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 approval: Goal is to model repaying/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:

- \( \mathbf{x} \) is the feature vector.
- \( \mathbf{y} \) is the class label. If it 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

\[ p(y | \mathbf{x}, s = 1) = p(y | \mathbf{x}) \]

- 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

\[ p(y | \mathbf{x}, s = 1) \neq p(y | \mathbf{x}) \]

- Labeled examples are selected from the general population only depending on features \( x \).
- A model \( p(y | \mathbf{x}) \) 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 \( p(y | \mathbf{x}, s = 1) = p(y | \mathbf{x}) \) implies decision boundaries are the same in the labeled and general populations, if the model is mis-specified a sub-optimal decision boundary may be learned under MAR bias.

\[ p(y | \mathbf{x}, s = 1) \neq p(y | \mathbf{x}) \]

- Labeled examples are selected from the general population depending on the label itself.
- The missing labels are said to be “missing not at random” (MNAR) in the literature.

Overcoming learnable bias

The training data consist of \( \{(y_i, \mathbf{x}_i, s_i) = 1\} \) and \( \{(y_i, \mathbf{x}_i, s_i) = 0\} \).

Two goals are possible:

- **General population modeling:** Learn \( p(y | \mathbf{x}) \) for loan application approval.

- **Unlabeled population modeling:** Learn \( p(s | \mathbf{x}) = 0 \), e.g. spam filtering.

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

General population modeling

**Lemma 1** ODER MAR bias in the labeling:

\[ p(y, s | \mathbf{x}) = \frac{p(s = 1)}{p(s = 1) + p(s = 0)} p(y | \mathbf{x}, s = 1) + \frac{p(s = 0)}{p(s = 1) + p(s = 0)} p(y | \mathbf{x}, s = 0) \]

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 \( p(s | \mathbf{x}) \) is learnable, we can estimate weights.

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

Unlabeled population modeling

**Lemma 2** ODER MAR bias in the labeling:

\[ p(y, s | \mathbf{x}) = \frac{1 - p(s = 1)}{1 - p(s = 1) + p(s = 0)} p(y | \mathbf{x}, s = 1) + \frac{p(s = 0)}{1 - p(s = 1) + p(s = 0)} p(y | \mathbf{x}, s = 0) \]

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 \( p(s | \mathbf{x}) \) is learnable, we can estimate weights.

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

Overcoming Arbitrary Bias — the Shifted Mixture Model

**Traditional ML semi-supervised learning**

Log-likelihood of semi-labeled data:

\[ f(\theta, \mathbf{X}) = \sum_{i=1}^{m} \log p(y_i | \mathbf{x}_i) + \sum_{i=1}^{n} \log p(x_i | y_i) \]

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

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

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

Extended:

\[ f(\theta, \mathbf{X}) = \sum_{i=1}^{m} \log p(y_i | \mathbf{x}_i, s_i = 1) p(y_i | s_i = 1) + \sum_{i=m+1}^{m+n} \log p(x_i | y_i, s_i = 0) p(s_i = 0) \]

This simplifies to two independent maximizations. But how do we maximize the likelihood in a sensible way for the unlabeled data?

**Experiment 1**

ADULT dataset

- \( k \): (AGE, EDUCATION, CAPITAL GAIN, CAPITAL LOSS, HOURS PER WEEK, SEX, native to US, ETHNICITY, FULL TIME)
- \( y \): INCOME > $50,000?
- MARRIED?

This dataset has information most of which could be used in a loan approval system. The target is analogous to a payable/bad behavior label. MMarital 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 MAR? No.

- \( p(y = 1 | s = 1) = 0.4726 \)
- \( p(y = 1 | s = 0) = 0.0092 \)

Is it M? 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%

CA-HOUSING dataset

- \( x \): CA census tract data: ME-DIAN INCOME, ME-DIAN HOUSE AGE, TOTAL ROOMS, TOTAL BEDROOMS, POP-ULATION, HOUSEHOLDS
- \( y \): house VALUES > Califor-nia median?
- \( x \): LATITUDE > 36 and within 0.4 degrees of coast?

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

Determining the type of bias

Is it MAR? No.

- \( p(y = 1 | s = 1) = 0.751 \)
- \( p(y = 1 | s = 0) = 0.445 \)

Is it M? 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%