Sep 18 presentation

- Liu et al. (2014): Robust Classification Under Sample Selection Bias

- Chen et al. (2016): Robust Covariate Shift Regression

- Reddi et al. (2015): Doubly Robust Covariate Shift Correction

- Slavutsky, Blei (2025): Quantifying Uncertainty in the Presence of Distribution Shifts

Presented by Tengyu Song

Liu et al. (2014): Robust Classification Under Sample Selection Bias

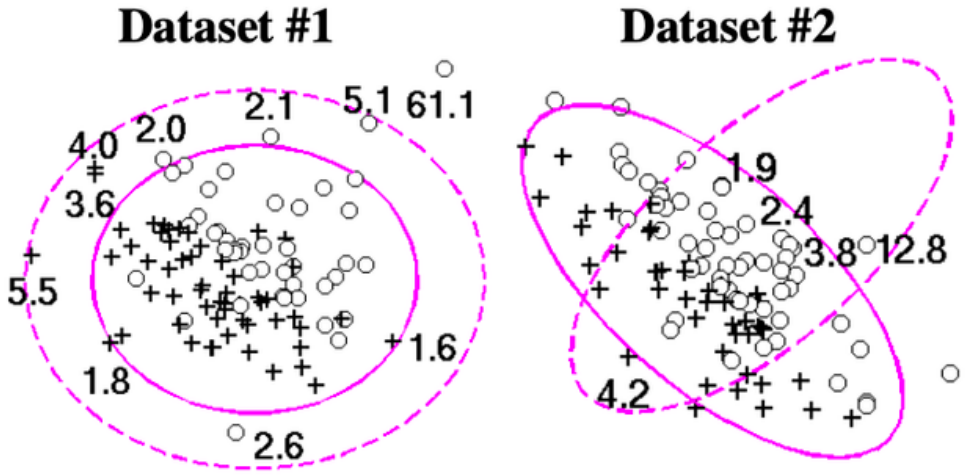
Background and Problem Setup

- For classification tasks under **covariate shift:** $p_{src}(y \mid x) = p_{tar}(y \mid x)$, traditional reweighting approach aims to minimize a reweighted loss on source distribution.
- This can be shown is be consistent under certain assumptions, i.e.

$$\mathbb{E}_{P_{\text{trg}}(x)P(y|x)}\left[\mathcal{L}\left(\hat{P}_\theta(Y \mid X), Y\right)\right] = \lim_{n\to\infty} \mathbb{E}_{\tilde{P}^{(n)}_{\text{scc}}(x)\tilde{P}(y|x)}\left[\frac{P_{\text{trg}}(X)}{P_{\text{src}}(X)}\mathcal{L}\left(\hat{P}_\theta(Y \mid X), Y\right)\right]$$

However, this approach has several limitations:

- Assumption may not hold, the first moment may not exist
- Weights may vary significantly across samples, leading to high variance
- Relies on accurate estimation of density ratio



Authors' approach

- The authors propose a novel minimax approach to minimizes the log loss of against the worst-case distribution subject to known properties of data from the source distribution.
- Importance weighting approach is a special case of their approach for a particularly strong assumption: that source statistics fully generalize to the target distribution

Notations and Problem Setup

We assume that a set of statistics, denoted as convex set Ξ, characterize the source distribution, $P_{\text{src}}(x, y)$.

Then we can define a robust minimax estimate of the conditional label distribution, $\hat{P}(Y \mid X)$, using a worst-case conditional label distribution, $\check{P}(Y \mid X)$.

Definition: robust bias-aware (RBA) probabilistic classifier is the saddle point of:

$$\min_{\hat{P}(Y|X)\in\Delta} \max_{\check{P}(Y|X)\in\Delta\cap\equiv} \text{logloss}_{P_{tr}(X)}(\check{P}(Y \mid X), \hat{P}(Y \mid X)),$$

where Δ is the conditional probability simplex:

$$\forall x \in \mathcal{X}, y \in \mathcal{Y} : P(y \mid x) \geq 0; \sum_{y' \in \mathcal{Y}} P\left(y' \mid x\right) = 1$$

Theorem 1.

Assuming Ξ is a set of moment-matching constraints,

$$\mathbb{E}_{P_{src}(x)\hat{P}(y|x)}[\mathbf{f}(X,Y)] = \mathbf{c} \triangleq \mathbb{E}_{P_{src}(x)P(y|x)}[\mathbf{f}(X,Y)]$$

The solution of the minimax logloss game maximizes the target distribution conditional entropy subject to matching statistics on the source distribution:

$$\max_{\hat{P}(Y|X)\in\Delta} H_{P_{tg}(x),\hat{P}(y|x)}(Y \mid X) \text{ such that: } \mathbb{E}_{P_{src}(x)\hat{P}(y|x)}[\mathbf{f}(X,Y)] = \mathbf{c}$$

By definition of conditional entropy, the solution to this optimization has low certainty where the target density is high by matching the source distribution statistics primarily where the target density is low.

Theorem 2.

The robust bias-aware (RBA) classifier for target distribution $P_{\text{trg}}(x)$ estimated from statistics of source distribution $P_{\text{src}}(x)$ has a form:

$$\hat{P}_\theta(y \mid x) = \frac{e^{\frac{P_{src}(x)}{P_{P_{\text{rg}}}(x)}\theta \cdot \mathbf{f}(x,y)}}{\sum_{y' \in \mathcal{Y}} e^{\frac{P_{sr}(x)}{P_{r_{tg}}(x)}\theta \cdot \mathbf{f}(x,y')}},$$

which is parameterized by Lagrange multipliers θ. The Lagrangian dual optimization problem selects parameters θ to maximize the target distribution log likelihood:

$$\max_\theta \mathbb{E}_{P_{\text{trg}}(x)P(y|x)} \left[\log \hat{P}_\theta(Y \mid X)\right].$$

Other details

1. The authors added regularization when estimating parameter θ since the characteristics of the source distribution Ξ are not precisely known.
2. The authors added regularization when estimating parameter θ since the characteristics of the source distribution Ξ are not precisely known.
3. The authors showed that if there is expert knowledge that reweighted source statistics are rep- resentative of the target distribution, then these strong generalization assumptions should be included as constraints in the RBA predictor and results in the sample reweighted approach
4. No theoretical analysis on the consistency of the proposed method is provided in the paper.

Remarks

1. The main challenge of approach is how to select the best statistics to characterize the original distribution.
2. The minimax approach protect against the worst case scenario, which might leads to poor performance when the shift is mild.

Chen et al. (2016): Robust Covariate Shift Regression

Chen et al. (2016) extend the work of Liu et al. (2014) to regression tasks, and proposed a robust covariate shift regression (RCSR) method.

Definitions

Log-loss for regression tasks is defined as:

$$\text{logloss}_{f_{\text{trg}}(\mathbf{x})}(f(y \mid \mathbf{x}), \hat{f}(y \mid \mathbf{x})) \triangleq \mathbb{E}_{f_{\text{trg}}(\mathbf{x}) f(y|\mathbf{x})}[-\log \hat{f}(Y \mid \mathbf{X})]$$

where

- $f_{\text{trg}}(\mathbf{x})$ is the target domain density of $\mathbf{x}$
- $f(y \mid \mathbf{x})$ is the true conditional likelihood of y given $\mathbf{x}$
- $\hat{f}(y \mid \mathbf{x})$ is the conditional likelihood of y given $\mathbf{x}$ for the estimator

Definitions

The **robust bias-aware regression estimator**, $\hat{f}(y \mid \mathbf{x})$, is the saddle point solution of the following minimax optimization:

$$\min_{\hat{f}(y|\mathbf{x})} \max_{f(y|\mathbf{x}) \in \Xi} \text{rel-loss}_{f_{\text{trg}}(\mathbf{x})} \Big(f(y \mid \mathbf{x}), \hat{f}(y \mid \mathbf{x}), f_0(y \mid \mathbf{x}) \Big).$$

where

$$\begin{aligned}
&\text{rel-loss}_{f_{\text{trg}}(\mathbf{x})} \Big(f(y \mid \mathbf{x}), \hat{f}(y \mid \mathbf{x}), f_0(y \mid \mathbf{x}) \Big) \\
&\triangleq \text{logloss}_{f_{\text{trg}}(\mathbf{x})} (f(y \mid \mathbf{x}), \hat{f}(y \mid \mathbf{x})) - \text{logloss}_{f_{\text{trg}}(\mathbf{x})} (f(y \mid \mathbf{x}), f_0(y \mid \mathbf{x})) \\
&= \mathbb{E}_{f_{\text{trg}}(\mathbf{x}) f(y|\mathbf{x})} \left[- \log \frac{\hat{f}(Y \mid \mathbf{X})}{f_0(Y \mid \mathbf{X})} \right]
\end{aligned}$$

Notes

- $f_0(y \mid \mathbf{x})$ is a **base conditional distribution** that can be estimated from the source distribution.
- In practice the authors proposed to use $\mu_o = \frac{y_{\min}+y_{\max}}{2}, \quad \sigma_o^2 = \left(\frac{y_{\max}-\mu_o}{2}\right)^2$ as the base distribution.
- The authors used moment-matching quadratic interaction features as characteristics of the convex set Ξ:

$$\mathbb{E}_{f_{\text{src}}(\mathbf{x})\hat{f}(y|\mathbf{x})}[\Phi(\mathbf{X}, Y)] = \mathbf{c},$$

$$\Phi(\mathbf{x}, y) = \text{vector}\left(\left[y\mathbf{x}^T 1\right]^T \left[y\mathbf{x}^T 1\right]\right).$$

where $\mathbf{c}$ is the empirical estimate of the above moment on the source distribution.

Rest of the paper

- The following theorems follow the same structure as Liu et al. (2014)
- the authors showed that the the proposed regression estimator can be solved by minimizing the target distribution conditional **KL divergence between the estimator conditional distribution and the base conditionaldistribution** subject to matching statistics on the source distribution.

Remarks

- This is minimax approach. The information from the source distribution is solely provided by the moment-matching constraints.
- When we incorporate the strong assumption that the feature expectation under the target distribution is equivalent to the expectation of reweighted features on source data, RBA is equivalent to importance weighted least squares

$$\mathbb{E}_{f_{\text{trg}}(\mathbf{x})\hat{f}(y|\mathbf{x})}[\Phi(X,Y)] \quad = \quad \tilde{\mathbf{c}}' \quad \triangleq \mathbb{E}_{\tilde{f}_{\text{src}}(\mathbf{x})\tilde{f}(y|\mathbf{x})}\left[\frac{f_{\text{tra}}(X)}{f_{\text{arc}}(X)}\Phi(X,Y)\right]$$

- Reddi et al. (2015): Doubly Robust Covariate Shift Correction

Reddi et al. (2015) proposed a doubly robust covariate shift correction by combining the weighted and unweighted estimates.

The goal is to minimize the expected risk with regard to target distribution p, $R_p[f] := \mathbf{E}_{(x,y)\sim p}[\ell(y, f(x))]$. Let $R_q[f] := \mathbf{E}_{(x,y)\sim q}[\ell(y, f(x))]$ denote expected risk with regard to source distribution q.

Ideas

- First, a regularizer Ω is introduced to prevent overfitting. It measures the complexity of a function f relative to a reference f', with the null function ($f' = 0$) commonly used as the baseline. In kernel methods, this takes the form $\Omega[f, f'] = \frac{1}{2}|f - f'|^2$
- **Unweighted estimator**: $\hat{f}_{q,\lambda} = \arg\min_{f \in \mathcal{F}} \widehat{R}[f \mid X, Y] + \lambda_1 \Omega[f, 0]$
- **Weighted estimator**: $\hat{f}_{\mathrm{w}} = \arg\min_{f \in \mathcal{F}} \widehat{R}[f \mid X, Y, \hat{\beta}] + \lambda_2 \Omega[f, 0]$, the weight can be estimated by various density ratio estimation methods.
- **Doubly robust estimator**: $\hat{f}_{\mathrm{DR}} := \underset{f \in \mathcal{F}}{\operatorname{argmin}} \widehat{R}[f \mid X, Y, \hat{\beta}] \text{ s.t. } \Omega\left[f, \hat{f}_{q,\lambda}\right] \leq \nu'$

  The doubly robust estimator has a prior around $\hat{f}_{q,\lambda}$ rather than 0 .

Effective sample size

- The authors defined the **effective sample size** of the weighted estimator as

$$m_{\text{eff}} := \|\beta(X)\|_1^2 / \|\beta(X)\|_2^2$$

where $\beta(X)$ is the vector of weights $\beta(x_1), \ldots, \beta(x_m)$.

- To gain better intuition for m_{eff}, consider the case where $p = q$. In this case, we have high effective sample size ($m_{\text{eff}} = m$). Whereas in the undesirable case of a single observation having very high weight, $m_{\text{eff}} \approx 1$. Hence, m_{eff} is a good indicator of the effect of $\beta(x)$ on variance of the weighted empirical averages.

Procedure

Step 1: Unweighted estimate Solve the unweighted inference problem using (X, Y) as training data to obtain $\hat{f}_{q,\lambda_q}$ (see Equation (2)).

Step 2: Covariate shift correction weights Using X and X' estimate the covariate shift correction weights. This can be done by any off-the-shelf (e.g. kernel mean matching) covariate shift procedure (Gretton et al. 2008; Agarwal et al. 2011).

Step 3: Doubly robust estimate If m_{eff} is much smaller than m, use unweighted estimate in Step 1 and covariate shift weights in Step 2 to obtain $\hat{f}_{DR}$ (see Equation 4).

Quantifying Uncertainty in the Presence of Distribution Shifts

Adaptive prior and posterior

Traditional Bayesian inference assumes that training and test data are drawn from the same distribution.

$$p\left(y^* \mid x^*, x_{1:n}, y_{1:n}\right) = \int p\left(y^* \mid x^*, \theta\right) p\left(\theta \mid x_{1:n}, y_{1:n}\right) d\theta$$

However, when there is a distribution shift between training and test data, the above posterior may be a poor estimate of the true posterior. The authors proposed an adaptive prior and posterior to address this issue.

$$p\left(y^* \mid x^*, x_{1:n}, y_{1:n}\right) = \int p\left(y^* \mid x^*, \theta\right) p\left(\theta \mid x^*, x_{1:n}, y_{1:n}\right) d\theta$$

Energy based prior

$$E\left(\theta; x_{1:N}, x^*\right) := \int \sum_{i=1}^{N} \log p\left(y \mid x_i, \theta\right) + \log p\left(y \mid x^*, \theta\right) dy$$

$$p\left(\theta \mid x_{1:N}, x^*\right) := \frac{1}{Z(\theta)} \exp\left(E\left(\theta; x_{1:N}, x^*\right)\right),$$

where $Z(\theta)$ is the normalizing constant.

The authors then estimated the posterior using variational inference.

Small sub-samples as synthetic shifts "Inverse Bootstrapping"

- Sample uniformly at random small samples of the training data
- Each will exhibit a different empirical distribution, simulating a shift
- Taking enough samples guarantees that with high probability, one will be close to the true unknown shift
- But which one?
- Since we don't know, we want to encourage good fit on all of them