANZ Journal of Surgery* 75(9):762–767.

Joachims, T. 2005. A support vector method for multivariate performance measures. In *International Conference on Machine Learning (ICML)*, 377–384.

Joachims, T, T Finley, and C Yu. 2009. Cutting-plane training of structural svms. *Machine Learning* 77(1):27–59.

Kawamoto, K, C Houlihan, E Balas, and D Lobach. 2005. Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *BMJ* 330(7494):765.

Kearney, P, C Baigent, J Godwin, H Halls, J Emberson, and C Patrono. 2006. Do selective cyclo-oxygenase-2 inhibitors and traditional non-steroidal anti-inflammatory drugs increase the risk of atherothrombosis? Meta-analysis of randomised trials. *BMJ* 332(7553):1302–1308.

Kent, D, and R Hayward. 2007. Limitations of applying summary results of clinical trials to individual patients: The need for risk stratification. *Journal of the American Medical Association* 298(10):1209–1212.

Koller, D, and N Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Kurian, A, L McClure, E John, P Horn-Ross, J Ford, and C Clarke. 2009. Second primary breast cancer occurrence according to hormone receptor status. *Journal of the National Cancer Institute* 101(15):1058–1065.

Kuusisto, F, I Dutra, M Elezaby, E Mendonca, J Shavlik, and E Burnside. 2015. Leveraging expert knowledge to improve machine-learned decision support systems. In *AMIA Joint Summits on Translational Science*. San Francisco.

Kuusisto, F, I Dutra, H Nassif, Y Wu, M Klein, H Neuman, J Shavlik, and E Burnside. 2013. Using machine learning to identify benign cases with non-definitive biopsy. In *IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 283–285. Lisbon, PT.

Kuusisto, F, V Santos Costa, H Nassif, E Burnside, D Page, and J Shavlik. 2014. Support vector machines for differential prediction. In *European Conference on Machine Learning (ECML-PKDD)*.

Lavrac, N, and S Dzeroski. 1994. *Inductive Logic Programming: Techniques and Applications*. New York: Ellis Horwood.

Liberman, L. 2000. Percutaneous imaging-guided core breast biopsy: State of the art at the millennium. *American Journal of Roentgenology* 174(5): 1191–1199.

Liberman, L, A Abramson, F Squires, J Glassman, E Morris, and D Dershaw. 1998. The breast imaging reporting and data system: Positive predictive value of mammographic features and final assessment categories. *American Journal of Roentgenology* 171:35–40.

Linn, R. 1978. Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology* 63:507–512.

Lo, V. 2002. The true lift model - A novel data mining approach to response modeling in database marketing. *SIGKDD Explorations* 4(2): 78–86.

Lucas, P. 2001. Expert knowledge and its role in learning Bayesian networks in medicine: An appraisal. In *Artificial Intelligence in Medicine*, ed. S Quaglini, P Barahona, and S Andreassen, vol. 2101 of *Lecture Notes in Computer Science*, 156–166. Springer Berlin Heidelberg.

Manning, C, P Raghavan, H Schütze, et al. 2008. *Introduction to Information Retrieval*, vol. 1. Cambridge University Press.

Manson, J, R Chlebowski, M Stefanick, A Aragaki, J Rossouw, R Prentice, G Anderson, B Howard, C Thomson, A LaCroix, et al. 2013. Menopausal hormone therapy and health outcomes during the intervention and extended poststopping phases of the Women's Health Initiative randomized trials. *Journal of the American Medical Association* 310(13):1353–1368.

Mendis, S, P Puska, B Norrving, et al. 2011. *Global Atlas on Cardiovascular Disease Prevention and Control*. World Health Organization.

Mitchell, T. 1997. *Machine Learning*. New York: McGraw-Hill.

Moskowitz, M. 1983. The predictive value of certain mammographic signs in screening for breast cancer. *Cancer* 51:1007–1011.

Muggleton, S. 1995. Inverse entailment and Progol. *New Generation Computing* 13:245–286.

Naci, H, J Brugts, and T Ades. 2013. Comparative tolerability and harms of individual statins: A study-level network meta-analysis of 246 955 participants from 135 randomized, controlled trials. *Circulation: Cardiovascular Quality and Outcomes* 6(4):390–399.

Nakayama, R, Y Uchiyama, R Watanabe, S Katsuragawa, K Namba, and K Doi. 2004. Computer-aided diagnosis scheme for histological classification of clustered microcalcifications on magnification mammograms. *Medical Physics* 31(4):789–799.

Narasimhan, H, and S Agarwal. 2013. A structural SVM based approach for optimizing partial AUC. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, ed. S Dasgupta and D Mcallester, vol. 28, 516–524. JMLR Workshop and Conference Proceedings.

Nassif, H, F Kuusisto, E Burnside, D Page, J Shavlik, and V Santos Costa. 2013a. Score As You Lift (SAYL): A statistical relational learning approach

to uplift modeling. In *European Conference on Machine Learning (ECML-PKDD)*, 595–611. Prague, CZ.

Nassif, H, F Kuusisto, E Burnside, and J Shavlik. 2013b. Uplift modeling with ROC: An SRL case study. In *International Conference on Inductive Logic Programming (ILP)*. Rio de Janeiro, BR.

Nassif, H, D Page, M Ayvaci, J Shavlik, and E Burnside. 2010. Uncovering age-specific invasive and DCIS breast cancer rules using Inductive Logic Programming. In *ACM International Health Informatics Symposium (IHI)*, 76–82. Arlington, VA.

Nassif, H, V Santos Costa, E Burnside, and D Page. 2012a. Relational differential prediction. In *European Conference on Machine Learning (ECML-PKDD)*, 617–632. Bristol, UK.

Nassif, H, Y Wu, D Page, and E Burnside. 2012b. Logical Differential Prediction Bayes Net, improving breast cancer diagnosis for older women. In *American Medical Informatics Association Symposium (AMIA)*, 1330–1339. Chicago.

National Research Council. 2011. *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease*. National Academies Press (US).

Ng, A, and M Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems*, vol. 14, 841.

NIH. 2013. What is a heart attack? `http://www.nhlbi.nih.gov/health/health-topics/topics/heartattack/`.

Page, D. 2015. Predicting health events from electronic health records using machine learning. Talk by David Page at the Center for Predictive Computational Phenotyping. University of Wisconsin.

Pedregosa, F, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Petticrew, M, A Sowden, and D Lister-Sharp. 2001. False-negative results in screening programs: Medical, psychological, and other implications. *International Journal of Technology Assessment in Health Care* 17(2):164–170.

Prentice, R. 1976. Use of the logistic model in retrospective studies. *Biometrics* 599–606.

Qian, M, and S Murphy. 2011. Performance guarantees for individualized treatment rules. *Annals of Statistics* 39(2):1180.

Radcliffe, N. 2007. Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Journal* 1:14–21.

Radcliffe, N, and R Simpson. 2008. Identifying who can be saved and who will be driven away by retention activity. *Journal of Telecommunications Management* 1(2):168–176.

Radcliffe, N, and P Surry. 1999. Differential response analysis: Modeling true response by isolating the effect of a single action. In *Credit Scoring and Credit Control VI*. Edinburgh, Scotland.

———. 2011. Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions.

Rauch, G, B Dogan, T Smith, P Liu, and W Yang. 2012. Outcome analysis of 9-gauge MRI-guided vacuum-assisted core needle breast biopsies. *American Journal of Roentgenology* 198(2):292–299.

Robins, J. 1989. The control of confounding by intermediate variables. *Statistics in Medicine* 8(6):679–701.

Robins, J, M Hernan, and B Brumback. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11(5):550–560.

Rosenbaum, P, and D Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Roshanov, P, N Fernandes, J Wilczynski, B Hemens, J You, S Handler, R Nieuwlaat, N Souza, J Beyene, H Van Spall, A Garg, and R Haynes. 2013. Features of effective computerised clinical decision support systems: Meta-regression of 162 randomised trials. *BMJ* 346.

Rothwell, P. 1995. Can overall results of clinical trials be applied to all patients? *The Lancet* 345(8965):1616–1619.

Russell, R. 2001. Non-steroidal anti-inflammatory drugs and gastrointestinal damage—problems and solutions. *Postgraduate Medical Journal* 77(904):82–88. `http://pmj.bmj.com/content/77/904/82.full.pdf+html`.

Rzepakowski, P, and S Jaroszewicz. 2010. Decision trees for uplift modeling. In *IEEE International Conference on Data Mining*, 441–450. Sydney, Australia.

———. 2012. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems* 32:303–327.

Santos Costa, V, R Rocha, and L Damas. 2012. The YAP Prolog system. *Journal of Theory and Practice of Logic Programming* 12(1 & 2):5–34.

Schnitt, S. 2010. Local outcomes in ductal carcinoma in situ based on patient and tumor characteristics. *Journal of the National Cancer Institute Monogrographs* 2010(41):158–161.

Simard, P, B Victorri, Y LeCun, and J Denker. 1992. Tangent prop—A formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems*, 895–903.

Srinivasan, A. 2007. *The Aleph Manual*, 4th ed. `http://www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph.html`.

Swets, J, D Getty, R Pickett, C D'Orsi, S Seltzer, and B McNeil. 1991. Enhancing and evaluating diagnostic accuracy. *Medical Decision Making* 11:9–18.

Tabar, L, H Tony Chen, M Amy Yen, T Tot, T Tung, L Chen, Y Chiu, S Duffy, and R Smith. 2004. Mammographic tumor features can predict long-term outcomes reliably in women with 1-14-mm invasive breast carcinoma. *Cancer* 101(8):1745–1759.

Taylor, F, M Huffman, A Macedo, T Moore, M Burke, G Davey Smith, K Ward, and S Ebrahim. 2013. Statins for the primary prevention of cardiovascular disease. *Cochrane Database of Systematic Reviews* 1(1).

Thain, D, T Tannenbaum, and M Livny. 2005. Distributed computing in practice: The Condor experience. *Concurrency - Practice and Experience* 17(2-4):323–356.

Therneau, T, and P Grambsch. 2000. *Modeling survival data: Extending the cox model*. Springer Science & Business Media.

Thurfjell, M, A Lindgren, and E Thurfjell. 2002. Nonpalpable breast cancer: Mammographic appearance as predictor of histologic type. *Radiology* 222(1):165–170.

Towell, G, and J Shavlik. 1994. Knowledge-based artificial neural networks. *Artificial Intelligence* 70(1-2):119 – 165.

Tufféry, S. 2011. *Data Mining and Statistics for Decision Making*. 2nd ed. John Wiley & Sons.

Vapnik, V. 1998. *Statistical Learning Theory*. New York: John Wiley & Sons.

Velikova, M, P Lucas, M Samulski, and N Karssemeijer. 2013. On the interplay of machine learning and background knowledge in image interpretation by Bayesian networks. *Artificial Intelligence in Medicine* 57(1): 73 – 86.

Vickers, A, M Kattan, and D Sargent. 2007. Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials* 8(1):14.

Weiss, J. 2014. Statistical Timeline Analysis for Electronic Health Records. Ph.D. thesis, University of Wisconsin-Madison department of Computer Sciences.

Weiss, J, F Kuusisto, K Boyd, J Liu, and D Page. 2015. Machine learning for treatment assignment: Improving individualized risk attribution. In *AMIA Annual Symposium*. San Francisco.

World Health Organization. 2015. *World Health Statistics 2015*. World Health Organization.

Young, J. 2001. Differential Validity, Differential Prediction, and College Admissions Testing: A comprehensive Review and Analysis. Research Report 2001-6, The College Board, New York.

Zadrozny, B, and C Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 609–616. Morgan Kaufmann.

Zaniewicz, L, and S Jaroszewicz. 2013. Support vector machines for uplift modeling. In *IEEE ICDM Workshop on Causal Discovery (CD 2013)*.

Zelezný, F, and N Lavrac. 2006. Propositionalization-based relational subgroup discovery with RSD. *Machine Learning* 62(1-2):33–66.

Zhang, H. 2004. The optimality of naive Bayes. In *Proceedings of the FLAIRS Conference*, 3–9.

Zhang, S, M Hossain, M Hassan, J Bailey, and K Ramamohanarao. 2009. Feature weighted SVMs using receiver operating characteristics. In *Proceedings of the SIAM International Conference on Data Mining*, 497–508. SIAM.

Zhao, Y, D Zeng, A Rush, and M Kosorok. 2012. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* 107(499):1106–1118.