Modern Topics on Statistical Learning Theory Spring 2023

Lecture 10 — Domain Adaptation (Part II) and Domain Generalization

Prof. Qi Lei *Scribe: Wenbo Jing, Aman Singhal, Qin Wen*

1 Review

In the last class, we introduced the domain adaptation and generalization problem, where the model is trained on the source dataset and applied to the target dataset different from the source data. There are three main settings below for domain adaptation dealing with the discrepancy between the source and the target dataset.

- Covariate shift: $\mathbb{P}_X^S \neq \mathbb{P}_X^T, \mathbb{P}_{Y|X}^S = \mathbb{P}_{Y|X}^T$.
- Label shift: $\mathbb{P}_Y^S \neq \mathbb{P}_Y^T, \mathbb{P}_{X|Y}^S = \mathbb{P}_{X|Y}^T$.
- Concept/Model shift: $\mathbb{P}_{Y|X}^S \neq \mathbb{P}_{Y|X}^T$. The relation of $\mathbb{P}_X^S$ and $\mathbb{P}_X^T$ is not specified.

We focus on the covariate shift in this note. Here are three different approaches to solve the problem:

- Reducing the selection bias by sample reweighting
- Invariant representation
- Label propagation

2 Sample Reweighting

Instead of using

$$\widehat{\mathcal{L}}_S(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f_\theta(x_i), y_i\right), \tag{1}$$

the sample weighting uses

$$\widehat{\mathcal{L}}_w(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} w_i \ell\left(f_\theta(x_i), y_i\right) \tag{2}$$

as the loss function, where $w_i = P_T(x_i)/P_S(x_i)$. The P_T, P_S here are density functions if x_is are continuous random variables and probability mass functions if x_is are discrete.

In this way,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x_i\sim P_S}\left[\mathop{\mathbb{E}}_{y_i\sim P_{Y|X}}\widehat{\mathcal{L}}_w(f_\theta)\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{x_i\sim P_S}\left[w_i\mathop{\mathbb{E}}_{y_i\sim P_{Y|X}}\ell\left(f_\theta(x_i),y_i\right)\right]\\
&= \frac{1}{n}\sum_{i=1}^{n}\int_x P_S(x_i)\frac{P_T(x_i)}{P_S(x_i)}\mathop{\mathbb{E}}_{y_i\sim P_{Y|X}}\left[\ell\left(f_\theta(x_i),y_i\right)\right]\mathrm{d}x_i\\
&= \mathop{\mathbb{E}}_{x_i\sim P_T}\left[\mathop{\mathbb{E}}_{y_i\sim P_{Y|X}}\ell\left(f_\theta(x_i),y_i\right)\right] = \mathcal{L}_T(f_\theta),
\end{aligned}$$

where $\mathcal{L}_T(f_\theta)$ is the population risk in the target domain.

Although sample reweighting can help reduce the bias, it faces three major problems below when it comes to the application.

Problem a. The sample reweighting method implies $P_T/P_S < \infty$, so $P_S(x_i) \neq 0, \forall x_i \in \operatorname{supp}(P_T)$. This requires $\operatorname{supp}(P_T) \subset \operatorname{supp}(P_S)$. That is to say, sample reweighting method only applies to the case where the support set of the target domain is already in the source domain, which is a relatively simple task compared to the general cases.

Problem b. It reduces the bias, but it potentially can amplify the variance, i.e., P_T/P_S can be large.

Problem c. It is not clear whether estimating P_T/P_S is simpler than the original problem.

3 Invariant Representation

This is the focus of the last lecture.

$$\mathcal{X} \xrightarrow[f_\theta]{} \mathcal{Y}$$

$$\mathcal{X} \xrightarrow[h\in\mathcal{H}]{} \mathcal{Z} \xrightarrow[w]{} \mathcal{Y}, \quad \text{where } \mathcal{Z} \text{ denotes the representation domain.}$$

$$\mathcal{L}_T(w\circ h) \leq \mathcal{L}_S(w\circ h) + d\left(P^S_{h(x)}, P^T_{h(x)}\right) + \lambda^*_h,$$ (Ben-David et al., 2006)

$$\text{where} \quad \lambda^*_h = \min_{\widehat{w}}\left[\mathcal{L}_T\left(\widehat{w}\circ h\right) + \mathcal{L}_S\left(\widehat{w}\circ h\right)\right], h \text{ is a fixed representation.}$$

- **Method**: Minimize the first 2 terms, i.e., $\mathcal{L}_S(w\circ h) + d\left(P^S_{h(x)}, P^T_{h(x)}\right)$ (Ben-David et al., 2006).
- **Problem**: λ^*_h might explode.

4 Label Propagation

Cai et al. (2021) proposes a new model for subpopulation shift based on label propagation. They introduce a consistency regularization method to ensure the samples with similar semantic meanings to predict similarly (among all unlabeled samples). Suppose we have a good representation function, where samples with similar semantic meaning or can be directly/indirectly connected by data augmentations are close to each other (in L_2 distance). Then the measure of inconsistency is defined by:

$$R_B(g) := \mathbb{P}_{\frac{1}{2}(S+T)}\left[\exists x' \in B(x), \text{ s.t. } g(x) \neq g(x')\right],$$

where $B(x)$ is a set of data augmentations on x that can be viewed as a neighborhood of x. This $R_B(g)$ serves as a consistency regularizer defined for all classifier g. Let μ be the consistency error of the ground truth function g^* , i.e., $\mu = R_B(g^*)$, then the algorithm is given by

$$\widehat{g} = \operatorname*{argmin}_g \mathbb{P}_S\left[g(x) \neq g_{tc}(x)\right], \text{ s.t. } R_B(g) \leq \mu,$$

where g_{tc} denotes the teacher classifier learned on the source dataset. In practice, one can use an alternative objective function by adding a penalty term:

$$\widehat{g} = \operatorname*{argmin}_g \mathbb{P}_S\left[g(x) \neq g_{tc}(x)\right] + \lambda R_B(g).$$

Proof by looking at illustration. A toy example is given by Figure 1. In this picture, $\mu = R_B(g^*) = 0$. As long as g_{tc} is at least 51% correct in each connected component of source distribution, label propogation is perfect on the target domain.

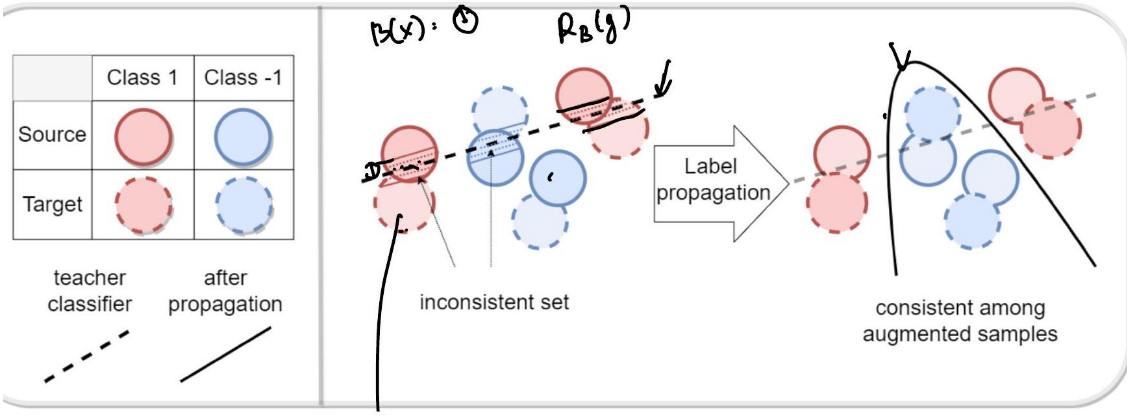


Figure 1: Illustration of the label propagation framework (Cai et al., 2021).

Recall that $\mu = R_B(g^*)$, where g^* is the optimal classifier on $\frac{1}{2}(S+T)$.

$$R_B(g^*) := \mathbb{P}_{\frac{1}{2}(S+T)}\left[\exists x' \in B(x), \text{ s.t. } g^*(x) \neq g^*(x')\right].$$

It quantifies, for all image x, how likely $B(x)$ contains some x' from another class. (μ: misspecification of the optimal classifier under B.)

When $\mu = R_B(g^*) > 0$, Domain Adaptation or B can be too aggressive to connect two classes by mistake. Therefore, we need a technical assumption:

Assumption 1. *Let C_i denote a connected component in $\frac{1}{2}(P_S + P_T)$. Assume that data augmentation is $(1/2, c)$ expansive, i.e., for all set $A \subseteq C_i$,*

$$\mathbb{P}_{C_i}(A) \leq \frac{1}{2} \Longrightarrow \mathbb{P}_{C_i}\left(\bigcup_{x \in A} B(x)\right) \geq (c+1)\mathbb{P}_{C_i}(A),$$

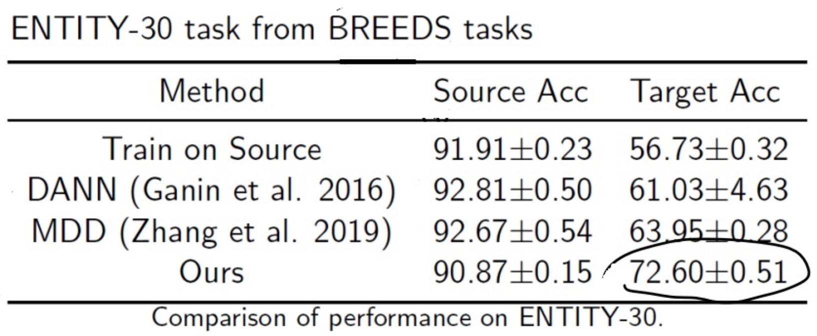
where $\mathbb{P}_{C_i}$ denotes the conditional distribution of $\frac{1}{2}(P_S + P_T)$ on C_i, and $c \leq 1$.

Under Assumption 1, the Label Propagation ensures that the target error of the classifier $\widehat{g}$ returned by the algorithm is bounded by

$$\text{Risk} := \mathop{\mathbb{P}}_{x \sim P_T} [\widehat{g}(x) \neq g^*(x)] = O\left(\frac{\mu}{c}\right).$$

Notice that, since $c \leq 1$, the above bound cannot be better than the ground truth consistency error μ. This result can also extend to finite sample analysis with standard analysis, i.e., the generalization bound results introduced in the first part of this course.

Figure 2 presents an experiment in Cai et al. (2021) showing that label propagation via consistency regularization works well for ENTITY-30, a subpopulation shift benchmark introduced in Santurkar et al. (2020). As compared to the competitors, the target accuracy is improved by around 10% using label propagation (referred to as "ours" in the table).



ENTITY-30 task from BREEDS tasks

Method	Source Acc	Target Acc
Train on Source	91.91±0.23	56.73±0.32
DANN (Ganin et al. 2016)	92.81±0.50	61.03±4.63
MDD (Zhang et al. 2019)	92.67±0.54	63.95±0.28
Ours	90.87±0.15	72.60±0.51

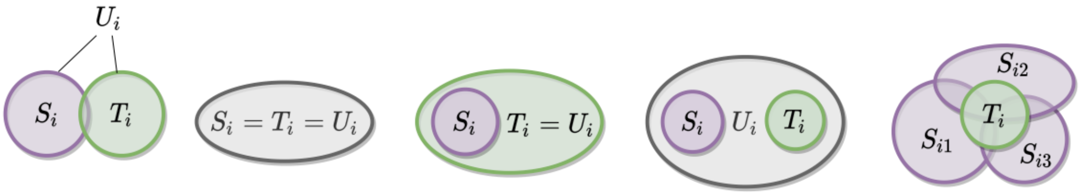
Comparison of performance on ENTITY-30.

Figure 2: Comparison of performance on the ENTITY-30 task (Cai et al., 2021).

5 Data Generalization

In this section, we first introduce some more general domain shift settings. Denote the source distribution by S and the target distribution by T. Moreover, let U be a "covering" distribution with only unlabeled data. Figure 3 (Cai et al., 2021) summarizes some of the generalized settings, including:

- **Unsupervised domain adaptation**. This is the special case discussed in the previous section, i.e., $U = \frac{1}{2}(S + T)$.
- **Semi-supervised learning or self-supervised denoising**. When $S = T = U$, the setting is identical to semi-supervised learning in a single domain.
- **Domain expansion**. The source distribution S, where we have labled data, is only a subset of a larger target distribution T with unlabeled data.
- **Domain extrapolation**. The source distribution S and the target distribution T are not directly connected, which means Assumption 1 is violated. However, they can be connected through a larger distribution U.
- **Multi-Source domain adaptation or domain generalization**. There are multiple different but related source domains S_1, S_2, S_3, and the goal is to learn a model that can generalize to a test distribution T.



(a) Unsupervised domain adaptation (b) Semi-supervised learning or self-supervised denoising (c) Domain expansion (d) Domain extrapolation (e) Multi-source domain adaptation or domain generalization

Figure 3: Settings of generalized subpopulation shift (Cai et al., 2021). The figures only draw one subpopulation i for each model.

Now we focus on the last setting: **domain generalization**. Suppose that we have e different datasets $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{n_t}$ where $(x_i, y_i)_{\in \mathcal{D}_t} \sim P_t$, $t = 1, \ldots, e$. (Here e stands for environment.) We want to learn a learner that is robust to distribution shift or has a good out-of-distribution generalization. The strategies include:

- Learning features that are invariant to all source tasks.
- Ensemble models or representation. Let h_t be a representation function from the t-th source domain S_t to a feature space and w_t be a classifier on this feature space. Ensemble methods build a model of the form $\frac{1}{e}\sum_{t=1}^{e}(w_t \circ h_t)$, while representation learning uses $\frac{1}{e}\sum_{t=1}^{e} h_t$ as an invariant representation.
- "Model soup" proposed by Wortsman et al. (2022). Given multiple models $f_{\theta_1}, \ldots, f_{\theta_e}$, Wortsman et al. (2022) suggest using a single model with averaging parameters, i.e., $f_{\frac{1}{e}(\theta_1 + \cdots + \theta_e)}$. The motivation of model soup comes from an observation that (1) fine-tuned models often appear to lie in a single low error basin and (2) interpolating the parameters of fine-tuned models can improve accuracy compared to any individual model. Unlike conventional ensemble methods, model soup saves inference or memory costs since it only builds a single model.

References

Ben-David, S., J. Blitzer, K. Crammer, and F. Pereira (2006). Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman (Eds.), *Advances in Neural Information Processing Systems*, Volume 19. MIT Press.

Cai, T., R. Gao, J. Lee, and Q. Lei (2021, 18–24 Jul). A theory of label propagation for subpopulation shift. In *Proceedings of the 38th International Conference on Machine Learning*, Volume 139 of *Proceedings of Machine Learning Research*, pp. 1170–1182. PMLR.

Santurkar, S., D. Tsipras, and A. Madry (2020). BREEDS: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*.

Wortsman, M., G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, et al. (2022, 17–23 Jul). Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, Volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR.