Oct 9 presentation

Summary of divergence metrics used in domain adaptation

Presented by Tengyu Song

$\mathcal{A}$-distance and implied $\mathcal{H}$-distance (Ben-David et al., 2006)

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

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

For a binary function class $\mathcal{H}$ we will define $\mathcal{H}$-distance, denoted as $d_{\mathcal{H}}(\cdot, \cdot)$, to indicate the $\mathcal{A}$-distance on the class of subsets whose indicator functions are in $\mathcal{H}$,

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

- The distance was used to derive a generalization bound for domain adaptation. The key ingredient in the bound is $d_{\mathcal{H}}\left(\tilde{\mathcal{D}}_S, \tilde{\mathcal{D}}_T\right)$

KL divergence (Liu et al., 2014; Chen et al., 2016, etc.)

For two distributions P and Q defined on the same probability space, with P absolutely continuous with respect to Q, i.e., P is dominated by Q, the KL divergence is defined as

$$D_{KL}(P\|Q) = \int \log\left(\frac{dP}{dQ}\right) dP$$

The KL divergence was used in the Robust bias-aware (RBA) probabilistic classifier (Liu et al., 2014) and the Robust Covariate Shift Regression (Chen et al., 2016).

The log-loss the author defined is as follows:

$$\operatorname{logloss}_{P_{\text{ug}}(X)}(P(Y \mid X), \hat{P}(Y \mid X)) \triangleq \mathbb{E}_{P_{\text{ug}}(x)P(y|x)}[-\log \hat{P}(Y \mid X)],$$

Rényi divergence (Mansour et al., 2009)

Rényi Divergence which is parameterized by α is defined by

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\log \sum_x P(x)\left(\frac{P(x)}{Q(x)}\right)^{\alpha-1}.$$

- For $\alpha = 1, D_1(P\|Q)$ coincides with the standard relative entropy or KL-divergence.
- For $\alpha = 2, D_2(P\|Q) = \log \mathrm{E}_{x\sim P}\frac{P(x)}{Q(x)}$ is the logarithm of the expected probabilities ratio.
- For $\alpha = \infty, D_\infty(P\|Q) = \log \sup_{x\in\mathcal{X}}\frac{P(x)}{Q(x)}$, which bounds the maximum ratio between the two probability distributions.

The authors usually denote by $d_\alpha(P\|Q)$ the exponential in base 2 of the Rényi divergence:

$$d_\alpha(P\|Q) = 2^{D_\alpha(P\|Q)} = \left[\sum_x \frac{P^\alpha(x)}{Q^{\alpha-1}(x)}\right]^{\frac{1}{\alpha-1}}.$$

A key lemma in the paper is the following:

Lemma 1 For any distributions P and Q, functions f and h and loss L and $\alpha > 1$, the following inequalities hold:

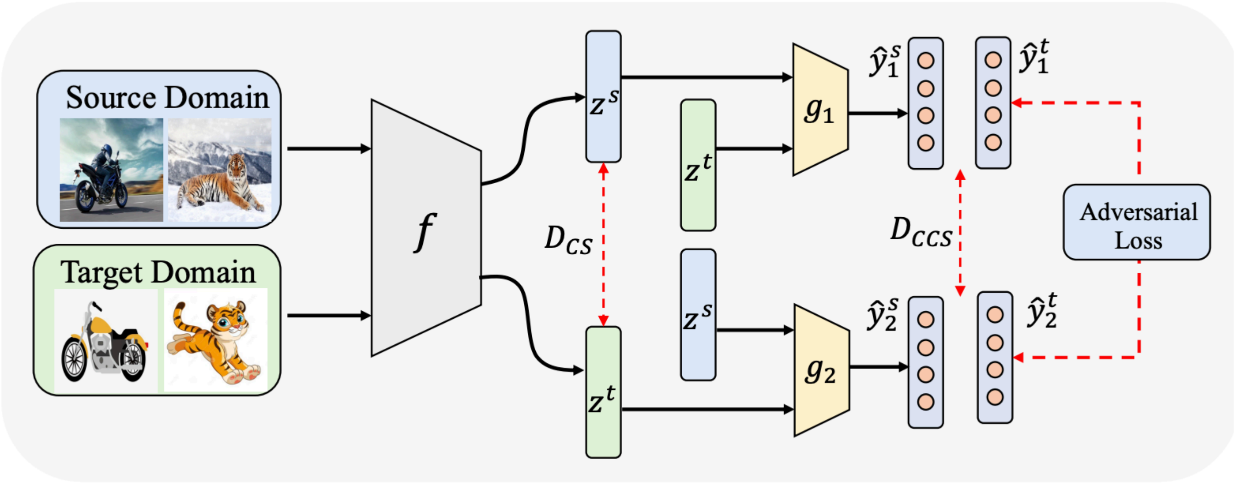
$$\begin{aligned}\mathcal{L}_P(h,f) &\leq \left(d_\alpha(P\|Q) E_{x\sim Q}\left[L^{\frac{\alpha}{\alpha-1}}(h(x), f(x))\right]\right)^{\frac{\alpha-1}{\alpha}} \\ &\leq \left(d_\alpha(P\|Q)\mathcal{L}_Q(h,f)\right)^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}\end{aligned}$$

The lemma bounds the loss of a hypothesis h with respect to a distribution P in terms of the loss of h to another distribution Q and the Rényi divergence between P and Q.

Feature alignment and Adversarial training (Tzeng et al., 2014; Long et al., 2015; Ganin et al., 2016, etc.)

Feature alignment methods try to learn a feature representation such that the source and target distributions are close in the new feature space.

This requires a distance metric to measure the distance between the two distributions in the feature space.



The distance metrics used in these papers include Maximum Mean Discrepancy (MMD) (Tzeng et al., 2014; Long et al., 2015) and domain classifier loss (Ganin et al., 2016), Wasserstein distance (Courty et al., 2017), Jensen-Shannon divergence (Shui et al., 2022), Cauchy-Schwarz (CS) divergence (Yin et al., 2024), Stein discrepancy (Seeger et al., 2025).

Maximum Mean Discrepancy (Tzeng et al., 2014; Long et al., 2015)

The MMD between two distributions P and Q is defined as

$$\mathrm{MMD}(P, Q) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{x \sim Q}[f(x)] \right)$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

Cauchy-Schwarz divergence (Yin et al., 2024)

Cauchy-Schwarz Divergence Motivated by the wellknown Cauchy-Schwarz (CS) inequality for squareintegrable functions:

$$\left(\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}\right)^2 \leq \int p(\mathbf{x})^2 d\mathbf{x} \int q(\mathbf{x})^2 d\mathbf{x},$$

with equality if and only if $p(\mathbf{x})$ and $q(\mathbf{x})$ are linearly dependent, the CS divergence [Principe et al., 2000a,b] defines the distance between probability density functions by measuring the tightness (or gap) of the left-hand side and right-hand side of Eq. (1) using the logarithm of their ratio:

$$D_{\mathrm{CS}}(p;q) = -\log\left(\frac{\left(\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}\right)^2}{\int p(\mathbf{x})^2 d\mathbf{x} \int q(\mathbf{x})^2 d\mathbf{x}}\right).$$

Wasserstein distance (Courty et al., 2017)

The Wasserstein distance between two distributions P and Q is defined as

$$W(P, Q) = \inf_{\gamma \in \Pi(P,Q)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\gamma(x, y)$$

where $\Pi(P, Q)$ is the set of all joint distributions $\gamma(x, y)$ whose marginals are P and Q, and $c(x, y)$ is a cost function.

- The Wasserstein distance can be viewed as the minimum cost of transporting mass in transforming the distribution P into Q.

Jensen-Shannon divergence (Shui et al., 2022)

The Jensen-Shannon (JS) divergence between two distributions P and Q is defined as

$$D_{\text{JS}}(P\|Q) = \frac{1}{2}(D_{\text{KL}}(P\|M) + D_{\text{KL}}(Q\|M))$$

where $M = \frac{1}{2}(P + Q)$ is the mixture distribution of P and Q, and D_{KL} is the KL divergence

- The JS divergence is a symmetrized and smoothed version of the KL divergence.

Stein discrepancy (Seeger et al., 2025)

The Stein discrepancy between two distributions P and Q is defined as

$$D_{\mathrm{Stein}}(P \| Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{x \sim P}[\mathcal{A}_Q f(x)] \right|$$

where $\mathcal{A}_Q$ is the Stein operator associated with Q, and $\mathcal{F}$ is a class of test functions.

Stein operator is defined as

$$\mathcal{A}_Q f(x) = \nabla_x \log q(x) f(x) + \nabla_x f(x)$$

- The Stein discrepancy tests whether q behaves like p using “smart probes” f.
- The Stein discrepancy is zero if and only if $P = Q$.