Orchestrating a brighter world

NEC

Multi-Source Domain Adaptation with von Neumann Conditional Divergence

Ammar Shaker and Shujian Yu

NEC Labs Europe

Problem Definition & Motivation

- Feature similarity includes:
  - the invariance of marginal distributions,
  - and the closeness of conditional distributions given the desired response y
- Traditional methods learn features without fully taking into consideration the information in y, which in turn may lead to a mismatch of the conditional distributions.
- Our main contributions are summarized as follows:
  - We introduce $\boldsymbol{D_{vN}}$ to the problem of domain adaptation (DA) and generalization,
  - and develop a novel learning objective and a generalization bound for multi-source DA.
  - The $\boldsymbol{D_{vN}}$-based objective improves the unsupervised robustness for new tasks.

Background Knowledge

- Let $\mathcal{X}$ and $\mathcal{Y}$ be the input and the output spaces. Given K source domains $\{D_i\}_{i=1}^K$, N_i training samples $\{\boldsymbol{x}_i^j, y_i^j\}_{i=1}^{N_i}$ from $\mathrm{P_i}(\mathbf{x}, \mathrm{y})$ over $\mathcal{X} \times \mathcal{Y}$.
- A network with feature extractor $f_\theta: \mathcal{X} \to \mathcal{T}$ (param. by θ) and a predictor $h_\varphi: \mathcal{T} \to \mathcal{Y}$ (param. by φ).
- The similarity of the latent representation $\boldsymbol{t}$ includes:
  - the invariance of marginal distributions (i.e., $\mathrm{P}(f_\theta(\boldsymbol{x}))$ across different domains, and
  - the functional closeness of using $\boldsymbol{t}$ to predict $\boldsymbol{y}$.
- Goal: generalize a parametric model learned from samples in $\{D_i\}_{i=1}^K$ to an unseen target domain D_{K+1}.

von Neumann Conditional Divergence

- Yu et al. (2020) [1] defines the relative divergence between $P_1(y|\boldsymbol{x})$ and $P_2(y|\boldsymbol{x})$ as:

  $D(P_1(y|\boldsymbol{x})||P_2(y|\boldsymbol{x})) = D_{vN}(\sigma_{xy}||\rho_{xy}) - D_{vN}(\sigma_x||\rho_x)$,

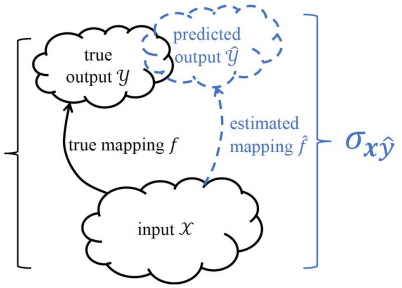
  where σ_{xy}, ρ_{xy}, σ_x and ρ_x denote the covariance matrices. D_{vN} is the von Neumann divergence [2,3]. $D_{vN}(\sigma||\rho) = \mathrm{tr}(\sigma \log \sigma - \sigma \log \rho - \sigma + \rho)$, operating on symmetric positive definite (SPD) matrices σ and ρ.
- To achieve symmetry, the following is formulated:

  $D\big(P_1(y|\boldsymbol{x}): P_2(y|\boldsymbol{x})\big) =$
  $D(P_1(y|\boldsymbol{x})||P_2(y|\boldsymbol{x})) + D(P_2(y|\boldsymbol{x})||P_1(y|\boldsymbol{x}))$

Interpreting the von Neumann Conditional Divergence as a Loss Function

- $P_1(\boldsymbol{x}, y)$ and $P_2(\boldsymbol{x}, y)$ share same marginal $P(\boldsymbol{x})$ i.e., $(\sigma_x = \rho_x)$.
- The symmetric von Neumann conditional divergence reduces to:

  $D(P_1(\boldsymbol{x}, y)||P_2(\boldsymbol{x}, y)) = \frac{1}{2}\,\mathrm{Tr}(\sigma_{\mathrm{xy}} - \rho_{xy})(\log \sigma_{\mathrm{xy}} - \log \rho_{\mathrm{xy}})$
- We term the r.h.s. as the Jeffery von Neumann divergence on σ_{xy} and ρ_{xy} and denote it as $J_{vN}(\sigma_{\mathrm{xy}} : \rho_{xy})$.
- The squared root $\sqrt{J_{vN}(\sigma_{\mathrm{xy}} : \rho_{xy})}$ can be interpreted and used as a loss function to train a deep neural network.
- This is because $J_{vN}(\sigma_{\mathrm{f(x)y}} : \rho_{xy})$ measures the closeness between the true mapping (or labeling) function f and the estimated predictor $\hat{f}$.



Our Method: Multi-source Domain Adaptation With Matrix-Based Discrepancy Distance

Definition 1:

The matrix-based discrepancy distance ($\mathrm{D_{M-disc}}$) measures the longest distance between two domains (with respect to $\mathcal{H}$) in a metric space equipped with the square root of $\mathrm{J_{vN}}$ as a distance function. Given $\mathrm{D_s}$ and $\mathrm{D_t}$ and their distributions $\mathrm{P_s}$ and $\mathrm{P_t}$, for any two hypotheses $\mathrm{h, h'_s} \in \mathcal{H}$, $\mathrm{D_{M-disc}}$ takes the form: $\mathrm{D_{M-disc}}\,(P_s, P_t) = \max_{h,h' \in \mathcal{H}} \left| \sqrt{J_{vN}(\sigma^S_{x,h(x)} : \sigma^S_{x,h'(x)})} - \sqrt{J_{vN}(\sigma^t_{x,h(x)} : \sigma^t_{x,h'(x)})} \right|$, with $a \in \{s, t\}$ and $g \in \{h, h'\}$, the matrix $\sigma^a_{x,g(x)}$ is the covariance matrix for the pair of variables $\boldsymbol{x}, g(\boldsymbol{x})$ in domain $\mathrm{D_a}$.

Theorem 2:

Given a set of K source domains $S = \{D_{s_1}, \dots, D_{s_K}\}$, for any hypothesis $h \in \mathcal{H}$, the square root of J_{vN} on the target domain D_t is bound as follows: $\sqrt{J_{vN}(\sigma^t_{x,h(x)} : \sigma^t_{x,f_t(x)})} \leq$

$\sum_{i=1}^K w_i \left(\sqrt{J_{vN}\left(\sigma^{s_i}_{x,h(x)} : \sigma^{s_i}_{x,f_{s_i}(x)}\right)} \right) + \mathrm{D_{M-disc}}(P_t, P_\alpha; h) +$

$\eta_Q(f_\alpha, f_t)$, where f_t is the ground truth mapping function for D_{s_i} associated with the weight w_i, f_α is the convex combination of all functions f_{s_i}, $\eta_Q(f_\alpha, f_t)$ is the minimum joint empirical losses on the source D_α and the target D_t, achieved by an optimal hypothesis h^*.

Optimization by Adversarial Min-Max Game

We explicitly implement the idea exhibited by Theorem 2 and combine a feature extractor $f_\theta: \mathcal{X} \to \mathcal{T}$ and a class of predictor $h: \mathcal{T} \to \mathcal{Y}$ in a unified learning framework:

$$\min_{f_\theta, h \in \mathcal{H}} \max_{||\alpha||=1} \left(\sum_{i=1}^K w_i \left(J_{vN}\left(\sigma^{s_i}_{x,h(f_\theta(x))} : \sigma^{s_i}_{x,y}\right)\right) + \left| \sum_{i=1}^K w_i \sqrt{J_{vN}(\sigma^{s_i}_{f_\theta(x),h(f_\theta(x))} : \sigma^{s_i}_{f_\theta(x),h'(f_\theta(x))})} - \sqrt{J_{vN}(\sigma^{t}_{f_\theta(x),h(f_\theta(x))} : \sigma^{t}_{f_\theta(x),h'(f_\theta(x))})} \right| \right).$$

The general idea is to find a feature extractor $f_\theta(\boldsymbol{x})$ that for any given pair of h and h', it is hard to discriminate the target domain P_t from the weighted combination of source distribution P_α.

Experiment I: Amazon Review Data

	AHD-MSDA	MDAN-Max	MDAN-Dyn	MDD
ba	0.586(0.003)	0.591(0.015)	0.711(0.006)	**0.583**(0.007)
be	0.608(0.005)	0.628(0.003)	0.656(0.004)	**0.591**(0.005)
ca	0.534(0.006)	0.522(0.005)	0.598(0.006)	**0.508**(0.004)
co	0.61(0.004)	0.682(0.016)	0.829(0.055)	**0.605**(0.008)
el	0.657(0.002)	0.654(0.001)	0.670(0.003)	**0.651**(0.002)
go	0.566(0.003)	0.552(0.003)	0.553(0.003)	**0.549**(0.001)
gr	0.527(0.002)	0.519(0.002)	0.538(0.003)	**0.514**(0.007)

Performance comparison in terms of MAE over 5 iterations on the Amazon data.
ba:baby, **be**:beauty, **ca**:camera, **co**:computer, **el**:electro., **go**:gourmet, **gr**:grocery.

Experiment 2: Year Prediction MSD Data

	AHD-MSDA	MDAN-Max	MDAN-Dyn	MDD
Dom1	7.04(0.07)	18.1(9.2)	16.8(8.6)	**6.69**(0.18)
Dom2	8.28(0.02)	42.8(14.7)	43.4(14.7)	**8.14**(0.11)
Dom3	7.8(7.4)	33.4(9)	33.8(9.4)	**7.69**(0.12)
Dom4	**7.61**(0.04)	28.5(10.2)	29.9(11)	7.63(0.04)
Dom5	7.5(0.05)	23.5(8.3)	24.6(8)	**7.45**(0.09)

Performance comparison in terms of mean absolute error (MAE) over five iterations on YearPredictionMSD data.

Conclusion

- Introduced the von Neumann conditional divergence D_{vN} to match the functional similarity of latent representation t to response variable y across domains.
- Derived a new generalization bound based on a new loss induced by D_{vN} that gives robustness guarantees.

References

[1] Y. Yu, A. Shaker, F. Alesiani, and J. C. Principe. Measuring the discrepancy between conditional distributions: Methods, properties and applications. In IJCAI 2020.
[2] M. A. Nielsen and I. Chuang. Quantum computation and quantum information, 2002.
[3] B. Kulis, M. A. Sustik, and I. S. Dhillon. Low-rank kernel learning with bregman matrix divergences. JMLR 2009.