Making Generative Classifiers Robust to Selection Bias

Andrew Smith
University of California, San Diego
9500 Gilman Drive, Mail Code 0404
La Jolla, CA 92093-0404
atsmith@cs.ucsd.edu

Charles Elkan
University of California, San Diego
9500 Gilman Drive, Mail Code 0404
La Jolla, CA 92093-0404
elkan@cs.ucsd.edu

ABSTRACT

This paper presents approaches to semi-supervised learning when the labeled training data and test data are differently distributed. Specifically, the samples selected for labeling are a biased subset of some general distribution and the test set consists of samples drawn from either that general distribution or the distribution of the unlabeled samples. An example of the former appears in loan application approval, where samples with repay/default labels exist only for approved applicants and the goal is to model the repay/default behavior of all applicants. An example of the latter appears in spam filtering, in which the labeled samples can be outdated due to the cost of labeling email by hand, but an unlabeled set of up-to-date emails exists and the goal is to build a filter to sort new incoming email.

Most approaches to overcoming such bias in the literature rely on the assumption that samples are selected for labeling depending only on the features, not the labels, a case in which provably correct methods exist. The missing labels are said to be "missing at random" (MAR). In real applications, however, the selection bias can be more severe. When the MAR conditional independence assumption is not satisfied and missing labels are said to be "missing not at random" (MNAR), and no learning method is provably always correct.

We present a generative classifier, the shifted mixture model (SMM), with separate representations of the distributions of the labeled samples and the unlabeled samples. The SMM makes no conditional independence assumptions and can model distributions of semi-labeled data sets with arbitrary bias in the labeling. We present a learning method based on the expectation maximization (EM) algorithm that, while not always able to overcome arbitrary labeling bias, learns SMMs with higher test-set accuracy in real-world data sets (with MNAR bias) than existing learning methods that are proven to overcome MAR bias.

KDD'07, August 12–15, 2007, San Jose, California, USA.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: statistical computing; H.2.8 [**Database Management**]: Database Applications—*data mining*; I.2.6 [**Artificial Intelligence**]: Learning—*parameter learning*

General Terms

Algorithms, Economics, Theory

Keywords

semi-supervised learning, sample selection bias, reject inference, generative classifiers

1. INTRODUCTION

It is often the case that labeled data available for training in a particular data mining task is not representative of the population in which the resulting model is to be used. Such a data set is said to suffer from "sample selection bias," since data from the general population were labeled in a biased way. In addition to the biased labeled data set, it is often easy to obtain an unlabeled data set randomly sampled from the general population. We present methods for incorporating the unlabeled data into the learning process to improve the accuracy of generative classifiers that ignore selection bias.

Situations necessitating learning in the face of sample selection bias can arise in many contexts. For example, lending institutions create models of who is likely to repay a loan from training sets consisting of people in their records who were given loans in the past; however, the institution only approved loan applications of those it judged likely to repay a loan. Learning from only approved applicants yields an incorrect model because the training set is a biased sample of the general population of applicants, which is the population in which the model is to be used. The lending institution also has the dataset consisting of all applicants, rejected and approved, which is assumed to be a random sample of people who apply for loans. Algorithms that overcome the bias of the labeled data set by incorporating the rejected applicants in the learning process are called "reject-inference" algorithms [4] [6] [7] [8].

Another common example of a learning problem that must deal with sample selection bias occurs in medical and epidemiological domains [1] [24] [18] [17]. We would like to develop a mathematical model for a particular treatment (or exposure–as in epidemiology) that predicts the extent

to which it will affect a particular patient, measured perhaps by the expected lifetime increase, or a probability of survival. To create such a model, we would use a database describing many patients and their responses to the treatment. However, unless the study is a carefully controlled randomized study, any such database will be biased by sample selection, since it only contains patients whom doctors recommended for the treatment; certainly whether or not a doctor recommends someone for a particular treatment is related to how much that treatment is expected to benefit the patient (Epidemiological studies are also subject to related but different bias, such as the different income levels of exposed and unexposed people, which is correlated with access to high-quality care [24].) Learning a response model for the general population should take into account differences between the distribution of patients exposed to treatment and the general distribution of all potential patients.

This bias in the labeling can also arise when the act of labeling data is expensive, and therefore datasets are infrequently updated. For example, to build a spam-detecting classifier, a human must hand-label each email. The distribution of emails may change over time and the spam filter will lose optimality, as it is optimal for old data. This is the "concept-drift" scenario, and the decrease in performance due to the out-dated training set can be mitigated by utilizing unlabeled up-to-date data, which will be drawn from the population in which the spam filter is to be used (since they are up-to-date), and easy to obtain (since they are unlabeled). Unlike the previous two examples in which the goal was to build a model of the general population (i.e. the labeled and unlabeled samples), the goal in overcoming concept-drift is to build an up-to-date model that best classifies new samples, not samples from the general population, which consists of old and new samples.

The next section describes different types of sample selection bias, which are distinguished by their conditional independence assumptions, and intuitive reasons for why these particular types of bias might be present in different real world situations. Section 3 clarifies the distinction between modeling the general population and modeling the unlabeled population, and provides methods for overcoming MAR bias in both scenarios. Section 4 presents a novel method for that can overcome bias that does not assume missing labels are MAR. Section 5 discusses prior work related to the methods presented here. Sections 7 and 8 demonstrate that our methods show improvement over existing methods on real-world data sets.

2. TYPES OF SAMPLE SELECTION BIAS

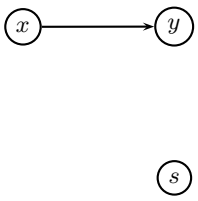
Learning problems that require overcoming sample selection bias are usually presented in terms of a semi-labeled data set in which the labeled samples are not a random subset. The bias present in a semi-labeled data set is characterized by the conditional independence relationships between the features, labels, and presence/absence of a label. We use Bayesian networks as an intuitive tool to organize the different conditional independence relationships [16] [14]. Following are the definition of each variable in our networks, with an example in parentheses of what the variable would represent in the example of loan applications [21].

- y is the class label. If y is binary, we use the notation $y = 1$, (or "good"– loan repaid; the applicant was a good borrower), otherwise $y = 0$.
- s is the selection variable indicating whether that sample was selected for labeling (whether a loan application was approved). $s = 1$ when the data point is selected for observation and y is observable, otherwise $s = 0$ and we cannot observe y.
- x is the vector of observed variables, also called covariates or features, available for training the new model (credit history, income level, age, etc.)

In [21], an additional variable h represented hidden features not available in either the labeled training data or the unlabeled training data. h is omitted here because these networks only include variables and interactions that are directly modeled in our methods.

2.1 No sample selection bias

In the case that samples are selected from the general population for labeling at random, no sample selection bias is present, and the labels of the unlabeled data are said to be "missing completely at random" (MCAR) [13]. This situation is represented by the following Bayesian network.



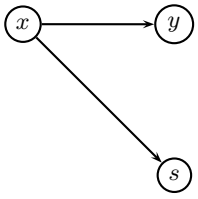
Labeling is only influenced by observable features.

Note that this is the traditional "semi-supervised learning" scenario, where the goal is to use the unlabeled data to reduce the variance in the estimated model, as opposed to overcoming differences in the distributions.

2.2 Learnable sample selection bias

In this type of bias, the selection variable s is influenced only by the observable features. Therefore, the labeled data may constitute a biased sample of the general population, but the mechanism for labeling samples is learnable. In the literature, this type of bias case is called "missing at random" (MAR) [10] [13].

When labels are MAR, whether a sample is labeled, s, is independent from the actual label y, given the features x. This is represented by the most general Bayesian network:



Labeling is only influenced by observable features.

Since we allow only the observable features to influence the selection. Thus the class and selection are conditionally independent:

$$s \perp y|x,$$

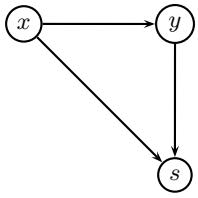
which implies the constraints

$$\begin{aligned} p(s|x,y) &= p(s|x) \\ p(y|x,s) &= p(y|x). \end{aligned}$$

In this graph, observing x d-separates y and s, so the conditional independence relationships are preserved. Selection may depend on x, but given x, y adds no additional information about selection, or, equivalently, given x, knowing s gives no information about the outcome (for example, whether or not someone is a bad borrower).

This could arise in practice if samples are selected for labeling with a formal selection model (such as a logistic regression), as a lending institution might use to approve loan applications. Alternatively, this bias is present when no actual model has been used to select samples, but the concept discriminating between labeled and unlabeled data is still learnable.

2.3 Arbitrary bias



Arbitrary bias in the labeling.

This type of bias in the labeling is completely general: there are no assumptions about conditional independence relationships between the features, labels, and labeling. In this case, the missing labels are said to be "missing not at random" (MNAR) [13].

There are two distinct goals to learning an unbiased model given a semi-labeled data set with biased labeling. The first, **general population modeling**, is represented in the loan application example in which the goal is to model the general population of accepted and rejected applicants given a data set containing only accepted applicants, i.e. to learn a classifier of the form $p(y|x)$, given only labeled examples $(x, y, s = 1)$ and unlabeled examples $(x, s = 0)$.

The other, **unlabeled population modeling**, appears in the concept-drift scenario, for example, in spam filtering. We are given labeled examples of out-dated email, and unlabeled examples of up-to-date email. Unlike general population modeling, we assume the object is to learn a classifier that accurately models incoming email, which is drawn from the same distribution as the up-to-date, unlabeled samples. The goal is to learn an up-to-date model, $p(y|x, s = 0)$ given only labeled examples $(x, y, s = 1)$ and unlabeled examples $(x, s = 0)$.

3. OVERCOMING MAR BIAS - TWO LEMMAS

In both modeling tasks, general population modeling and unlabeled population modeling, the labeled samples can be weighted to approximate a sample drawn from the appropriate distribution, assuming MAR bias.

3.1 General population modeling

The following lemma shows how the joint distribution of the labeled samples is related to the general population:

LEMMA 1. *Under MAR bias in the labeling,*

$$p(x, y) = \frac{p(s = 1)}{p(s = 1|x)} p(x, y|s = 1) \tag{1}$$

if all probabilities are non-zero.

PROOF. Use the assumed conditional independence relationship of MAR bias, then apply Bayes' Rule:

$$\begin{aligned} p(x,y) &= \frac{p(s=1)}{p(s=1|x)} \frac{p(s=1|x)}{p(s=1)} p(x,y) \\ &= \frac{p(s=1)}{p(s=1|x)} \frac{p(s=1|x,y)p(x,y)}{p(s=1)} \\ &= \frac{p(s=1)}{p(s=1|x)} p(x,y|s=1). \end{aligned}$$

□

In medical observational studies, the weight $\frac{p(s=1)}{p(s=1|x)}$, is known as the "inverse probability of treatment weight" (IPTW), where "treatment" means the sample is selected for labeling [24] [17]. This lemma is explored in the context of data mining in [26].

3.2 Unlabeled population modeling

The following lemma shows how the joint distribution of the labeled samples is related to the distribution of unlabeled samples:

LEMMA 2. *Under MAR bias in the labeling,*

$$p(x, y|s = 0) = \frac{p(s = 1)}{1 - p(s = 1)} \frac{1 - p(s = 1|x)}{p(s = 1|x)} p(x, y|s = 1) \tag{2}$$

if all probabilities are non-zero.

PROOF. Apply Bayes' rule, use the assumed conditional independence relationship, then substitute equation 1.

$$\begin{aligned} p(x,y|s=0) &= \frac{p(s=0|x,y)}{p(s=0)} p(x,y) \\ &= \frac{p(s=0|x)}{p(s=0)} p(x,y) \\ &= \frac{p(s=1)}{1-p(s=1)} \frac{1-p(s=1|x)}{p(s=1|x)} p(x,y|s=1). \end{aligned}$$

□

This is related to importance sampling [11], which calculates expectations over a distribution different from the distribution of the actual samples using importance weights. If

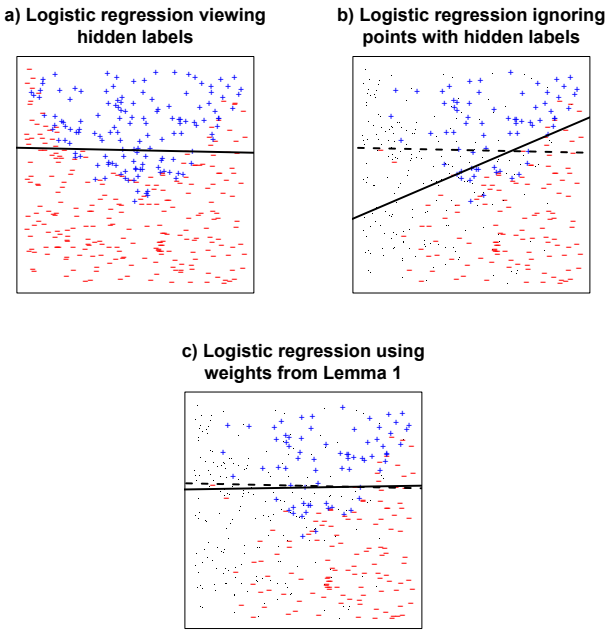


Figure 1: General population modeling with Lemma 1. a) A synthetic labeled data set is generated with 5000 points and a curved decision boundary (for clarity, only some points are shown). A logistic regression classifier is learned viewing the hidden labels to demonstrate the theoretical best performance (solid line, reproduced in plots b) and c) as a dashed line for comparison). b) Some labels are hidden (plotted as black dots), on the left with high probability and on the right with low probability. The classifier estimated from only the labeled points (solid line) is not an accurate boundary in the general population. c) Lemma 1 provides weights for the logistic regression loss function and estimates a boundary nearly identical to the one estimated from all points (viewing hidden labels), therefore minimizing loss in the general population.

samples X are distributed according to $p(x)$, the expectation of a function $f(x)$ over a different distribution $p'(x)$ is given by the weighted sum of $f(x)$ over the samples:

$$E_{p'(x)}f(x) = \sum_{x \in X} f(x)\frac{p'(x)}{p(x)},$$

where the density ratio is the importance weight. Under MAR sample selection bias, the density ratio $\frac{p(x|s=0)}{p(x|s=1)}$ which would be used for importance sampling reduces to the example weight given by Lemma 2. This lemma is derived directly from the importance weight in [2]. Similarly, Lemma 1 is derived from importance sampling weights in [3].

3.3 Using the lemmas with discriminative classifiers

Selecting samples based only on the features preserves the decision boundary because $y \perp s|x$. This would seem to imply that discriminative classifiers are robust to MAR bias since they model decision boundaries; however, an analysis of the loss function of discriminative classifiers shows that the lemmas are useful for improving expected performance in the general and unlabeled populations. This is useful under a common data mining assumption, that the the model is misspecified, i.e. the decision boundary is not known to conform to the particular parametric model in use.

The goal of training a discriminative classifier $f(x)$, modeling $p(y = 1|x)$ is to minimize the expectation of the loss function $L(f(x), y)$ over the distribution of some test set:

$$\mathrm{E}_{x,y}L(f(x), y) = \int_{(x,y)} L(f(x), y)p(x, y)dxdy$$

where the expectation is over the joint distribution of x and y, in this case, the general population, and the sum is over the sampled data, indexed by i. Note that when y is binary, this formulation of loss encompasses maximum-likelihood, $L(f(x), y) = y \log f(x) + (1-y)\log(1-f(x))$, and minimum squared error, $L(f(x), y) = (f(x) - y)^2$.

Since labels are missing for some samples, this loss cannot be directly estimated; however, Lemma 1 shows that it is equal to the loss over the weighted distribution of labeled samples:

$$\begin{aligned}\mathrm{E}_{x,y}L(f(x), y) &= \int_{(x,y)} L(f(x), y)p(x, y)dxdy \\ &= \int_{(x,y)} L(f(x), y)p(x, y|s = 1)\frac{p(s = 1)}{p(s = 1|x)}dxdy \\ &= \mathrm{E}_{x,y|s=1}L(f(x), y)\frac{p(s = 1)}{p(s = 1|x)}\end{aligned}$$

where $\mathrm{E}_{x,y|s=1}$ indicates the joint distribution of x and y among the labeled samples. Similarly, Lemma 2 provides weights to estimate the loss over the distribution of unlabeled samples. The expected loss over the target distribution can be minimized by minimizing the weighted loss over the distribution of labeled samples. [25] explores the issues of maximum likelihood estimation under bias and the use of a weighted loss function for overcoming it.

This weighted loss method first learns a model of the labeling mechanism $p(s|x)$ from the labeled and unlabeled samples. The model is then calibrated[1] and used with the appropriate lemma to create example weights for each labeled sample. An unbiased discriminative classifier is then learned by minimizing the loss of these weighted (labeled) samples, ignoring the unlabeled samples.

Figure 1 demonstrate how a semi-labeled data set with MAR bias in the labeling can result in a discriminative classifier that is not optimal for the target distributions (in the figure, the general population is used as an example), and how the lemmas can provide weights for the loss function to overcome the bias.

Interactions between model misspecification and MAR bias are explored in [20]. Specifically it is shown that when the model is correctly specified, maximum-likelihood estimation

[1]Here, a model is calibrated if the estimate $\hat{p} = p(s = 1|x)$ is equal to the proportion of samples whose value of s is 1, among those samples for which the model output is (near) $\hat{p}$. [27]

is asymptotically unbiased. When the model is misspecified, the asymptotically unbiased maximum-likelihood estimate is shown to maximize the log-likelihood equation weighted with the importance sampling weight (density ratios), which reduce to lemmas 1 and 2.

3.4 Using the lemmas with generative classifiers

Generative classifiers model the joint density $p(x, y)$, factored into a class prior probability $p(y)$ and a class-conditional density model $p(x|y)$. Unlike discriminative classifiers, generative classifiers are not inherently robust to MAR bias in the labeling. This is because the density of features within each class among the labeled samples, $p(x|y, s = 1)$, is not guaranteed to be the same in the general population, $p(x|y)$. The specific density model used in our experiments are Gaussian mixture models (GMMs) and are explained in Section 6; however, our methods are general and can be used with whichever density model is appropriate for the learning task at hand.

Both lemmas have straight forward applications to overcoming MAR bias with generative classifiers:

- Learn a classifier for $p(s|x)$, distinguishing labeled samples from unlabeled samples. Use it to assign to each labeled sample x_i, a weight w_i, according to the reweighting lemma appropriate for the test set distribution.
- The prior probability in the test set for class c is proportional to the sum of weights of the labeled samples in that class in the training set:

  $$p(y = c) \propto \sum_{y_i = c} w_i.$$

  Similarly, an unbiased model of the class conditional density for class c in the test set, $p(x|y = c)$ can be estimated from the weighted labeled samples in that class, for example, using Gaussian mixture models (Section 6).

The class priors and class conditional densities are combined to form a classifier using Bayes' Rule:

$$p(y = c_i|x) = \frac{p(x|y = c_i)p(y = c_i)}{\sum_{j=1}^{|C|} p(x|y = c_j)p(y = c_j)}$$

This elegant method is provably correct when the missing labels are MAR. Practically, though, it is only useful to the extent that the classifier used to create the weights can provide accurate probability estimates. Similarly, if one of the probabilities in the denominator of equations 1 or 2 is zero, the weight for that sample is infinite. This can be avoided in practice with a maximum sample weight.

4. OVERCOMING MNAR BIAS – THE SHIFTED MIXTURE MODEL (SMM)

The lack of guarantees regarding the relationship between the labeled and unlabeled samples under MNAR bias conflicts with the intuition that the underlying concepts should be similar. For example, the map from words in emails to the spam/ham label is expected to drift over time, but not completely change. The approach presented here is motivated by this intuition.

We can use Lemma 2 to create a generative classifier for the unlabeled data, $p(x|y, s = 0)$ and $p(y|s = 0)$, that is robust to MAR bias; however, this classifier will have the same decision boundary as one for the labeled data and for the general population, since $p(y|x, s = 0) = p(y|x, s = 1) = p(y|x)$ under the conditions of MAR bias. The shifted mixture model (SMM) uses Lemma 2 to estimate the generative classifier $p(x|y, s = 0)$ and $p(y|s = 0)$, then allows the parameters of this model to shift to increase the likelihood of the unlabeled data.

The complete SMM consists of two generative classifiers: a model of the labeled data, $p(x|y, s = 1)$ and $p(y|s = 1)$, which is estimated directly from the labeled samples and a model of the unlabeled data, $p(x|y, s = 0)$ and $p(y|s = 0)$, which is initialized with Lemma 2 and shifted with the EM algorithm. In addition, the probability of selection $p(s)$ can be estimated directly from the labeled and unlabeled data.

The log-likelihood of model parameters Θ given the unlabeled data $(x_i, s_i = 0)$, for $i = 1..N$ is

$$\begin{aligned} l(\Theta; X, s = 0) &= \sum_{i=1}^{N} \log p_\Theta(x_i, s_i = 0) \\ &= \sum_{i=1}^{N} \log \sum_y [p_\Theta(x_i|y, s_i = 0) \\ &\quad p_\Theta(y|s = 0)p_\Theta(s_i = 0)]. \end{aligned}$$

Since no label is given for these samples, their probability is the sum over the possible labels y. The logarithm of the sum makes direct maximization difficult, so the likelihood it is iteratively increased with the EM algorithm. [9] showed that this is bounded from below by the expectation of the so-called "complete data" log-likelihood:

$$\begin{aligned} l(\Theta; X|s = 0) &= \sum_{i=1}^{N} \sum_{y \in Y} z_{iy} \log[p_\Theta(x_i|y, s_i = 0) \\ &\quad p_\Theta(y|s_i = 0)p_\Theta(s_i = 0)] \end{aligned} \quad (3)$$

where z_{iy} is the probability that unlabeled sample i was generated by class y. Re-estimating parameters to increase this lower bound also increases the log-likelihood. Each iteration j of the EM algorithm uses the current estimate of the parameters Θ^j to estimate Z_{iy}, and then uses that matrix to update the parameters.

If run to convergence, the EM algorithm finds a good clustering of the data; however, in real-world data mining tasks, it is not necessarily the case that the natural clusters in the data correspond well to the class labels. For this reason, it is important not to allow the parameters to shift too far from their initialization, or the mixture components will represent the natural clustering of the data as opposed to the density of the classes of unlabeled data. This is accomplished by only running the EM algorithm for a few iterations (we use 5), as well as updating the parameters using an "inertia" factor α to slow the parameter evolution. In iteration j of the EM algorithm, when the M-step estimates new parameters Θ', the actual parameters used in the next iteration are a combination of the current parameters and the new parameters:

$$\Theta^{j+1} \leftarrow \Theta^j \alpha + \Theta'(\alpha - 1).$$

After improving the likelihood of the model $p(x|y, s = 0)$

and $p(y|s=0)$ given the unlabeled data, the model better represents the classes within the population of unlabeled samples. Our experiments used $\alpha = 0.99$.

For the unlabeled population modeling task, the generative model of the unlabeled data can be used directly. For the general population modeling task, the two generative classifiers are combined to model the joint

$$\begin{aligned} p(x_i, y_i) &= p(x_i|y_i, s_i=0)p(y_i|s_i=0)p(s_i=0) \\ &\quad + p(x_i|y_i, s_i=1)p(y_i|s_i=1)p(s_i=1), \end{aligned}$$

and test samples can be classified using the definition of conditional probability

$$p(y_i|x_i) = \frac{p(x_i, y_i)}{\sum_y p(x_i, y)}.$$

Since shifting the parameters potentially changes the decision boundary, the resulting generative classifier can generate samples with MNAR bias. In Sections 7 and 8, the SMM is demonstrated to partially overcome MNAR bias in practice.

5. RELATED WORK

Sample selection bias is also known as "covariate shift," especially in regression literature [22] [23]. An approach to overcoming covariate shift in regression tasks is given in [22]. Their model assumes (x, y) pairs in the training set are generated by two independent processes, $P_1(x, y)$ and $P_2(x, y)$, but that the test data is only generated by the first process, $P_1(x, y)$. The goal is to estimate mixture model parameters that separate effects that are expected to be characteristic only of the test distribution from those characteristic of both test and training data. As with the SMM, the parameters are estimated by maximizing the likelihood of a mixture model with the EM algorithm; however, this framework assumes training instances are drawn from the general population, and test data are selected not at random from that population.

A different approach to unlabeled population modeling, in this case under MAR bias, is given in [2]. Here, Lemma 2 is derived directly from the importance weights as opposed to our derivation from the conditional independence assumptions of MAR bias. A discriminative classifier for the unlabeled population and selection probabilities are learned with a single convex optimization. The convexity requirement, however, precludes the use of more flexible semi- and nonparametric models.

The idea of learning model parameters from a training set and then adjusting them to better account for the samples from a different target distribution is used in [5]. A small set drawn from the target distribution is used to adjust the parameters learned from a large training set with a different distribution; however, in their approach both sets are labeled. In a maximum entropy framework, they use the large training set to learn a prior distribution over model parameters, which are updated using data drawn from the target distribution; this is analogous to our initialization of the shifted mixture components based on the labeled data. To limit parameter evolution, each model parameter is alowed to be shifted away from its prior up to a maximum distance that is based on that parameter's variance estimated from the large training set. This is used in natural language processing for so-called "domain adaptation," learning a model when the target and training domains are different.

A general approach to overcoming MNAR bias is given in [19]. Similar to the method suggested by Lemma 1, their method creates examples weights based on a selection model learned from the data; however, their selection model $p(s = 1|x, y)$ is dependent on the class label. If these selection probabilities were known exactly, the weights would correct for MNAR bias; however, it is not proven that this method can always estimate reliable selection models. Despite this lack of a guarantee, the method is demonstrated to be useful for overcoming MNAR bias in a real-world data mining task.

To learn a selection model $p(s = 1|x, y, \theta)$, where θ is the set of model parameters, they observe that for any function $g(x)$ of the features, the expectation of $g(x)$ in the general population matches the weighted expectation over the distribution of labeled samples:

$$E_{P(x)}g(x) = E_{P(x|s=1)}\frac{g(x)}{p(s=1|x, y, \theta)}.$$

With respect to a data set in which samples 1 through M are labeled and $M+1$ through $M+N$ are unlabeled, this is expressed empirically as

$$\sum_{i=1}^{M+N} g(x) = \sum_{i=1}^{M} \frac{g(x)}{p(s=1|x, y, \theta)}. \quad (4)$$

The method estimates $\theta \in \Re^k$ by selecting k functions, $g_1(x), ..., g_k(x)$, and demanding equation 4 be satisfied for all $g_i(x)$. This method has the potential to overcome MNAR bias, though it requires solving nonlinear simultaneous equations, requires a parametric selection model, and requires careful selection of the functions $g_i(x)$ to insure the uniqueness, stability, and robustness of the estimate.

A similar method based on kernel means matching is given in [12], however it requires missing labels to be MAR. This method does not require any model of the selection mechanism, but estimates optimal example weights β_i for each labeled sample directly.

Let $\Phi : X \rightarrow F$ be a map from the feature space X to the kernel feature space F, and let $\mu : \wp \rightarrow F$ be the operator

$$\mu(P) = E_{x\sim P(x)}\Phi(x)$$

which maps probability distribution $P \in \wp$ to its expectation under transformation Φ, called the kernel mean. It is shown in [12] that weights $\beta(x)$ for each example x are optimal when the kernel mean of the distribution of the general population Pr' is equal to the weighted mean of the distribution Pr of the labeled samples:

$$\mu(Pr') = E_{x\sim Pr(x)}\beta(x)\Phi(x).$$

Convergence results and proof of convexity are also given.

Both methods correct differences between the distribution of the labeled data and the general distribution using example weights. In [19] these weights are determined by an explicit selection model, and in [12] are estimated directly, but both are found by requiring that some expectation in the general population be equal to the weighted expectation over the labeled data.

6. GAUSSIAN MIXTURE MODELS

The methods for overcoming selection bias presented here are built on generative classifiers, which require accurate

multivariate density estimation. Gaussian mixture models (GMMs) comprise a class of semiparametric density estimators, and are flexible enough to model many real-world distributions. GMMs estimate the density at x, $p(x)$ as a weighted sum over the $|C|$ component densities,

$$p(x) = \sum_{c \in C} p(c)p(x|c).$$

Each component in C has a multivariate normal distribution:

$$p(x|c) = \frac{1}{(2\pi)^{(d/2)}|\Sigma_c|^{(1/2)}} \exp^{-\frac{1}{2}(x-\mu_c)^T \Sigma_c^{-1}(x-\mu_c)}$$

where d is the dimensionality, and the parameters of component c are the mean μ_c and the covariance matrix Σ_c.

The probabilistic formulation of this density model allows the application of maximum likelihood (ML) methods, such as the expectation maximization (EM) algorithm [9]. Similar to the application of EM to the shifted mixture model, estimating a GMM with EM iteratively improves the log-likelihood of the N data points,

$$\sum_{i=1}^{N} \log \sum_{c \in C} p(c)p(x_i|c),$$

by bounding it from below and improving the bound, which is the expectation of the so-called "complete data" log-likelihood :

$$\sum_{i=1}^{N} \sum_{c \in C} z_{ic} \log(p(c)p(x_i|c)) \quad (5)$$

where z_{ic} is the probability that point x_i is generated by component c, which is analogous to the class variable y used in estimating the SMM.

To fit a GMM, the parameters for each component, $p(c)$, μ_c, and Σ_c are initialized and the following two steps are iterated: **E-step:** Use the current parameters to estimate the matrix $Z = \{z_{ic}\} = p(c|x_i)$. **M-step:** Use Z to maximize equation 5 over the parameters.

Typically, this is iterated for some maximum number of steps or until a convergence criterion is satisfied.

6.1 Computational issues

There are several known difficulties applying EM to Gaussian mixture models.

Initialization. We start with a (uniform) random distribution over component responsibilities. This is equivalent to starting at the M-step with a random Z matrix (with rows that sum to 1). We found that in general, classifiers could be estimated to have reliably high test-set accuracy by initializing them with a general mixture model of the features. For example, in a binary classification task mapping features x to label $y \in \{1, 0\}$, a GMM modeling $p(x)$ would be estimated using EM and its parameters would be used to initialize the class-conditional densities $p(x|y = 0)$ and $p(x|y = 1)$, which would then both be improved with EM.

Local minima. Our solution to avoid local minima in the solution is to run the EM algorithm several times under different random initializations, and pick the model with the highest likelihood. Each model was estimated by running EM until the relative change in the log-likelihood of the data was less than 10^{-7}. We found picking the best of 25 models to provide sufficiently stable results.

Non-invertible covariance matrix. There are two common ways for EM to produce a singular Σ. Data sets in which some samples have the exactly the same value for a particular feature are particularly prone to this problem because it is possible for one component to take responsibility only for a subset of these samples and contain no variance for that feature. To overcome this issue, uniform random noise in $[-0.25, 0.25]$ is added to the binary and integer-valued features. Initial tests with logistic regression showed this to have minimal impact on class-separability.

Another reason a covariance matrix might be singular is if that component claims responsibility for too few points; a d dimensional covariance matrix estimated from fewer than d samples is singular. Our solution was to stop updating components in the M-step which were found to have responsibility for fewer than $2d$ points. The probabilistic responsibility of component j for sample i is given by Z_{ij}, and for the purpose of avoiding singular covariance matrices, each point is assigned to the component with highest partial responsibility. Our experiments used only 6 components and this rarely occurred. When it does, the resulting mixture model usually has lower likelihood than the other 25 and is rejected.

7. EXPERIMENTS – ADULT DATA SET

The ADULT data set contains demographic information collected for a census. The target value is predicting whether income exceeds $50,000 annually, using features measuring education, household statistics, reported capital gains and losses, employment and marital status. [1]

This data set contains features similar to what might be in a loan application approval system. The target variable is analogous to a repay/default label. Marital status is used to select which samples are labeled (married) and which have hidden labels (unmarried), analogous to which loan applications were approved (selected for labeling). In this sense, marital status is interpreted as a unquantifiable measure of responsibility which would not be officially recorded in a real loan application (and, in fact, may not be legal), but which nevertheless might influence the ultimate decision to accept or reject the application. It is not immediately clear how the distribution of income of married people is related to the general distribution of income, even given the other demographic information.

Unlike data sets collected from actual loan approval records, this data set contains labels for every sample, so test set accuracy can be directly measured.

Testing the type of bias. Which type of bias is induced by this method of sample selection? It is possible that marital status is not related at all to income and therefore that the labels are MCAR; however, this is easily shown to be

[1]The categories in each categorical feature were combined to form a binary feature. For example, the categorical feature "country of origin" is transformed into the binary feature "is native-born US citizen." "Capital gains" and "capital losses" were transformed by $\log(1 + x)$. After all transformations, noise is added as described in the previous section, to prevent Gaussian components in the mixture models from losing all variance in a particular feature. The labels of the unmarried samples are hidden, the "marital status" feature is removed, and 40% of the samples are held out as a test set for validation. This yields 12798 labeled samples, 14342 unlabeled samples, and 18092 test samples (married and unmarried).

Table 1: ADULT data set: General population modeling

Model	Test set accuracy
Log. reg. (uncorrected)	74.0% (.19%)
Log. reg. + Lemma 1	74.2% (.22%)
Log. reg. (viewing hidden labels)	80.7% (.4%)
GMM + Lemma 1	79.9% (.38%)
SMM	81.4% (.38%)

Table 2: ADULT dataset: unlabeled population modeling

Model	Test set accuracy
Log. reg. (uncorrected)	77.3% (.27%)
Log. reg. + Lemma 2	76.8% (.27%)
Log. reg. (viewing hidden labels)	93.9% (.18%)
GMM + Lemma 2	85.8% (.69%)
SMM	90.2% (.58%)

false by comparing the two probabilities:

$$p(y=1|s=1) = 0.4556$$
$$p(y=1|s=0) = 0.0692.$$

Another possibility is that marital status is random given the observable features, in which case the selection for labeling is still only dependent on observable features (MAR). This is testable using a discriminative classifier. The accuracy of a logistic regression trained on the biased (married) data but tested on unbiased data shows the MAR scenario to be not plausible, even when using Lemma 1 to weight the likelihood function: the accuracy of a model learned from the biased training set is 74.2%, but the accuracy a similar classifier learned from an unbiased training set (viewing the "hidden" labels) which is 80.7% on the same test set. The increase in the training set size is not to be cause of the increase in performance, rather, it is because the observing the hidden labels removes all bias.

Since the decision boundaries are different, $p(y|x, s=1) \neq p(y|x)$, the missing labels are MNAR.

Main results. Table 1 shows the test set accuracy of the SMM is significantly better than the accuracy of a logistic regression that ignores the bias in both general population modeling and unlabeled population modeling. Generative classifiers estimated using the reweighting lemmas also show improved performance over logistic regression; however the SMM is better able to model the unlabeled data, improving classification accuracy of the test sets. For these tests mixture models with 6 components are used as density estimators. The test-set accuracies are averages over 10 random test/train splits (using 60% for training, 40% for testing), with the standard deviation in parentheses.

8. EXPERIMENTS – CA-HOUSING DATA SET

The California Housing (CA-HOUSING) is based on US Census data and contains 20640 records about house values in California [15]. Each record describes a localized area in California and contains the following features: median INCOME, median house AGE, total ROOMS, total BEDROOMS, POPULATION, HOUSEHOLDS, LONGITUDE, and LATITUDE, in addition to the median VALUE of houses in each area [19]. The target is to predict whether the house value is above the median house value in all of California or the rest of California to test general population modeling or unlabeled population modeling, respectively.

With the CA-HOUSING data set, labels in the training set are hidden depending on the geographic location where that sample was observed. Samples are labeled only if they are approximately within 0.40 degrees of longitude (approximately 22.4 miles) of the ocean, and above the 36th parallel (which is below the San Francisco Bay area). The majority of these houses are concentrated in the San Francisco peninsula and in Marin County. The longitude and latitude features are then removed from the data set.[2]

This represent a scenario in which a model of house prices is being built for California and census data are available for the entire state, however only house prices in the coastal northern California are available.

Testing the type of bias. Houses on the San Francisco peninsula and in Marin County are on average more expensive than in the rest of the state, indicating bias in the labeling is present:

$$p(y=1|s=1) = 0.751$$
$$p(y=1|s=0) = 0.443.$$

With the biased training set, the test accuracy of a logistic regression classifier weighted with Lemma 1 is 74.8%. A similar classifier learned from labeled and unlabeled samples (viewing the "hidden" labels) yields 80.5% on the same test set. Since the decision boundaries are different, $p(y|x, s=1) \neq p(y|x)$, and therefore the missing labels are MNAR.

Main result. Tables 3 and 4 shows that the SMM has significantly higher test-set classification accuracy in both the general population modeling task and the unlabeled population modeling task. As with the ADULT data set, generative classifiers using GMMs and weighted data shows higher test set accuracy than logistic regression. The SMM is a better model of the unlabeled population than re-weighted GMMs, resulting in highest test set accuracy classifying houses not in coastal northern California (unlabeled population), and an increase of classifying houses throughout the state (general population). For these tests mixture models with 6 components are used as density estimators. The test-set accuracies are averages over 10 random test/train splits (using 60% for training, 40% for testing), with the standard deviation in parentheses

9. DISCUSSION

Both experiments show that the shifted mixture model improves the generative classifier estimated using the weights

[2]Coastal homes are found by creating 50 strata in the latitude with equal numbers of samples. Within each stratum the coast is defined as the sample with lowest longitude (most western), and the samples within 0.4 degrees of longitude of the western most house in each stratum were defined as "coastal". After the samples not from coastal northern California were hidden, LONGITUDE and LATITUDE were removed from the data set. ROOMS, BEDROOMS, POPULATION, and HOUSEHOLDS were log-transformed. Noise is added to AGE as with the ADULT data set, to avoid singular covariance matrices in the Gaussian components of the mixture model. Again, 40% of the data are held out in a test set, yielding 2159 labeled samples, 10228 unlabeled samples, and a test set size of 8253 samples.

Table 3: CA-HOUSING data set: general population modeling

Model	Test set accuracy
Log. reg. (uncorrected)	75.1% (.37%)
Log. reg. + Lemma 1	74.8% (.38%)
Log. reg. (viewing hidden labels)	80.5% (.33%)
GMM + Lemma 1	75.9% (.63%)
SMM	77.4% (.45%)

Table 4: CA-HOUSING data set: unlabeled population modeling

Model	Test set accuracy
Log. reg. (uncorrected)	72.9% (.40%)
Log. reg. + Lemma 2	72.7% (.35%)
Log. reg. (viewing hidden labels)	80.3% (.25%)
GMM + Lemma 2	77.3% (.67%)
SMM	79.8% (.57%)

from the lemmas. This more accurate model of unlabeled data results in better test-set classification accuracies in both the general population modeling task, and the unlabeled population modeling task. Lemmas 1 and 2 allowed estimation of generative classifiers that had better test-set accuracies than logistic regressions trained on the same biased data, but to a lesser extent than the SMM.

These experiments are both plausible real-world modeling scenarios. The experiment with the ADULT data set is analogous to the loan application approval problem, where income level represents repay/default behavior. Marital status is analogous to the decision of a lending institution to approve a loan. It was shown to be partially predictable from the other features in the data set, but also to have additional influence on the target variable, given those features. The SMM can partially capture this additional influence and improve test-set classification accuracy.

The CA-HOUSING experiments showed that the SMM can learn a model of housing prices throughout California based on a biased labeled subset that is more accurate than a logistic regression that ignores the bias in the labeling.

10. CONCLUSIONS AND FUTURE WORK

Most approaches to overcoming selection bias in the literature rely on the MAR assumption, which implies the decision boundary is the same for the training and test sets. In real world applications, however, the conditional independence requirements that define MAR are rarely perfectly satisfied. Despite this shortcoming, we demonstrated that the re-weighting lemmas can be used to estimate generative classifiers that are more robust to selection bias than simple decision boundaries, even when the selection is demonstrably MNAR.

By shifting the parameters of a generative classifier modeling the unlabeled data to increase the likelihood, the SMM can achieve accuracy higher than what is guaranteed by the lemmas, in both the general population modeling task and the unlabeled population modeling task.

Effective use of the SMM requires controlling the evolution of parameters as the likelihood given the unlabeled data is improved. This was acomplished by limiting EM to 5 iterations and using the inertia parameter, $\alpha = .99$. Future work will include more sophisticated methods for controlling parameter shifts, as well as investigating the effects of only allowing some parameters to change (e.g. mixture component means but not covariance matrices).

Future work will investigate the effects of modulating the flexibility of the learning algorithm, as manifested in the number of components, both on creating classifiers for $p(s = 1|x)$ to learn the weights, and for the final classification task, learning $p(y|x)$ or $p(y|x, s = 0)$. In particular, the effect of the quality of the weights on the resulting classifier will be investigated.

11. ACKNOWLEDGEMENTS

This work was supported by Fair Isaac.

12. REFERENCES

[1] K. Benson and A. J. Hartz. A comparison of observational studies and randomized controlled trials. *The New England Journal of Medicine*, 342(25):1878–1886, 2000.

[2] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[3] S. Bickel and T. Scheffer. Dirichlet-enhanced spam filtering based on biased samples. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 161–168. MIT Press, Cambridge, MA, 2007.

[4] W. J. Boyes, D. J. Hoffman, and S. A. Low. An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*, 40(1):3–14, 1989.

[5] C. Chelba and A. Acero. Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, 20(4):382–399, 2006.

[6] G. Chen and T. Astebro. The economic value of reject inference in credit scoring. In J. N. C. L. C. Thomas and D. B. Edelman, editors, *Credit Scoring and Credit Control VII: Proceedings of Conference held at University of Edinburgh, Edinburgh, Scotland, 5-7 September*, 2001.

[7] D. A. Cobb-Clark and T. Crossley. Econometrics for evaluations: An introduction to recent developments. *The Economic Record*, 79(247):491–511, 2003.

[8] J. Crook and J. Banasik. Does reject inference really improve the performance of application scoring models? Technical Report Working Paper Series No. 02/3, Credit Research Centre, 2002.

[9] A. Dempster, N. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[10] A. J. Feelders. An overview of model based reject inference for credit scoring. Technical report, Utrecht University, Institute for Information and Computing Sciences, (unpublished). http://www.cs.uu.nl/people/ad/mbrejinf.pdf.

[11] Glynn, Peter W. and Iglehart, Donald L. Importance sampling for stochastic simulations. *Management Science*, 35(11):1367–1392, nov 1989.

[12] J. Huang, A. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by

unlabeled data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

[13] R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, Inc., New York, NY, USA, second edition, 1986.

[14] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002.

[15] R. Pace and R. Barry. Sparse spatial autoregressions. *Statistics and Probability Letters*, 33:291–297, 1997.

[16] J. Pearl. Graphical models for probabilistic and causal reasoning. In D. M. Gabbay and P. Smets, editors, *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Volume 1: Quantified Representation of Uncertainty and Imprecision*, pages 367–389. Kluwer Academic Publishers, Dordrecht, 1998.

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

[18] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–56, 1983.

[19] S. Rosset, J. Zhu, H. Zou, and T. Hastie. A method for inferring label sampling mechanisms in semi-supervised learning. *Advances in Neural Information Processing Systems*, 17:1161–1168, 2005.

[20] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[21] A. Smith and C. Elkan. A bayesian network framework for reject inference. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 286–295, New York, NY, USA, 2004. ACM Press.

[22] A. Storkey and M. Sugiyama. Mixture regression for covariate shift. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1337–1344. MIT Press, Cambridge, MA, 2007.

[23] M. Sugiyama and K.-R. Müller. Model selection under covariate shift. In W. Duch, J. Kacprzyk, E. Oja, and S. Zadrozny, editors, *Artificial Neural Networks: Formal Models and Their Applications*, volume 3697 of *Lecture Notes in Computer Science*, pages 235–240, Berlin, 2005. Springer.

[24] A. J. Treno, P. J. Gruenewald, and F. W. Johnson. Sample selection bias in the emergency room: an examination of the role of alcohol in injury. *Addiction*, 93(1):113–29, 1998.

[25] K. Yamazaki, M. Kawanabe, S. Watanabe, M. Sugiyama, and K.-R. Müller. Asymptotic Bayesian generalization error when training and test distributions are different. In *Proceedings of 24nd International Conference on Machine Learning*, Corvallis, Oregon, USA, Jun. 20–24 2007.

[26] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 114, New York, NY, USA, 2004. ACM Press.

[27] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, New York, NY, USA, 2002. ACM Press.