READING REPORT ON OPTIMAL TRANSPORT

for 3 papers

Guangyu Hou

Contents

1 Joint distribution optimal transportation for domain adaptation

1.1 Summary

The paper deals with the unsupervised domain adaptation problem with the assumption: The non-linear transformation $\mathcal{T}$ between the source and target domain can be estimated with optimal transport. They propose find a prediction function f that minimizes the optimal transport loss between the joint source distribution $\mathcal{P}_f$ and an estimated target joint distribution $\mathcal{P}_{\sqcup}^{\{} = (X, f(X))$. They've proved an upper bound of the target error of f (Theorem 3.1) **but is not uniform w.r.t.** f. It can be seen as an extension of the paper *Optimal Transport for Domain Adaptation* I read last week.

> Q1: The writer mentions that previous approach under **covariate shift** assumption requires that the distributions share a common *support* to be defined. What is the support here?
> Q2: It seems the definition of the *image measure* is a typo where μ_t should be μ_s

1.2 Detail

The main idea is to adapt both marginal feature and conditional distributions by minimizing a global divergence between them. (Minimize simultaneously)Here the problem is shifted to the derive

$$\gamma_0 = \underset{\gamma \in \Pi(\mathcal{P}_s, \mathcal{P}_t)}{\operatorname{argmin}} \int_{(\Omega \times \mathcal{C})^2} \mathcal{D}\left(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2\right) d\gamma\left(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2\right)$$

where $\mathcal{D}\left(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2\right) = \alpha \mathrm{d}\left(\mathbf{x}_1, \mathbf{x}_2\right) + \mathcal{L}\left(y_1, y_2\right)$ and the first term is to control the marginal distributions while the latter one is for conditional distributions. **(The joint cost is separable here)**.

They use $Figure1$ as an example, but how to derive the matrix link and when the transformation is non-linear, can we still get the matrix?

Given the definition of the error and Probabilistic Transfer Lipschitzness, they derive a upper bound under the strong assumption.

Finally, by adding a regularization term, we get a optimization problem with smoothness (either fixed f or γ).

1.3 Pictures

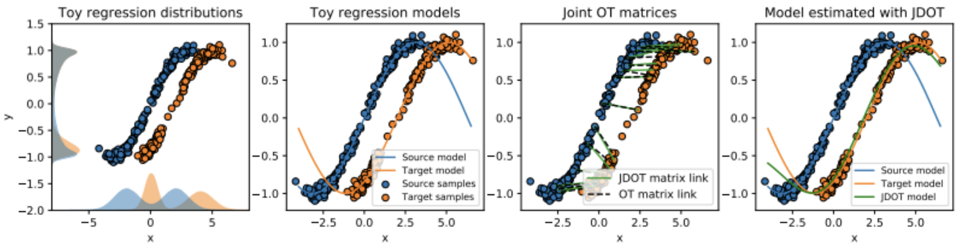

Figure 1: Illustration of JDOT on a 1D regression problem. (left) Source and target empirical distributions and marginals (middle left) Source and target models (middle right) OT matrix on empirical joint distributions and with JDOT proxy joint distribution (right) estimated prediction function f.

2 Sinkhorn Distances: Lightspeed Computation of Optimal Transport

2.1 Summary

Compute the optimal problem with regularization term through Sinkhorn's matrix scaling algorithm as the execution of Sinkhorn's algorithm only relies on matrix-vector products.

2.2 Entropic Constraints on Joint Probabilities

definition 1. Transport Polytope

$$U(r,c) := \left\{P \in \mathbb{R}_+^{d\times d} \mid P\mathbf{1}_d = r, P^T\mathbf{1}_d = c\right\}, \Sigma_d := \left\{x \in \mathbb{R}_+^d \mid x^T\mathbf{1}_d = 1\right\}$$

$$\forall r, c \in \Sigma_d, \forall P \in U(r,c), h(P) \le h(r) + h(c),$$
$$U_\alpha(r,c) := \left\{P \in U(r,c) | \mathbf{KL}\left(P \| rc^T\right) \le \alpha\right\} = \{P \in U(r,c) \mid h(P) \ge h(r) + h(c) - \alpha\} \subset U(r,c).$$

Claim that $h\left(\boldsymbol{r}\boldsymbol{c}^\top\right) = h(\boldsymbol{r}) + h(\boldsymbol{c})$. Since $h(\boldsymbol{p}) = -\sum_i p_i \log p_i$, then

$$\begin{aligned}
h\left(\boldsymbol{r}\boldsymbol{c}^\top\right) &= -\sum_{i,j} r_i c_j \log\left(r_i c_j\right) \\
&= -\sum_{i,j} r_i c_j \left(\log\left(r_i\right) + \log\left(c_j\right)\right) \\
&= -\sum_{i,j} r_i c_j \log\left(r_i\right) - \sum_{i,j} r_i c_j \log\left(c_j\right) \\
&= -\sum_{i} r_i \log\left(r_i\right) - \sum_{j} c_j \log\left(c_j\right) \\
&= h(\boldsymbol{r}) + h(\boldsymbol{c})
\end{aligned}$$

. Thus, we only need to prove $h(P) \le h\left(\boldsymbol{r}\boldsymbol{c}^\top\right)$ (By Lagrange Multiplier).

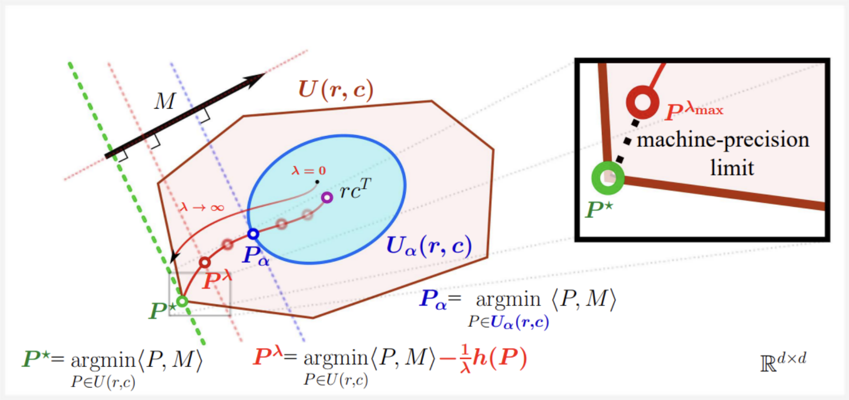
M
$U(r,c)$
$\lambda = 0$
rc^T
$\lambda \to \infty$
$U_\alpha(r,c)$
P_α
P^λ
$P^\star$
$P^{\lambda_{\max}}$
machine-precision limit
$P^\star$
$P_\alpha = \underset{P\in U_\alpha(r,c)}{\operatorname{argmin}} \langle P, M\rangle$
$P^\star = \underset{P\in U(r,c)}{\operatorname{argmin}} \langle P, M\rangle$
$P^\lambda = \underset{P\in U(r,c)}{\operatorname{argmin}} \langle P, M\rangle - \frac{1}{\lambda}h(P)$
$\mathbb{R}^{d\times d}$

Figure 1: Transport polytope $\boldsymbol{U(r,c)}$ and Kullback-Leibler ball $\boldsymbol{U_\alpha(r,c)}$ of level α centered around rc^T. This drawing implicitly assumes that the optimal transport $\boldsymbol{P^\star}$ is unique. The Sinkhorn distance $d_{M,\alpha}(r,c)$ is equal to $\langle \boldsymbol{P_\alpha}, M\rangle$, the minimum of the dot product with M on that ball. For α large enough, both objectives coincide, as $\boldsymbol{U_\alpha(r,c)}$ gradually overlaps with $\boldsymbol{U(r,c)}$ in the vicinity of $\boldsymbol{P^\star}$. The dual-sinkhorn distance $d_M^\lambda(r,c)$, the minimum of the transport problem regularized by minus the entropy divided by λ, reaches its minimum at a unique solution $\boldsymbol{P^\lambda}$, forming a regularization path for varying λ from rc^T to $\boldsymbol{P^\star}$. For a given value of α, and a pair (r,c) there exists $\lambda \in [0,\infty]$ such that both $d_M^\lambda(r,c)$ and $d_{M,\alpha}(r,c)$ coincide. d_M^λ can be efficiently computed using Sinkhorn's fixed point iteration (1967). Although the convergence to $\boldsymbol{P^\star}$ of this fixed point iteration is theoretically guaranteed as $\lambda \to \infty$, the procedure cannot work beyond a problem-dependent value $\lambda_{\max}$ beyond which some entries of $e^{-\lambda M}$ are represented as zeroes in memory.

Combined the two given constraints, we know $h(\boldsymbol{r}) + h(\boldsymbol{c}) - \alpha \le h(P) \le h(\boldsymbol{r}) + h(\boldsymbol{c})$ i.e. $h(P)$ is constrained within a α-length interval. By definition, *mutual information* $I(\boldsymbol{r}\|\boldsymbol{c}) = KL\left(P\|\boldsymbol{rc}^\top\right) \le \alpha$, s.t. r and c become less independent. Also, one can know $h(P) \ge \frac{1}{2}(h(r) + h(c))$

Possible reason for adding constraints: the matrix in the OT problem is quite sparse, then the optimal transport is usually a vertex of the whole $U(r,c)$. By doing so, one can sufficiently smooth the problem which is better to solve.

definition 2. Sinkhorn Distance:

$$d_{M,\alpha}(r,c) := \min_{P\in U_\alpha(r,c)} \langle P, M\rangle$$

Q3: why does the author mention the negative definite kernel and negative definite distance when $\alpha = 0$?

Since for α small enough $d_{M,\alpha}(r,r) > 0$ for any r such that $h(r) > 0$, $d_{M,0}(\boldsymbol{r},\boldsymbol{r}) = \boldsymbol{r}^\top M\boldsymbol{r} = \sum_{ij}^n r_i r_j \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$. If $r_k \ne 0, r_l \ne 0, k \ne l$, then $r_k r_l > 0$ and $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 > 0$, hence $d_{M,0}(\boldsymbol{r},\boldsymbol{r}) > 0$. Thanks to indicator function, we regain the coincide axiom.

2.3 Computing Regularized Transport with Sinkhorn's Algorithm

They consider in this section a Lagrange multiplier for the entropy constraint of Sinkhorn distances:

$$\text{For } \lambda > 0, d_M^\lambda(r, c) := \left\langle P^\lambda, M \right\rangle, \text{ where } P^\lambda = \underset{P \in U(r,c)}{\operatorname{argmin}} \langle P, M \rangle - \frac{1}{\lambda} h(P).$$

By duality theory they have that to each α corresponds a $\lambda \in [0, \infty]$ such that $d_{M,\alpha}(r, c) = d_M^\lambda(r, c)$ holds for that pair (r, c). They call d_M^λ the dual-Sinkhorn divergence and show that it can be computed for a much cheaper cost than the original distance d_M.

Computing d_M^λ with Matrix Scaling Algorithms. Based on Sinkhorn theorem, one can decompose the solution and derive *lemma* 2 therefore realise a efficient algorithm.

Q4: Since the entropy of P^λ decreases monotonically with λ, computing $d_{M,\alpha}$ can be carried out by computing d_M^λ with increasing values of λ until $h\left(P^\lambda\right)$ ***reaches*** $h(r) + h(c) - \alpha$***.(?)***

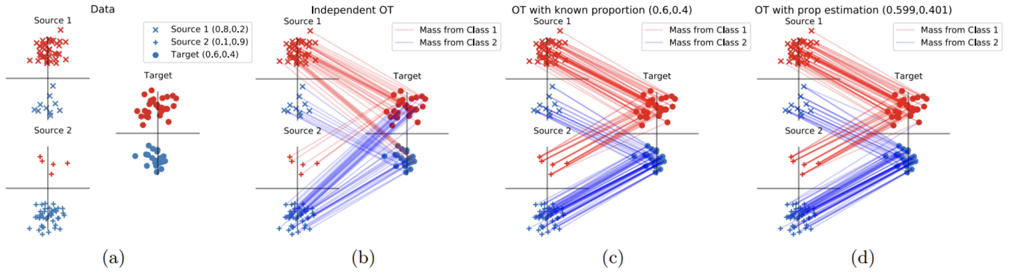

Figure 1: Illustration of the importance of proportion estimation for target shift: (a) the data of 2 source and 1 target domains with different class proportions is visualized; (b) DA method based on OT (Courty *et al.* , 2014) transports instances across different classes due to class proportions imbalance; (c) the transportation obtained when the true class proportions are used to reweigh instances; (d) transportation obtained with JCPOT that is nearly identical to the one obtained with an a priori knowledge about the class proportions.

3 Optimal Transport for Multi-source Domain Adaptation under Target Shift

3.1 Summary

In this paper, they tackle the problem of reducing discrepancies between multiple domains, i.e. *multi-source domain adaptation*, and consider it under the **target shift** assumption: in all domains we aim to solve a classification problem with the same output classes, but with different labels proportions (conditional distributions are same).

Approach: Joint Class Proportion and Optimal Transport (JCPOT), which performs multi-source adaptation and target shift correction simultaneously, while former DA method fails to restrict the transportation of mass across instances of different classes when the class proportions of source and target domains differ.

Q5: When given multi-source data, is it possible to find a common state space where source and target data have best classification performance instead of doing projection towards to target domain.

3.2 DOMAIN ADAPTATION UNDER THE TARGET SHIFT

After setting up a binary classification problem with K source domains, they define a domain as a pair consisting of a distribution P_D on some space of inputs Ω and a labeling function $f_D : \Omega \to [0, 1]$. A hypothesis class $\mathcal{H}$ is a set of functions so that $\forall h \in \mathcal{H}, h : \Omega \to \{0, 1\}$. Given a convex loss-function l, the true risk with respect to the distribution P_D, for a labeling function f_D (which can also be a hypothesis) and a hypothesis h is defined as

$$\epsilon_D\left(h, f_D\right) = \mathbb{E}_{x \sim P_D}\left[l\left(h(x), f_D(x)\right)\right].$$

In the multi-source case, when the source and target error functions are defined w.r.t. h and $f_S^{(k)}$ or f_T, they use the shorthand $\epsilon_S^{(k)}\left(h, f_S^{(k)}\right) = \epsilon_S^{(k)}(h)$ and $\epsilon_T\left(h, f_T\right) = \epsilon_T(h)$. The ultimate goal of multi-source DA then is to learn a hypothesis h on K source domains that has the best possible performance in the target one.

To this end, they define the combined error of source domains as a weighted sum of source domains error functions:

$$\epsilon_S^{\boldsymbol{\alpha}} = \sum_{k=1}^{K} \alpha_k \epsilon_S^{(k)}, \sum_{k=1}^{K} \alpha_k = 1, \alpha_k \in [0,1] \forall k \in [1, \ldots, K].$$

ps. denote by f_S^{α} the labeling function associated to the distribution mixture $P_S^{\alpha} = \sum_{k=1}^{N} \alpha_j P_S^j$. Here different weights α_k can be seen as measures reflecting the proximity of the corresponding source domain distribution to the target one.

Q6: The definition of error needs true label functions, therefore, it can only be used under supervised learning assumption? (At the end of the paper, the author offers 2 ways to classify: **Barycentric mapping**(directly learning) and **Label propagation**(by OT matrix know the source of each target point))

Let $\operatorname{disc}_l\left(P_S, P_T\right) = \max_{h, h' \in \mathcal{H}} \mid \epsilon_S\left(l\left(h, h'\right)\right) - \epsilon_T\left(l\left(h, h'\right)\right) \mid$ be the discrepancy distance between two probability distributions P_S and P_T. Then, for any fixed $\boldsymbol{\alpha}$ and for any $h \in \mathcal{H}$ the following holds:

$$\epsilon_T(h) \leq \epsilon_S^{\boldsymbol{\alpha}}(h) + \left| \pi_T - \sum_{j=1}^{N} \alpha_j \pi_S^j \right| \operatorname{disc}\left(P_0, P_1\right) + \lambda,$$

where $\lambda = \min_{h \in \mathcal{H}} \epsilon_S^{\alpha}(h) + \epsilon_T(h)$ represents the joint error between the combined source error and the target one

The second term in the bound can be minimized for any α_k when $\pi_T = \pi_S^k, \forall k$. This can be achieved by using a proper reweighting of the class distributions in the source domains, but **requires to have access to the target proportion** which is assumed to be unknown. Therefore, they propose to estimate optimal proportions by minimizing the sum of the Wasserstein distances between all reweighted sources and the target distribution. **First**, they prove below that the minimization of the Wasserstein distance between a weighted source distribution and a target distribution yields the optimal proportion estimation. To proceed, let us consider the multi-class problem with C classes, where the target distribution is defined as

$$P_T = \sum_{i=1}^{C} \pi_i^T P_i$$

with P_i being a distribution of class $i \in \{1, \ldots, C\}$. As before, the source distribution with weighted classes can be then defined as

$$P_S^{\pi} = \sum_i \pi_i P_i$$

where $\pi \in \Delta_C$ are coefficients lying in the probability simplex $\Delta_C \overset{\text{def}}{=} \left\{\alpha \in \mathbb{R}_+^C : \sum_{i=1}^{C} \alpha_i = 1\right\}$ that reweigh the corresponding classes. As the proportions of classes in the target distribution are unknown, our goal is to reweigh source classes distributions by solving the following optimization problem:

$$\pi^\star = \underset{\pi\in\Delta_C}{\arg\min} W\left(P_S^\pi, P_T\right).$$

We can now state the following proposition.

Assume that $\forall i, \nexists \alpha \in \left\{\Delta_C \mid \alpha_i = \ 0, P_i = \sum_j \alpha_j P_j\right\}$. Then, for any distribution P_T, the unique solution π^* minimizing the above equation is given by π^T.

3.3 JCPOT

For every source domain, they assume that its data points follow a probability distribution function or probability measure $\mu^{(k)}\ \left(\int \mu^{(k)} = 1\right)$.
In real-world situations, $\mu^{(k)}$ is only accessible through the instances $\mathbf{x}_i^{(k)}$ that we can use to define a distribution $\mu^{(k)} = \sum_{i=1}^{n^{(k)}} m_i^{(k)} \delta_{\mathbf{x}_i^{(k)}}$, where $\delta_{\mathbf{x}_i^{(k)}}$ are Dirac measures located at $\mathbf{x}_i^{(k)}$, and $m_i^{(k)}$ is an associated probability mass. Then one can write $\mu^{(k)} = (\mathbf{m}^{(k)})^T \delta_{\mathbf{X}^{(k)}}$. (Note that when the data set is a collection of independent data points, the weights of all instances in the sample are usually set to be equal) be equal.
In this work, however, they use different weights for each class of the source domain so that they can adapt the proportions of classes w.r.t. the target domain. As $\mu^{(k)} = \sum_{c=1}^{C} \mu_c^{(k)}$ among the C classes, they denote by $h_c^{(k)} = \int \mu_c^{(k)}$ the proportion of class c in $\mathbf{X}^{(k)}$. By construction, we have $h_c^{(k)} = \sum_{i=1}^{n^{(k)}} \delta\left(y_i^{(k)} = c\right) m_i^{(k)}$. Since they chose to have equal weights in the classes, they define two linear operators $\mathbf{D}_1^{(k)} \in \mathbb{R}^{C\times n^{(k)}}$ and $\mathbf{D}_2^{(k)} \in \mathbb{R}^{n^{(k)}} \times C$ that allow to express the transformation from the vector of mass $\mathbf{m}^{(k)}$ to the class proportions $\mathbf{h}^{(k)}$ and back:

$$\mathbf{D}_1^{(k)}(c,i) = \begin{cases} 1 & \text{if } y_i^{(k)} = c, \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathbf{D}_2^{(k)}(i,c) = \begin{cases} \frac{1}{\#\left\{y_i^{(k)}=c\right\}_{i=\left\{1,\ldots,n^{(k)}\right\}}} & \text{if } y_i^{(k)} = c \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbf{D}_1^{(k)}$ allows to retrieve the class proportions with $\mathbf{h}^{(k)} = \mathbf{D}_1^{(k)}\mathbf{m}^{(k)}$ and $\mathbf{D}_2^{(k)}$ returns weights for all instances for a given vector of class proportions with $\mathbf{m}^{(k)} = \mathbf{D}_2^{(k)}\mathbf{h}^{(k)}$, where the masses are distributed equiproportionnally among all the data points associated to one class. Then the optimization problem can be written as follows:

$$\underset{\mathbf{h}\in\Delta_C}{\arg\min} \sum_{k=1}^{K} \lambda_k W_{\epsilon, C^{(k)}}\left(\left(\mathbf{D}_2^{(k)}\mathbf{h}\right)^T \delta_{\mathbf{X}^{(k)}}, \mu\right),$$

where regularized Wasserstein distances are defined as

$$W_{\epsilon, C^{(k)}}\left(\mu^{(k)}, \mu\right) \overset{\text{def}}{=} \min_{\gamma^{(k)}\in\Pi\left(\mu^{(k)},\mu\right)} \mathrm{KL}\left(\gamma^{(k)} \mid \zeta^{(k)}\right),$$

provided that $\zeta^{(k)} = \exp\left(-\frac{C^{(k)}}{\epsilon}\right)$ with λ_k being convex coefficients ($\sum_k \lambda_k = 1$) accounting for the relative importance of each domain. Here, they define the set $\Gamma = \left\{\gamma^{(k)}\right\}_{k=1\ldots K} \in \left(\mathbb{R}^{\ltimes^{(k)} \times n}\right)^K$ as the set of couplings between each source and the target domains. This problems leads to K marginal constraints $\gamma_k^T \mathbf{1}_n = \mathbf{1}_n/n$ w.r.t. the uniform target distribution, and K marginal constraints $\mathbf{D}_1^{(k)} \gamma_k \mathbf{1}_n = \mathbf{h}$ related to the unknown proportions $\mathbf{h}$.

After calculating independently at first, the remaining K constraints require to solve the proposed optimization problem for Γ and $\mathbf{h}$, simultaneously. To do so, wthey formulate the problem as a Bregman projection with prescribed row sum $\left(\forall k \; \mathbf{D}_1^{(k)} \gamma^{(k)} \mathbf{1}_n = \mathbf{h}\right)$, i.e.,

$$\mathbf{h}^\star = \underset{\mathbf{h} \in \Delta_C, \Gamma}{\arg\min} \sum_{k=1}^{K} \lambda_k \mathrm{KL}\left(\gamma^{(k)} \mid \zeta^{(k)}\right)$$
$$\text{s.t. } \forall k \quad \mathbf{D}_1^{(k)} \gamma^{(k)} \mathbf{1}_n = \mathbf{h}.$$

$\mathbf{D}_1^{(k)}$ gives information about class in source domain, while $\gamma^{(k)}$ provides the coupling information about source and target data. Thus, $\mathbf{h}$ is based on classes instead of samples directly. Γ relies on the choice of $\mathbf{h}$ and the algorithm guarantees a better outcome.

This problem admits a closed form solution that we establish in the following result.

The solution of the projection defined above is given by:

$$\forall k, \gamma_k = \mathrm{diag}\left(\frac{\mathbf{D}_2^{(k)} \mathbf{h}}{\zeta^{(k)} \mathbf{1}_n}\right) \zeta^{(k)}, \mathbf{h} = \Pi_{k=1}^{K} \left(\mathbf{D}_1^{(k)} \left(\zeta^{(k)} \mathbf{1}_n\right)\right)^{\lambda_k}$$