Optimal transport for dimension reduction and domain adaptation

2023-11-14

Introduction

Optimal Transport (OT) problems have recently raised interest in several fields such as domain adaptation, in particular because OT theory can be used for computing distances between probability distributions. By exploiting the geometry of the underlying metric space, they can provide meaningful distances even when the supports of the distributions do not overlap. There exist state-of-the-art methods for dimension reduction and domain adaptation using the framework of OT.

Our choice of the method is motivated by the problem of mining user behavior in web (learning) application in form of their cursor movements. Indeed, trajectories of pointer over an application layout can be viewed as probability distributions over this layout. Note that the supports of those distributions can vary with layout content and dimensions (see example of a recorded cursor trajectory below).

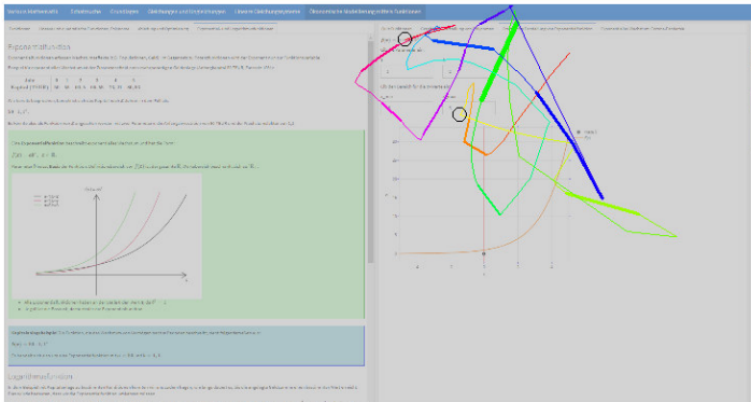

Figure 1: An example of a cursor trajectory within a learning web application. Time is coded by rainbow colors (starting by red). Speed is coded by line thickness (the thicker, the faster). Clicks are marked by circles.

A further challenge in modelling the data lies in the nature of the trajectories which correspond to objects that are intrinsically infinite-dimensional and the sampled traces represent discretized noisy versions of the underlying functional objects.

The ultimate task of the cursor trajectories mining is to forecast user states such as attention, confusion or similar. The cues for such states can be contained in the traces over different layout parts and, whereas the traces depend on the layout itself. For example, interactive parts where users are able to click or drag and draw items naturally contain different traces compared to text or pictures. To cluster or classify users according to their interactions across different layout (a simple example is laptop versus mobile versions in Figure 2) not only dimension reduction but also domain adaptation techniques are essential. Both can be accomplished in the OT framework under the consideration of particular challenges of our data discussed above.

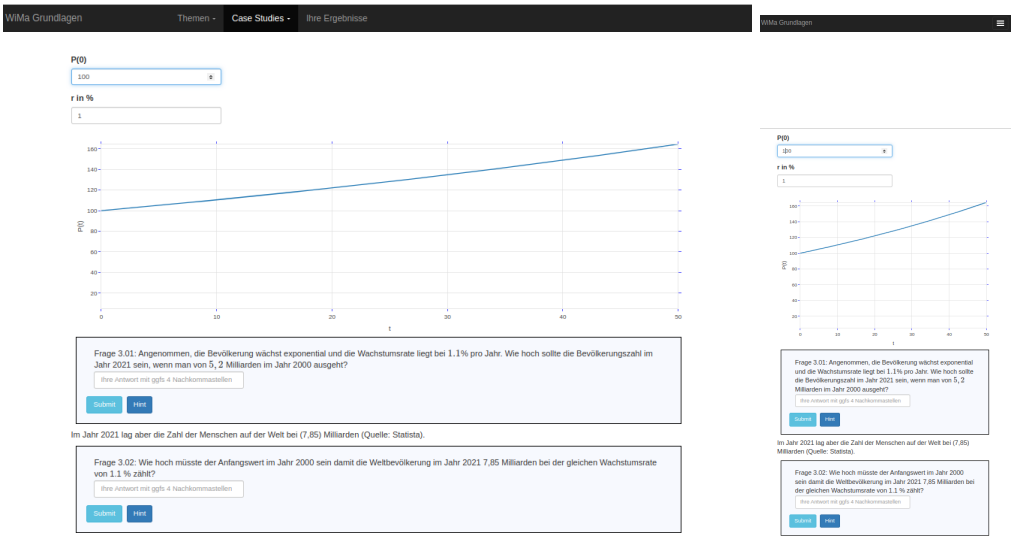

Figure 2: Different layouts of a learning app on a laptop (left) and on a mobile phone (right).

For training a machine learning algorithms to predict emotional states based on the cursor traces, only limited number of labels is usually available due to high labor- and time costs, and some times only possible in experimental setting. In this regard, pooling the data across web layouts (different domains) would increase data availability for training machine learning algorithms and allow for knowledge transfer.

Background on OT

- Definition : Let $\Omega \subseteq \mathbb{R}^d$ be a measurable space, given $\Omega_s, \Omega_t \subseteq \Omega$ with marginal distributions μ_s, μ_t, the Monge Optimal transport problem is finding a mapping $T : \Omega_s \to \Omega_t$ to minimize the overall transportation cost

$$C(T) = \int_{\Omega_t} c(x, T(x)) d\mu_s(x)$$

  where the overall transportation cost can be interpreted as energy required to move a source probability mass $\mu_s(x)$ from x to $T(x)$ with given target probability mass $\mu_t(x)$, and $c : \Omega_s \times \Omega_t \to \mathbb{R}^+$ is a distance function over the metric space Ω. Formally the Monge optimal transport mapping is defined as :

$$T^* = \arg \min_{T:T\#\mu_s=\mu_t} \int_{\Omega_t} c(x, T(x)) d\mu_s(x)$$

  where $T\#$ is a push forward operator defined by $T\#\mu(x) = \mu(T^{-1}(x))$, where μ is a probability measure.

- However nonconvexity of Monge formulation makes optimization difficult, Kantorovich proposed a convex relaxation of Monge Problem, which seeks the best "coupling" of joint distribution for marginals μ_s and μ_t :

$$\pi^* = \arg \min_{\pi \in \Pi(\mu_s, \mu_t)} \int_{\Omega_s \times \Omega_t} c(x_s, x_t) d\pi(x_s, x_t)$$

  where $\Pi(\mu_s, \mu_t)$ is a collection of all possible joint distributions on $\Omega_s \times \Omega_t$ with marginals μ_s and μ_t.

1 OT for dimension reduction

High-dimensional data are in general well approximated or modeled by a low-dimensional manifold. Dimension reduction techniques can assist, in this case, by removing redundant dimensions and stabilizing machine

learning algorithms build upon the resulting low-dimensional embeddings. Manifolds usually arise from data generated in a continuous process. The generated manifold is often embedded in a high-dimensional Euclidean space. (In most cases, the manifold is represented as a discrete data set.) An intuitive example of this is the set of images generated by a continuously changing set of facial expressions. This set of data points can be accurately represented by a low dimensional set of features.

A pioneering work, Wasserstein Isometric Mapping ***Wassmap*** (Hamm, Henscheid, and Kang (2023)), represents images via probability measures in Wasserstein space, then uses pairwise Wasserstein distances between the associated measures to produce a low-dimensional, approximately isometric embedding (a mapping which preserves distances). The authors interpret images (x) as samples of infinite-dimensional data via a discretization operator $\mathcal{H} : X \to \mathbb{R}^n$ on a Banach space X (usually $L_p(\mathbb{R}^m)$):

$$x = \mathcal{H}[\mu] + \eta$$

where η is some noise. The authors assume images correspond to probability measures with finite p-th moment, i.e. $x = \mathcal{H}(\mu)$ where $\mu \in \mathcal{W}_p(\mathbb{R}^m)$, the p-Wasserstein space of probability measures with finite p-th moment $M_p(\mu) := \int_{\mathbb{R}^m} |x|^p \, d\mu(x) < \infty$. Such p-Wasserstein space is equipped with Wasserstein metric arising from Optimal Transport Theory. To define the metric, we consider two measures $\mu, \nu \in \mathbb{W}_p(\mathbb{R}^m)$, and a set of couplings $\Gamma(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathbb{R}^{2m}) : \pi_1 \# \gamma = \mu, \pi_2 \# \gamma = \nu\}$ with $\mathcal{P}(\mathbb{R}^{2m})$ being the set of all probability measures on $\mathbb{R}^{2m}$, and $\pi_{1(2)}$ are the projections on the first (last) m coordinates. That is, $\Gamma(\mu, \nu)$ is the set of joint probability measures with marginals μ and ν. Then, we can give the Wasserstein distance as :

$$W_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}^{2m}} |x - y|^p d\pi(x, y) \right)^{\frac{1}{p}} \tag{1}$$

Hamm, Henscheid, and Kang (2023) stress the advantage of using the Wasserstein distance with an example from object recognition, where one would expect a model to understand that two images of a car are the same object even if the car is translated in the frame of one of the images. These two images can have large Euclidean distance, even though they are semantically identical. The idea of replacing pairwise Euclidean distances with pairwise Wasserstein distances in common manifold learning algorithms has been explored in many settings; for example in Zelesko et al. (2020) to study shape spaces of proteins and in Mathews et al. (2019) to analyze gene expression data. This property is also desirable in our cursor trajectories context, where the "emotional cues" in the traces should be recognized independent of their absolute position.

The algorithm of Hamm, Henscheid, and Kang (2023) builds upon the Wasserstein distance in (1) $W_p(\mu_s, \mu_t)$ is presented in Algorithm 1:

Algorithm 1 Wassmap

Require: probability measures $\{\mu_i\}_{i=1}^N \subset \mathbb{W}_p(\mathbb{R}^m)$; embedding dimension d

1: Compute pairwise Wasserstein distance matrix W with elements $W_{i,j} = W_p^2(\mu_i, \mu_j)$.
2: Compute $B = -\frac{1}{2} HWH$, where $H = \mathcal{I}_N - \frac{1}{N} \mathbf{1}_{N \times N}$.
3: Compute truncated eigendecomposition $B_d = V_d \Lambda_d V_d^\top$.
4: **return** $\{z_i\}_{i=1}^N$ with $z_i = (V_d \Lambda_d^{\frac{1}{2}})_{i,:}$, for $i = 1, \ldots, N$.

In the algorithm above $\mathcal{I}_N$ denotes an $(N \times N)-$ identity matrix and $\mathbf{1}_{N \times N}$ denotes an $(N \times N)-$ matrix of ones.

We explore the potential of the above OT method for mining our cursor trajectories data for the tasks of user clustering and classification using different cost specifications. In particular, the weighted cost is interesting in order to ensure that certain layout elements are transported in the most cases to the same layout elements.

2 Domain adaptation within the OT based dimension reduction

In our context of cursor movements, we have to take into account that traces recorded in different application layouts can have different feature dimensions, such that X_s and X_t belong to different metric spaces. However, the algorithm (Algorithm 3.1) of Hamm, Henscheid, and Kang (2023) relies on a classical formulation of the Wasserstein distance in (1) $W_p(\mu_s, \mu_t)$ which in turn requires the computation of the cost function. The later is not feasible in case of varying feature dimensions.

In this regard, Redko et al. (2020) propose joint optimal transport (COOT) between samples (denoted as π^s) and features (denoted as π^f), which optimizes two transport maps (between samples and features) simultaneously. The authors define the COOT problem for two datasets (the source one indexed with s and the target one with t) $X_s \in \mathbb{R}^{n_s \times d_s}$ and $X_t \in \mathbb{R}^{n_t \times d_t}$ as follows:

$$\min_{\pi^s \in \Pi(\mu_t^s, \mu_s^s) \pi^f \in \Pi(\mu_t^f, \mu_s^f)} \sum_{i,j,k,l} c(X_{s,i,k}, X_{t,j,l}) \pi^s_{i,j} \pi^f_{k,l} = \min_{\pi^s \in \Pi(\mu_t^s, \mu_s^s) \pi^f \in \Pi(\mu_t^f, \mu_s^f)} \langle \mathbf{c}(X_s, X_t) \otimes \pi^s, \pi^f \rangle$$

with $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ being a one-dimensional divergence measure, $\mathbf{c}(\mathbf{x}, \mathbf{x}')$ is a $(d_s \times d_t \times n_s \times n_t)$ tensor of all pairwise divergences. As a result of their optimization, projection $\tilde{X}_t = \pi^{s\top} X_s \pi^f$ is obtained.The corresponding algorithm is presented in Algorithm 2 below:

Algorithm 2 COOT

Require: $\pi^s_{(0)}$, $\pi^f_{(0)}$, $k \leftarrow 0$

1: **while** $k <$max.iter and $err > 0$ **do**
2: $\quad \pi^s_{(k)} \leftarrow OT\left(\pi^s; c(X_s, X_t) \otimes \pi^f_{(k-1)}\right)$//OT problem on the samples
3: $\quad \pi^f_{(k)} \leftarrow OT\left(\pi^f; c(X_s, X_t) \otimes \pi^s_{(k-1)}\right)$//OT problem on the features
4: $\quad err \leftarrow ||\pi^f_{(k-1)} - \pi^f_{(k)}||_F$
5: $\quad k \leftarrow k + 1$
6: **end while**
7: **return** π^s, π^f from the last iteration.

The approach of can be cast as a sort of domain alignment or domain adaptation between different source and target domains without any labels in play. Our goal here is to combine the method of Hamm, Henscheid, and Kang (2023) with COOT of Redko et al. (2020), where also the feature transport is addressed and therefore dimension mismatch is taken into consideration.

3 OT for classification with domain adaptation

A classification problem with domain adaptation assumes existence of two distinct joint probability distributions labeled source domain and unlabeled (with unknown labels or only with some labelled samples) target domain. We use $\mathcal{D}^s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ to denote source domain dataset with n_s labeled samples, where x_i^s is source domain sample and y_i^s is associated label. Let the target domain dataset be unlabeled and denoted as $\mathcal{D}^t = \{(x_i^t)\}_{i=1}^{n_t}$ with n_t unlabeled samples. Domain adaptation is applied, when discrepancies occur in data distribution.

Formally, we want to look for a classifier f and we define expected loss in target domain as $err_T(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}^t} \mathcal{L}(y, f(x))$.

Overall there are two types of domain shifts we focus on: feature shift and label shift. Feature shift occurs when marginal probability distributions of x differ, $\mu_s(x) \neq \mu_t(x)$, while conditional probability distributions remain constant across domains, $\mu_s(y|x) = \mu_t(y|x)$. Label shift occurs when marginal probability distributions of labels differ, i.e. $\mu_s(y) \neq \mu_t(y)$, it causes so called negative transfer. Both shifts are relevant in our

cursor trajectories data. Covariate shifts can occur as a consequence of layout changes. Label shifts occur due to group heterogeneity, instructor heterogeneity or application subject and content.

To solve the problem of simultaneous covariate and label shifts, Le et al. (2022) propose their approach LDROT (***L***abel and ***D***ata Shift ***R***eduction via ***O***ptimal ***T***ransport) based on imitation Learning. Imitation learning includes the following steps:

- Employ a generator (feature generator) to map source and target data into a latent space.
- On the latent space a teacher classifier is trained using labeled source domain.
- Construct student classifier which imitates the prediction of teacher on some source sample, when it predicts an unlabeled sample in target domain that "corresponds" to this source sample (i.e. aiming to reduce label shift).

In particular, the authors in Le et al. (2022) set the target training set as source set to enforce student to not only predict well on source training set, but also generalize to predict well on the unlabeled source set in presence of label shifts, so that this strategy yields a regularizer to mitigate the overfitting problem.

Concretely, the authors consider two data domains A and B with two data distributions μ_A and μ_B respectively, let $h_A : \mathcal{X}_A \to \mathcal{Y}_\Delta$ be a well-qualified classifier that gives accurate prediction for data instances on $\mathcal{X}_A$ and $\mathcal{Y}_\Delta := \left\{\pi \in \mathbb{R}^M : ||\pi||_1 \text{ and } \pi \geq 0\right\}$ and M is number of classes. A classifier h_B learns to predict data from $\mathbb{P}_B$ by imitating h_A on domain A.

The training features are based on data distributions μ_A and μ_B and classifiers h_A, h_B. To enforce h_B to imitate behavior of h_A, Wasserstein distance between the distributions μ_{A,h_A} and μ_{B,h_B} over $\mathcal{X}_A \times \mathcal{Y}_\triangle$ and $\mathcal{X}_B \times \mathcal{Y}_\triangle$, respectively, is considered. The authors propose to add a loss component which minimizes the Wasserstein distance $W(\mu_{s,h_s}, \mu_{t,h_t})$ to mitigate the data shift and label shift. The Wasserstein distance to minimize in Le et al. (2022) is computed with respect to the ground metric:

$$d(z^s, z^t) = \lambda \cdot d_X(x^s, x^t) + d_Y(h^s(x^s), h^t(x^t)),$$

which represents a combined distance between the covariates and labels.

Since in our cursor movements data, we have heterogeneous domains, in the sense, that the source and the target domains may belong to different metric spaces, we need to consider this heterogeneity also in classification scenario. Furthermore, due to the label shift, as Zhao et al. (2019) showed in theorem 4.3, minimizing only the divegence between marginal distributions and expected error of a hypothesis in the source domain will enlarge the expected error of this hypothesis in target domain, therefore alignment of label distributions is suggested. Motivated by the problem, we plan to extend the approach in Le et al. (2022) to heterogeneous domains based on COOT of Redko et al. (2020).

4 References

Hamm, Keaton, Nick Henscheid, and Shujie Kang. 2023. "Wassmap: Wasserstein Isometric Mapping for Image Manifold Learning." https://arxiv.org/abs/2204.06645.

Le, Trung, Dat Do, Tuan Nguyen, Huy Nguyen, Hung Bui, Nhat Ho, and Dinh Phung. 2022. "On Label Shift in Domain Adaptation via Wasserstein Distance." https://arxiv.org/abs/2110.15520.

Mathews, James, Maryam Pouryahya, Caroline Moosmueller, Ioannis Kevrekidis, Joseph Deasy, and Allen Tannenbaum. 2019. "Molecular Phenotyping Using Networks, Diffusion, and Topology: Soft Tissue Sarcoma." *Nature, Sci Rep* 9. https://doi.org/https://doi.org/10.1038/s41598-019-50300-2.

Redko, Ievgen, Titouan Vayer, Rémi Flamary, and Nicolas Courty. 2020. "CO-Optimal Transport." In *Proceedings. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. Available at: Https://Proceedings.neurips.cc/.* Vol. 34.

Zelesko, Nathan, Amit Moscovich, Joe Kileel, and Amit Singer. 2020. "Earthmover-Based Manifold Learning for Analyzing Molecular Conformation Spaces." In *Proceedings. IEEE International Symposium on Biomedical Imaging*, 2020:1715–19. https://doi.org/10.1109/ISBI45749.2020.9098723.

Zhao, Han, Remi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. 2019. "On Learning Invariant Representation for Domain Adaptation." https://arxiv.org/abs/1901.09453.