

Sep 18 presentation

- Lipton et al. (2018): Detecting and Correcting for Label Shift with Black Box Predictors (BBSE)
- Azizzadenesheli et al. (2020): Regularized Learning for Domain Adaptation under Label Shifts (RLLS)

Lipton et al. (2018): Detecting and Correcting for Label Shift with Black Box Predictors (BBSE)

Notations and Problem Setup

- $x \in \mathcal{X} = \mathbb{R}^d$ denotes the features, $y \in \mathcal{Y}$ to denote the label variables. For simplicity, we assume that \mathcal{Y} is a **discrete domain** equivalent to $\{1, 2, \dots, k\}$.
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$ drawn iid from a training (or source) distribution P , and $X' = [\mathbf{x}'_1; \dots; \mathbf{x}'_m]$ drawn iid from a test (or target) distribution Q .

Assumptions

- A1: The label shift (also known as target shift) assumption

$$p(\mathbf{x} \mid y) = q(\mathbf{x} \mid y) \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

- A2: For every $y \in \mathcal{Y}$ with $q(y) > 0$ we require $p(y) > 0$.
- A3: Access to a black box predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$ where the expected confusion matrix $\mathbf{C}_p(f)$ is invertible.

$$\mathbf{C}_P(f) := p(f(\mathbf{x}), y) \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}$$

Note: Assumption A3 requires that the expected predictor outputs for each class be linearly independent. This assumption is usually satisfied by a **non-degenerate** classifier.

Idea

Let $\hat{y} = f(\mathbf{x})$, where $f : \mathcal{X} \rightarrow \mathcal{Y}$ is a fixed function.

By the law of total probability and under assumption A1 (label shift) and A2 (common support)

$$\begin{aligned} q(\hat{y}) &= \sum_{y \in \mathcal{Y}} q(\hat{y} \mid y) q(y) \\ &= \sum_{y \in \mathcal{Y}} p(\hat{y} \mid y) q(y) = \sum_{y \in \mathcal{Y}} p(\hat{y}, y) \frac{q(y)}{p(y)}. \end{aligned}$$

- $p(\hat{y} \mid y)$ and $p(\hat{y}, y)$ can be estimated using f and data from source distribution P ,
- $q(\hat{y})$ can be estimated with unlabeled test data drawn from target distribution Q .

BBSE: Black Box Shift Estimation

$$\begin{aligned} [\boldsymbol{\nu}_y]_i &= p(y = i) & [\boldsymbol{\mu}_y]_i &= q(y = i) \\ [\boldsymbol{\nu}_{\hat{y}}]_i &= p(f(\mathbf{x}) = i) & [\boldsymbol{\mu}_{\hat{y}}]_i &= q(f(\mathbf{x}) = i) \\ [\hat{\boldsymbol{\nu}}_{\hat{y}}]_i &= \frac{\sum_j \mathbb{1}\{f(\mathbf{x}_j)=i\}}{n} & [\hat{\boldsymbol{\mu}}_{\hat{y}}]_i &= \frac{\sum_j \mathbb{1}\{f(\mathbf{x}'_j)=i\}}{m} \end{aligned}$$

and $[\boldsymbol{w}]_i = q(y = i)/p(y = i)$. Lastly define the covariance matrices $\mathbf{C}_{\hat{y},y}$, $\mathbf{C}_{\hat{y}|y}$ and $\hat{\mathbf{C}}_{\hat{y},y}$ in $\mathbb{R}^{k \times k}$ via

$$\begin{aligned} [\mathbf{C}_{\hat{y},y}]_{ij} &= p(f(\mathbf{x}) = i, y = j) \\ [\mathbf{C}_{\hat{y}|y}]_{ij} &= p(f(\mathbf{x}) = i \mid y = j) \\ [\hat{\mathbf{C}}_{\hat{y},y}]_{ij} &= \frac{1}{n} \sum_l \mathbb{1}\{f(\mathbf{x}_l) = i \text{ and } y_l = j\} \end{aligned}$$

We can now rewrite the equation in idea slide in matrix form:

$$\boldsymbol{\mu}_{\hat{y}} = \mathbf{C}_{\hat{y}|y} \boldsymbol{\mu}_y = \mathbf{C}_{\hat{y},y} \boldsymbol{w}$$

Using plug-in maximum likelihood estimates of the above quantities yields the estimators

$$\hat{\boldsymbol{w}} = \hat{\mathbf{C}}_{\hat{y},y}^{-1} \hat{\boldsymbol{\mu}}_{\hat{y}} \text{ and } \hat{\boldsymbol{\mu}}_y = \text{diag}(\hat{\boldsymbol{\nu}}_y) \hat{\boldsymbol{w}},$$

The weight vector $\hat{\boldsymbol{w}}$ can be used to reweight the training data and obtain a consistent estimate of the target distribution Q .

Algorithm

input Samples from source distribution X, \mathbf{y} . Unlabeled data from target distribution X' .
A class of classifiers \mathcal{F} . Hyperparameter $0 < \delta < 1/k$.

1. Randomly split the training data into two $X_1, X_2 \in \mathbb{R}^{n/2 \times d}$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^{n/2}$.
2. Use X_1, \mathbf{y}_1 to train the classifier and obtain $f \in \mathcal{F}$.
3. On the hold-out data set X_2, \mathbf{y}_2 , calculate the confusion matrix $\hat{\mathbf{C}}_{\hat{y}, y}$. If,
if $\sigma_{\min} \left(\hat{\mathbf{C}}_{\hat{y}, y} \right) \leq \delta$ then Set $\hat{\mathbf{w}} = \mathbf{1}$. (Method fails)
else Estimate $\hat{\mathbf{w}} = \hat{\mathbf{C}}_{\hat{y}, y}^{-1} \hat{\boldsymbol{\mu}}_{\hat{y}}$.
4. Solve the importance weighted ERM on the X_1, \mathbf{y}_1 with $\max(\hat{\mathbf{w}}, \mathbf{0})$ and obtain \tilde{f} .
output \tilde{f}

Theoretical Guarantees

The authors showed that the estimator performs well in high probability when the number of samples n and m are large.

Theorem (Error bounds). Assume that A3 holds robustly. Let σ_{\min} be the smallest eigenvalue of $\mathbf{C}_{\hat{y},y}$. There exists a constant $C > 0$ such that for all $n > 80 \log(n) \sigma_{\min}^{-2}$, with probability at least $1 - 3kn^{-10} - 2km^{-10}$ we have

$$\begin{aligned}\|\hat{\mathbf{w}} - \mathbf{w}\|_2^2 &\leq \frac{C}{\sigma_{\min}^2} \left(\frac{\|\mathbf{w}\|^2 \log n}{n} + \frac{k \log m}{m} \right) \\ \|\hat{\boldsymbol{\mu}}_y - \boldsymbol{\mu}_y\|^2 &\leq \frac{C \|\mathbf{w}\|^2 \log n}{n} + \|\boldsymbol{\nu}_y\|_{\infty}^2 \|\hat{\mathbf{w}} - \mathbf{w}\|_2^2\end{aligned}$$

Thoughts

- The method relies on **finite sample estimation** of confusion matrix $\mathbf{C}_{\hat{y},y}$ and the distribution of predicted labels on target domain $\mu_{\hat{y}}$, which can have **high variance** when k is large and n, m are small.
- If the **classifier** f is poor, the confusion matrix may be close to **singular** and estimation can be arbitrarily bad.
- From the theorem, the error is **linear** in k .

Regularized Learning for Domain Adaptation under Label Shifts (RLLS)

Azizzadenesheli et al. (2020) and proposed a two-step algorithm to correct for finite sample errors in BBSE, and provided better theoretical guarantees.

Method

In BBSE, we are solving the linear system $q = Cw$ to estimate weights w .

The author defines $\theta = w - \mathbf{1}$, the weight shift vector. Then let $b := q - C\mathbf{1} = C\theta$

Instead of using the finite sample estimate \hat{C} and \hat{b} directly, the authors proposed to solve a **regularized least square problem**:

$$\hat{\theta} = \arg \min_{\theta} \|\hat{C}\theta - \hat{b}\|_2 + \Delta_C \|\theta\|_2$$

where $\Delta_C > 0$ is a regularization parameter. The L2-penalty shrinks the weight shift vector towards zero.

Algorithm

1. calculating the measurement error adjusted $\hat{\theta}$
2. computing the regularized weight $\hat{w} = \mathbf{1} + \lambda \hat{\theta}$ where λ depends on the sample size $(1 - \beta)n_p$.
3. Using the estimated weights to solve the importance weighted ERM.

In particular, for step 2 of the algorithm, we choose $\lambda^* = 1$ whenever

$$n_q \geq \frac{1}{\theta_{\max}^2 \left(\sigma_{\min} - \frac{1}{\sqrt{n_p}} \right)^2} \text{ and 0 else, where } \theta_{\max} \text{ is an upper bound on } \|\theta\|_2 \text{ that we want to}$$

be robust against.

Estimation Error for θ

For $\hat{\theta}$ as defined above, we have with probability at least $1 - \delta$ that

$$\|\hat{\theta} - \theta\|_2 \leq \epsilon_{\theta}(n_p, n_q, \|\theta\|_2, \delta)$$

where

$$\epsilon_{\theta}(n_p, n_q, \|\theta\|_2, \delta) := \mathcal{O} \left(\frac{1}{\sigma_{\min}} \left(\|\theta\|_2 \sqrt{\frac{\log(k/\delta)}{(1-\beta)n_p}} + \sqrt{\frac{\log(1/\delta)}{(1-\beta)n_p}} + \sqrt{\frac{\log(1/\delta)}{n_q}} \right) \right).$$

Generalization Bound for proposed RLLS

Given n_p samples from the source data set and n_q samples from the target set, a hypothesis class \mathcal{H} and loss function ℓ , the following generalization bound holds with probability at least $1 - 2\delta$

$$\mathcal{L}(\hat{h}_{\hat{w}}) - \mathcal{L}(h^*) \leq \epsilon_{\mathcal{G}}(n_p, \delta, \beta) + (1 - \lambda)\|\theta\|_2 + \lambda\epsilon_{\theta}(n_p, n_q, \|\theta\|_2, \delta, \beta).$$

where

$$\epsilon_{\mathcal{G}}(n_p, \delta) := 2\mathcal{R}_n(\mathcal{G}) + \min \left\{ d_{\infty}(q\|p) \sqrt{\frac{\log(2/\delta)}{\beta n_p}}, \frac{2d_{\infty}(q\|p) \log(2/\delta)}{n} + \sqrt{2 \frac{d(q\|p) \log(2/\delta)}{n}} \right\}.$$

Papers on Covariate Shift

1. The papers are more focused on methodology and experiments, with less emphasis on theoretical analysis.
2. Many of them provide generalization bounds but are