CS 165: Foundations of Machine Learning **Winter 2019**

Lecture 16: Methods for Distribution Shift, 3/5/2019

Lecturer: Angie Liu *Scribes: Josef M. Sabuda, Michelle Zhao, Carl Folkestad, Bijan Mazaheri*

16.1 Introduction

Usually in machine learning we assume that the training data and testing data are from the same distribution (i.i.d.), which enables us to derive theoretical guarantees on the learning process. However, in reality, there are usually differences between the training and testing distributions.

The *source* domain is the set of data on which we are training our model. The *target* domain is the domain on which we will be using our model. We will have a phenomenon called *distribution shift* if:

$$P_{\text{source}}(x, y) \neq P_{\text{target}}(x, y) \tag{16.1}$$

Now there are typically two approaches in machine learning; We either want to perform some empirical risk minimization, or use a kind of Minimax approach.

Empirical risk minimization involves minimizing the expected loss over a given function space

$$\min_{f \in F} \mathbb{E}_{\tilde{p}(x,y)}[\text{Loss}(f(X), Y)] \tag{16.2}$$

While in the minimax approach the function f is unconstrained but the distribution p is restricted:

$$\min_{f} \max_{p \in \Gamma(p)} \mathbb{E}_{p(x,y)}[\text{Loss}(f(X), Y)] \tag{16.3}$$

	I.I.D	**Distribution Shift**
Epirical Risk Minimization	Logistic regressin, SVM, boosting	Importance weighting
Minimax Approach	Logistic Regression, Adv 0-1	Robust bias-aware Prediction

A special case of the minimax approach is robust learning, seen earlier in the course. In robust learning, p describes perturbations (possibly adversarial) that we want the learned model to be robust against.

16.2 Empirical Risk Minimization (ERM)

Before we minimize the loss function (16.2), we first need to specify a loss function to minimize. The most "natural" loss function is 0-1 loss. However, this results in a non-convex loss function which is difficult to optimize, so to ease computation different surrogate loss functions are employed instead (e.g. least squares, logloss).

There are two ways to make assumptions about where distribution shift can occur; either on the input variables (Covariate Shift), or on the output variables (Label Shift).

Input variable	Covariate shift	$P_{\text{source}}(x) \neq P_{\text{target}}(x)$ $P_{\text{source}}(y\|x) = P_{\text{target}}(y\|x)$
Output variable	Label shift	$P_{\text{source}}(y) \neq P_{\text{target}}(y)$ $P_{\text{source}}(x\|y) = P_{\text{target}}(x\|y)$

The assumptions on the type of shifts make it possible to simplify the weight estimation. For covariate shift we get the modified weight calculation $\frac{P_{\text{target}}(X,Y)}{P_{\text{source}}(X,Y)} \rightarrow \frac{P_{\text{target}}(X)}{P_{\text{source}}(X)}$. Similarly, for label shift we have $\frac{P_{\text{target}}(X,Y)}{P_{\text{source}}(X,Y)} \rightarrow \frac{P_{\text{target}}(Y)}{P_{\text{source}}(Y)}$.

16.3 Importance Weighting (IM)

Importance weighting seeks to re-weight the source data in order to make it "look like" the target data.

$$\mathbb{E}_{\tilde{P}_{\text{source}}(x,y)}\left[\frac{P_{target}(X,Y)}{P_{source}(X,Y)}\text{Loss}(\hat{f}_\theta(X|Y))\right] \tag{16.4}$$

For covariate shift, we mainly want to estimate the ratio $\frac{P_{\text{target}}(X,Y)}{P_{\text{source}}(X,Y)} \rightarrow \frac{P_{\text{target}}(X)}{P_{\text{source}}(X)}$. We will look at two methods for estimating this ratio; probabilistic classification and moment matching.

Probabilistic Classification

The idea behind probabilistic classification is to separate the numerator and denominator samples. We use Bayes' theorem to calculate the density ratio $r(x)$:

$$\begin{aligned} r(x) &= \frac{p_{\text{nu}}(x)}{p_{\text{de}}(x)} \\ &= \frac{p(x \mid \text{nu})}{p(x \mid \text{de})} \\ &= \frac{p(\text{de})P(\text{nu} \mid x)}{p(\text{nu})P(\text{de} \mid x)} \end{aligned}$$

Probabilistic classification is attractive because off-the-shelf software can be used directly. Furthermore, logistic regression achieves the minimum asymptotic variance for correctly specified models, though it can be unreliable for misspecified models.

Moment Matching

The idea of moment matching is to match the moments (mean, variance, etc) of the distributions. For example, we can match the means of the distributions by shifting the target data such that it gets the same mean as the source distribution. This can be done by finding $\hat{r}(x)$ such that

$$\int x\hat{r}(x)p_{de}(x)dx = \int xp_{nu}(x)dx$$

However, when we have a finite number of moments, completely capturing the relevant characteristics of the distributions is not guaranteed. Instead, kernel mean matching can be used to match all moments and Gaussian RKHS has been shown to efficiently capture all moments.

$$\min_{\hat{r}\in\mathcal{H}} \quad || \int K(x,\cdot)\hat{r}(x)p_{de}(x)dx - \int K(x,\cdot)p_{nu}(x)dx||^2_{\mathcal{H}}$$

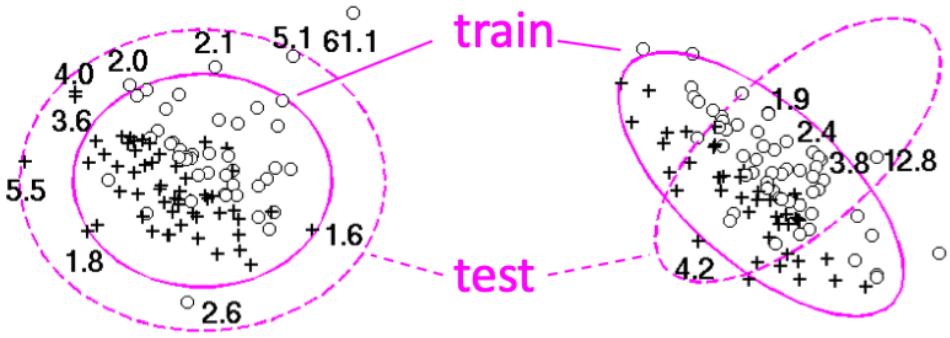
with $K(x,x')$ being the Gaussian kernel. This can be solved by solving an empirical optimization problem as follows.

$$\begin{aligned} \min_{\beta_1,\beta_2,...,\beta_{n_{de}}} \quad & \frac{1}{2}\sum_{j,j'=1}^{n_{de}} \beta_j\beta_{j'}K(x_j^{de},x_{j'}^{de}) - \frac{n_{de}}{n_{nu}}\sum_{j=1}^{n_{de}}\beta_j\sum_{i=1}^{n_{nu}}K(x_i^{nu},x_j^{de}) \\ \text{subject to} \quad & 0 \le \beta_1,\beta_2,...,\beta_{n_{de}} \le B \\ & |\frac{1}{n_{de}}\sum_{j=1}^{n_{de}}\beta_j - 1| \le \epsilon \end{aligned}$$

Note that this is a convex quadratic program (quadratic objective, linear constraints) and the solution directly gives the density ratio estimates as $\hat{\beta}_j = \hat{r}(x_j^{de})$. Obtaining good results with kernel mean matching relies on appropriate tuning of the Gaussian width required. This can be done heuristically by choosing it to be the median distance between samples in the data set. For multi-modal data this may still not be sufficient and does not guarantee good performance.

Problems of Importance Weighting under Covariate Shift

Some general problems with importance weighting for covariate shift problems is that they can result in high variance estimates and slow (or no) convergence. These problems are especially prevalent for small sample sizes.



Weight Estimation under Label Shift

Under label shift, we do not have access to the distributions of the inputs as in the covariate shift case. Instead, the weights rely on the ratio of the distributions of the labels $\frac{P_{\text{target}}(X,Y)}{P_{\text{source}}(X,Y)} \to \frac{P_{\text{target}}(Y)}{P_{\text{source}}(Y)}$. Because we cannot directly estimate this ratio, since we do not have access to the target label distribution, an indirect method is employed.

For any function f and $\forall y, y'$, these are equivalent:

$$f(X) = y'|Y = y, s.t. X \backsim P_{source}$$

$$f(X) = y'|Y = y, s.t. X \backsim P_{target}$$

Applying any f on the covariates of the test domain:

$$P_{target}(f(X) = y') = \sum P_{target}(f(X) = y'|Y = y)P_{target}(Y = y)$$

$$= \sum P_{source}(f(X) = y'|Y = y)P_{target}(Y = y)$$

$$= \sum P_{source}(f(X) = y', Y = y)\frac{P_{target}(Y = y)}{P_{source}(Y = y)}$$

Here, $P_{target}(f(X) = y')$ and $P_{source}(f(X) = y', Y = y)$ can be estimated using training and then testing for X. We can now solve for the weights to obtain

$$P_{target}(f(X) = y') = \sum P_{source}(f(X) = y', Y = y)\frac{P_{target}(y = y)}{P_{source}(Y = y)} \tag{16.5}$$

The remaining material about weight estimation for label shift was not covered in class due to time constraints. For details, see lecture slides for Lecture 16, p. 23-24.

16.4 Minimax Approach

The minimax was not covered in class due to time constraints. See lecture slides for Lecture 16, p. 25-31 for details.