Sep 10 presentation

Ben-David et al.(2006): Generalization bounds for covariate shift domain adaptation

Label shift: Seong-ho et al. (2024): Double flexible estimators for $E[Y]$ under label shift

Presented by Tengyu Song

Problem Setup in Ben-David (2006) - Binary classification under covariate shift

- **Instance space**: $\mathcal{X}$
- **Source distribution** $\mathcal{D}_S$ over $\mathcal{X}$
- **Target distribution** $\mathcal{D}_T$ over $\mathcal{X}$
- **Label set**: $\mathcal{Y} = \{0, 1\}$
- **Representation function** $\mathcal{R} : \mathcal{X} \to \mathcal{Z}$
- **Induced distributions** $\tilde{\mathcal{D}}_S$ and $\tilde{\mathcal{D}}_T$ over $\mathcal{Z}$
- **Labeling rule**: $f : \mathcal{X} \to [0, 1]$, common to both domains, and $\tilde{f}$ is the induced image of f under $\mathcal{R}$.

Error of a predictor

A predictor is a function, h, from the feature space, $\mathcal{Z}$ to $[0, 1]$. We denote the probability, according the distribution $\mathcal{D}_S$, that a predictor h disagrees with f by

$$\begin{aligned}\epsilon_S(h) &= \mathrm{E}_{\mathbf{z}\sim\tilde{\mathcal{D}}_S}\left[\mathrm{E}_{y\sim\tilde{f}(\mathbf{z})}[y \neq h(\mathbf{z})]\right] \\ &= \mathrm{E}_{\mathbf{z}\sim\tilde{\mathcal{D}}_S}|\tilde{f}(\mathbf{z}) - h(\mathbf{z})|\end{aligned}$$

Similarly, $\epsilon_T(h)$ denotes the expected error of h with respect to $\mathcal{D}_T$.

$\mathcal{A}$-distance

Let $\mathcal{A}$ be a family of subsets of $\mathcal{X}$. The $\mathcal{A}$-distance between two distributions $\mathcal{D}$ and $\mathcal{D}'$ is defined as

$$d_{\mathcal{A}}\left(\mathcal{D}, \mathcal{D}'\right) = 2 \sup_{A \in \mathcal{A}} \left|\Pr_{\mathcal{D}}[A] - \Pr_{\mathcal{D}'}[A]\right|$$

for a binary function class $\mathcal{H}$ we will write $d_{\mathcal{H}}(\cdot, \cdot)$ to indicate the $\mathcal{A}$-distance on the class of subsets whose characteristic (indicator) functions are functions in $\mathcal{H}$.

$$d_{\mathcal{H}}\left(\mathcal{D}, \mathcal{D}'\right) = 2 \sup_{A : 1_A \in \mathcal{H}} \left|\Pr_{\mathcal{D}}[A] - \Pr_{\mathcal{D}'}[A]\right|$$

λ-close

We say that a function $\tilde{f} : \mathcal{Z} \to [0, 1]$ is λ-close to a function class $\mathcal{H}$ with respect to distributions $\tilde{\mathcal{D}}_S$ and $\tilde{\mathcal{D}}_T$ if

$$\inf_{h \in \mathcal{H}} \left[\epsilon_S(h) + \epsilon_T(h) \right] \leq \lambda.$$

- Important assumption for controlling the error on the target domain.

Theorem 1 (Ben-David et al., 2006)

Let $\mathcal{H}$ be a hypothesis space of VC-dimension d. If a random labeled sample of size m is generated by applying R to a $\mathcal{D}_S$-i.i.d. sample labeled according to f, then with probability at least $1-\delta$, for every $h \in \mathcal{H}$:

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + \sqrt{\frac{4}{m}\left(d \log \frac{2em}{d} + \log \frac{4}{\delta}\right)} + d_{\mathcal{H}}\left(\tilde{\mathcal{D}}_S, \tilde{\mathcal{D}}_T\right) + \lambda,$$

where $\lambda = \min_{h \in \mathcal{H}} \left(\epsilon_S(h) + \epsilon_T(h)\right)$.

- first term is the empirical error on the source domain.
- second term bound the deviation of the empirical error from the true error.
- third term measures the divergence between the source and target distributions.
- The last term measures the adaptability of the hypothesis space to the two domains.

Then the authors extend the theorem to finite samples from both domains.

Thoughts

- The generalization bound showed that a good representation need to achieves low error rate on the source domain while also minimizing the $\mathcal{A}$-distance between the induced marginal distributions of the two domains.
- Representation function is fixed instead of learned.
- In practice the adaptation term λ is difficult to compute/verify. (labels unknown in target domain, optimization over h)

Problem setup in Seong-ho et al. (2024)

- i.i.d. observations $\{Y_i, \mathbf{X}_i\}, i = 1, \ldots, n_1$ from population $\mathcal{P}$, and iid observations $\mathbf{X}_j, j = n_1 + 1, \ldots, n_1 + n_0$ from population $\mathcal{Q}$.
- $\pi = n_1/\left(n_1 + n_0\right)$ is the proportion of labeled data. $R = 1$ if the observation is from $\mathcal{P}$ and $R = 0$ if from $\mathcal{Q}$. $\rho(y) = q(y)/p(y)$ is the density ratio of Y between two populations.
- **Label shift**: $p(y) \neq q(y)$, but $P(\mathbf{X} \mid Y) = Q(\mathbf{X} \mid Y)$.
- **Goal**: estimate the mean of Y in the target population $\mathcal{Q}$, i.e., $\theta_0 = E_{\mathcal{Q}}(Y)$ using the information from both populations.

Motivation

$$\theta = \mathrm{E}_q(Y) = \mathrm{E}_p[\rho(Y)Y] = \mathrm{E}[R\rho(Y)Y]/\pi$$

- Y is not observed in target domain
- $\rho(y)$ is hard to estimate, especially when Y is continuous.
- another important quantity is $E[y|X]$. we can try to estimate $E[y|X]$ using the labeled data. But the authors show that even this is not needed.

Idea

By designing an estimator $b^{**}(X)$ s.t. $E[b^{**}(X)|y] = y$, we have

$$\begin{aligned}
&\mathrm{E}\left[\frac{R}{\pi}\rho^*(Y)\left\{Y - b^{\star\star}(\mathbf{X})\right\} + \frac{1-R}{1-\pi}b^{\star\star}(\mathbf{X})\right] \\
=&\mathrm{E}_p\left[\rho^*(Y)\left\{Y - b^{\star\star}(\mathbf{X})\right\}\right] + \mathrm{E}_q\left\{b^{\star\star}(\mathbf{X})\right\} \\
=&\mathrm{E}_p\left[\rho^*(Y)\left[Y - \mathrm{E}\left\{b^{\star\star}(\mathbf{X}) \mid Y\right\}\right]\right) + \mathrm{E}_q\left[\mathrm{E}\left\{b^{\star\star}(\mathbf{X}) \mid Y\right\}\right] \\
=&E_q(Y).
\end{aligned}$$

- The third equality uses the label shift assumption: $X|y$ has the same distribution in both domains.

The authors showed that $b^{**}(X)$ can be estimated by solving the following integral equation:

$$b^{**}(\mathbf{x}) \equiv \frac{\mathrm{E}_p^{\star}\left\{a^{**}(Y)\rho^{*}(Y) \mid \mathbf{x}\right\}}{\mathrm{E}_p^{\star}\left\{\rho^{*2}(Y) \mid \mathbf{x}\right\} + \pi/(1-\pi)\mathrm{E}_p^{\star}\left\{\rho^{*}(Y) \mid \mathbf{x}\right\}}$$

$a^{**}(y)$ is a solution to

$$\mathrm{E}\left[\left.\frac{\mathrm{E}_p^{\star}\left\{a^{**}(Y)\rho^{*}(Y) \mid \mathbf{X}\right\}}{\mathrm{E}_p^{\star}\left\{\rho^{*2}(Y) \mid \mathbf{X}\right\} + \pi/(1-\pi)\mathrm{E}_p^{\star}\left\{\rho^{*}(Y) \mid \mathbf{X}\right\}}\right| y\right] = y.$$

The second equation can be evaluated using the labeled data only non-parametrically.

asymptotic results of the proposed estimator and finite sample experimental results are provided in the paper.

Thoughts

- The authors method is suitable for both classification and regression problems.
- The idea of constructing $b^{**}(X)$ is very interesting. It comes from the efficient influence function of θ. Is such construction necessary? Is there any general recipe to construct such $b^{**}(X)$ for other problems?
- Does the method work well when Y is high-dimensional?

Papers that are interesting

- (SCL) Blitzer, John, et al. "Domain adaptation with structural correspondence learning." EMNLP 2006.
- (Cconformal prediction) Tibshirani, Ryan J. "Conformal prediction under covariate shift." ICML 2019.
- (Influence function) Bickel, P. J., Klaassen, J., Ritov, Y., and Wellner, J. A. (1993), Efficient and Adaptive Estimation for Semiparametric Models, Baltimore: Johns Hopkins University Press