Making Generative Classifiers Robust to Selection Bias

Andrew Smith
University of California, San Diego

What is selection bias?
A generalization of semi-supervised learning:
- The training set and testing set may be differently distributed...
- but unlabeled examples from the testing set are available.

Examples
- Loan application approvals: Goal is to model repayment/default behavior of all applicants, but the training set only includes labels for people who were approved for a loan.
- Medical observational studies: Modeling the effect of an experimental treatment on the general population is complicated by training sets containing only the effects of the treatment on patients doctors approved for the experiment.
- Spam filtering: We want an up-to-date spam filter, but hand-labeled data sets are expensive and may be rarely updated.

Types of selection bias
The variables in our model of selection bias are similar to those of standard supervised learning:
- \( x \) is the feature vector.
- \( y \) is the class label. If \( y \) is binary, \( y \in \{0,1\} \).
- \( s \) is the binary selection variable; if \( y \) is observable then \( s = 1 \), otherwise \( s = 0 \).

The different types of selection bias are conditional independence relationships between these variables.

No selection bias
- The standard semi-supervised learning scenario.
- Labeled examples are selected completely at random from the general population.
- The missing labels are said to be “missing completely at random” (MCAR) in the literature.

Learnable selection bias
- Labeled examples are selected from the general population only depending on features \( x \).
- A model \( p(s|x) \) is learnable.
- The missing labels are said to be “missing at random” (MAR), or “ignorable bias” in the literature.

Model mispecification under learnable selection bias
- Although \( p(y|x, s = 1) \neq p(y|x) \) implies decision boundaries are the same in the labeled and general populations, if the model is mispecified a sub-optimal decision boundary may be learned under MAR bias.

Arbitrary selection bias
- Labeled examples are selected from the general population depending on the label itself.
- No assumptions about the bias can be made.
- The missing labels are said to be “missing not at random” (MNAR) in the literature.

Overcoming learnable bias
The training data consist of \( \{(y_i, x_i), s_i = 1\} \) and \( \{(y_i, x_i), s_i = 0\} \).
Two goals are possible:
- General population modeling: Learn \( p(y|x) \), e.g. loan approval.
- Unlabeled population modeling: Learn \( p(y|x), s = 0 \), e.g. spam filtering.

Both goals are attainable in multiple ways, assuming the conditional independence relationship of MAR bias.

General population modeling

**Lemma 1** Over MAR bias in the labeling,
- \( p(x,y) = p(x = 1) p(y|x = 1) + p(x = 0) p(y|x = 0) \)
- \( p(x,y) = p(y|x) p(x) \)

if all probabilities are non-zero.

The distribution of samples in the general population is a weighted version of the distribution of labeled samples. Since \( p(y|x) \) is learnable, we can estimate weights.

This lemma can be used to estimate class conditional density models \( p(y|x) \), or improve the loss of misspecified discriminative classifiers in the general population.

Unlabeled population modeling

**Lemma 2** Over MAR bias in the labeling,
- \( p(x,y) = p(x = 1) (1 - p(s = 1|x)) p(y|x = 1) + p(x = 0) p(y|x = 0) \)
- \( p(x,y) = p(y|x) p(x) \)

if all probabilities are non-zero.

Similarly, the distribution of samples in the unlabeled population is a weighted version of the distribution of labeled samples. Since \( p(y|x) \) is learnable, we can estimate weights.

This lemma can be used to estimate class conditional density models \( p(y|x, s = 0) \), or improve the loss of misspecified discriminative classifiers in the unlabeled population.

Conclusions
- Model mispecification, the reality in real-world applications, means that even under MAR bias, re-weighting can help discriminative classifiers.
- Even when the cluster assumption may not be true, improving the likelihood of the model parameters for the unlabeled data yields better classifiers in the unlabeled and general populations.
- The SMM approach is a practical method of allowing the unlabeled model parameters to better represent the unlabeled samples while remaining close to the parameters defining the decision boundary for the labeled data.

Overcoming Arbitrary Bias – the Shifted Mixture Model

Traditional ML semi-supervised learning

Log-likelihood of semi-labeled data:
- \( f(\theta, X) = \sum_i \log p(y_i|x_i, s_i = 1) + \sum_i \log p(x_i|\theta) + \sum_i \log p(s_i=0|x_i) \)

for labeled data \( i = 1 \ldots m \) and unlabeled data \( i = m + 1 \ldots n \).

The labeled and unlabeled data are constrained to be modeled by the same \( p(y|x) \), which isn’t the case under arbitrary bias.

For each sample, we have an extra bit to include in the model, \( s \):

Extended:
- \( f(\theta, X) = \sum_i \log p(y_i|x_i, s_i = 1) + \sum_i \log p(y_i|x_i, s_i = 0) + \sum_i \log p(s_i=0|x_i) \)

Improving the likelihood for unlabeled samples
- Solution: Assume the model parameters for the labeled data and the unlabeled data are “close.”
- \( \text{Learns density models } p(x_i, s_i = 1) \text{ and priors } p(y_i|x, s_i = 1) \text{ from labeled data.} \)
- Initialize models for unlabeled data with parameters learned from labeled data, possibly using lemma 2.
- Improve the likelihood of the parameters for the unlabeled (unsupervised learning).

The SMM approach is useful for both general and unlabeled population modeling, as it produces an explicit generative model of both populations.

Application of the SMM approach to EM:
- Limit to few (10) iterations to limit parameter changes.
- Use inertia parameter \( \alpha = 0.99 \) to slow parameter evolution.

Experiment 1

**ADULT dataset**

- \( x \): (AGE, EDUCATION, CAPITAL GAIN, CAPITAL LOSS, HOURS PER WEEK, SEX, NATIVE TO US, ETHNICITY, FULLTIME)
- \( y \): INCOME \( \sim \) $50,000
- \( m \): MARRIED?

This dataset has information most of which could be used in a loan approval system. The target is an analog of a repayability/default behavior label. Marital status is analogous to an unquantifiable measure of responsibility which wouldn’t be in a bank’s records, but which might influence the label.

Determining the type of bias

Is it MCAR? No
- \( p(y = 1|x = 1) = 0.4726 \)
- \( p(y = 1|x = 0) = 0.0092 \)

Is it MAR? Not as far as logistic regression can detect:
- Accuracy in gen. pop. based on labeled data \( = 74.2\% \)
- Accuracy in gen. pop. based on all data \( = 80.7\% \)

Experiment 2

**CA-HOUSING dataset**

- \( x \): CA census tract data: MEAN INCOME, MEAN HOME AGE, TOTAL ROOMS, TOTAL BEDROOMS, POPULATION, HOUSEHOLDS
- \( y \): OWN VALUE \( \sim \) California median
- \( k \): LATITUDE > 36 and within 0.4 degrees of coast

This task is to build a model of housing prices throughout California when price information is only available in a limited location.

Determining the type of bias

Is it MCAR? No
- \( p(y = 1|x = 1) = 0.75 \)
- \( p(y = 1|x = 0) = 0.443 \)

Is it MAR? Not as far as logistic regression can detect:
- Accuracy in gen. pop. based on labeled data \( = 74.8\% \)
- Accuracy in gen. pop. based on all data \( = 80.5\% \)