Structured Data: Learning and Prediction Differentiable Ranks and Sorting using Optimal Transport

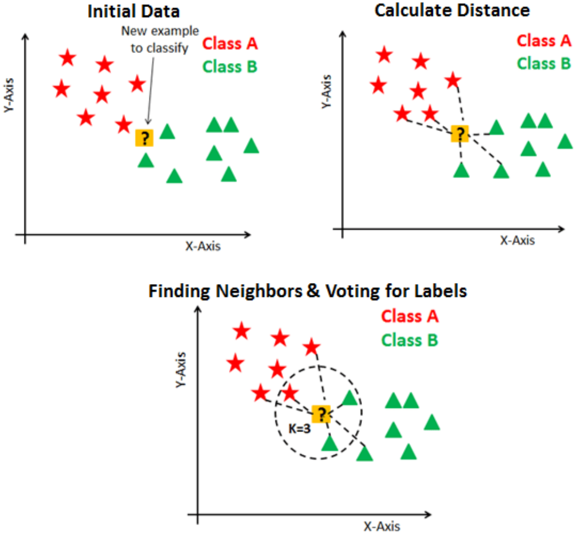
Benjamin Cohen, Vladimir Kondratyev, Jean Pachebat

IP Paris Master Data Science

28/03/2022

Abstract

Sorting is necessary for ML to create algorithms loke k-NN or losses based on the rank.

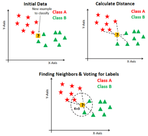


Structured Data: Learning and Prediction
Differentiable Ranks and Sorting using Optimal Transport

└─Abstract

Abstract

Sorting is necessary for ML to create algorithms loke k-NN or losses based on the rank.



It seems to be a difficult task for automatically differentiable pipelines in DL. Sorting gives us two vectors and this application is not differentiable as we are working with integer-valued permutation. In the paper they aim to implement a differentiable proxy of the basic approach.

The article conceive this proxy by thinking of an optimal assignment problem. We sort n values by matching them to a probability measure supported on any increasing family of n target values. Therefore we are considering Optimal Transport (OT) as a relaxation of the basic problem allowing us to extend rank and sort operators using probability measures. The auxiliary measure will be supported on m increasing values with $m \neq n$. Introducing regularization with an entropic penalty and applying Sinkhorn iterations will allow to gain back differentiable operators. The smooth approximation of rank and sort allow to use the 0/1 loss and the quantile regression loss.

Difficult problem for differentiable pipelines

Sorting $\Leftrightarrow$ finding the ranks and the sorted vector, which are not differentiable operations!
The article's goal: implement a differentiable proxy of the basic approach.
Thought as an optimal assignment problem: sort n values by matching them to a probability measure supported on m increasing values on any increasing family of the n target values.
Optimal Transport (OT) $\rightarrow$ relaxation of the basic problem allowing us to extend rank and sort operators using probability measures.
Regularization with entropic penalty then apply Sinkhorn iterations to gain back differentiable operators.
Smooth approximation of rank and sort for classification 0/1 loss and the quantile regression loss.

└─Difficult problem for differentiable pipelines

Difficult problem for differentiable pipelines

Sorting ⇔ finding the ranks and the sorted vector, which are not differentiable operations!
The article's goal: implement a differentiable proxy of the basic approach.
Thought as an optimal assignment problem: sort n values by matching them to a probability measure supported on m increasing values on any increasing family of the n target values.
Optimal Transport (OT) → relaxation of the basic problem allowing us to extend rank and sort operators using probability measures.
Regularization with entropic penalty then apply Sinkhorn iterations to gain back differentiable operators.
Smooth approximation of rank and sort for classification 0/1 loss and the quantile regression loss.

Notations

- $O_n \subset R^n$ - the set of increasing vectors of size n.
- $\Sigma_n \subset R^n_+$ - probability simplex.
- 1_n - n-vector of ones.
- Given $c = (c_1, \ . \ . \ . \ , c_n) \in R^n$, $\bar{c} = (\mathrm{c}_1 + \cdots + \mathrm{c}_i)_i$.
- Given two permutations $\sigma \in \mathrm{S}_n$, $\tau \in \mathrm{S}_m$ and a matrix $A \in \mathrm{R}^{nm}$, we write $A_{\sigma\tau}$ for the $n \times m$ matrix $[A_{\sigma_i \tau_j}]_{ij}$ obtained by permuting the rows and columns of A using σ and τ.
- $\forall x \in R$, δ_x - Dirac measure on x.
- ξ probability measure, then $\forall \xi \in P(R)$, F_ξ - cumulative distribution function (CDF), and Q_{x_i}: quantile function (generalized if x_i is discrete).

Notations

- $O_n \subset R^n$ - the set of increasing vectors of size n.
- $\Sigma_n \subset R^n_+$ - probability simplex.
- 1_n - n-vector of ones.
- Given $c = (c_1, \ldots, c_n) \in R^n$, $\bar{c} = (c_1 + \cdots + c_i)_i$.
- Given two permutations $\sigma \in S_n$, $\tau \in S_m$ and a matrix $A \in R^{nm}$, we write $A_{\sigma\tau}$ for the $n \times m$ matrix $[A_{\sigma_i \tau_j}]_{ij}$ obtained by permuting the rows and columns of A using σ and τ.
- $\forall x \in R$, δ_x - Dirac measure on x.
- ξ probability measure, then $\forall \xi \in P(R)$, F_ξ - cumulative distribution function (CDF), and Q_{x_i}: quantile function (generalized if x_i is discrete).

Sorting

Sorting can be seen a a function S:

$$x = (x_1, ..., x_n) \in R^n \xrightarrow{\text{find } \sigma} x_\sigma = (x_{\sigma_1}, ..., x_{\sigma_n})$$

where the array $x_\sigma = (x_{\sigma_1}, ..., x_{\sigma_n})$ is positioned in increasing order. We obtain two vectors:

- $S(x) := x_\sigma$ - vectors of sorted values
- $R(x) := \sigma^{-1}$ - the rank of each entry of x.

Problems of these functions: S not differentiable everywhere and R piecewise constant (i.e. Jacobian $\frac{\partial R}{\partial x} = 0$ a.e)

Sorting

Sorting can be seen a a function S:

$$x = (x_1, \ldots, x_n) \in R^n \xrightarrow{\text{find } \sigma} x_\sigma = (x_{\sigma_1}, \ldots, x_{\sigma_n})$$

where the array $x_\sigma = (x_{\sigma_1}, \ldots, x_{\sigma_n})$ is positioned in increasing order. We obtain two vectors:

- $S(x) := x_\sigma$ - vectors of sorted values
- $R(x) := \sigma^{-1}$ - the rank of each entry of x.

Problems of these functions: S not differentiable everywhere and R piecewise constant (i.e. Jacobian $\frac{\partial R}{\partial x} = 0$ a.e)

- point 1
- point 2

Idea: Link between Optimal Transport and smoothed operators R,S

Learn σ by solving an Optimal Assignment problem from a measure defined on the support of x to any increasing family y of same length.
Extend OT: target measures of different lengths supports (m $\neq$ n) then use OT $\implies$ convex combinations of ranks and sorted values.

Idea: Link between Optimal Transport and smoothed operators R,S

Learn σ by solving an Optimal Assignment problem from a measure defined on the support of x to any increasing family y of same length.
Extend OT: target measures of different lengths supports ($m \neq n$) then use OT $\implies$ convex combinations of ranks and sorted values.

Problem

Too costly, non- differentiable operators.
Solving this: regularize OT then use Sinkhorn algorithm with complexity O(nml);
n = length(x)
m = size of the target measure that we can choose small
l = Card(Iterations for Sinkhorn algorithm convergence)

Too costly, non- differentiable operators.
Solving this: regularize OT then use Sinkhorn algorithm with complexity O(nml);
n = length(x)
m = size of the target measure that we can choose small
l = Card(Iterations for Sinkhorn algorithm convergence)

Usual Sorting and OT complexity

$(\xi, \nu) \in (P(R))^2$ discrete with supports resp x, y and a, b vectors $s.t.$:

$$\xi = \sum_{i=1}^{n} a_i \delta_{x_i} \quad and \quad \nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

.

Wasserstein distance between $(\xi,\ \nu)$ univariate computed by: comparing quantile functions by inverting CDFs computed with the ordered values of the supports of those measures.

Complexity of $n.log(n)$ far less than $n^3.log(n)$ for OT problems.

2022-04-22

Usual Sorting and OT complexity

$(\xi, \nu) \in (P(R))^2$ discrete with supports resp x, y and a, b vectors $s.t.$:

$$\xi = \sum_{i=1}^{n} a_i \delta_{x_i} \quad and \quad \nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

.

Wasserstein distance between (ξ, ν) univariate computed by: comparing quantile functions by inverting CDFs computed with the ordered values of the supports of those measures.
Complexity of $n.log(n)$ far less than $n^3.log(n)$ for OT problems.

- point 1
- point 2

OT between 1D measures using sorting

We introduce a translation invariant, non-negative ground metric:

$$(x, y) \to h(y - x) \text{ with } h : \mathbb{R} \to \mathbb{R}_+$$

the OT problem between ξ and ν is the linear program:

$$OT_h(\xi, \nu) = \min_{P \in U(a,b)} \langle P, C_{xy} \rangle \tag{1}$$

where $U(a, b) := \{P \in {R_+}^{n \times m} \mid P\mathbf{1}_m = a, \; P^\top \mathbf{1}_m = b\}$.
Write $C_{xy} = [h(y_j - x_i]_{ij}$ When h is supposed convex, we get a closed form solution using quantile functions:

$$OT_h(\xi, \nu) = \int_0^1 h(Q_\nu(u) - Q_\xi(u))\, du \tag{2}$$

OT between 1D measures using sorting

We introduce a translation invariant, non-negative ground metric:

$$(x, y) \to h(y - x) \text{ with } h : \mathbb{R} \to \mathbb{R}_+$$

the OT problem between ξ and ν is the linear program:

$$OT_h(\xi, \nu) = \min_{P \in U(a,b)} \langle P, C_{xy} \rangle \tag{1}$$

where $U(a, b) := \{P \in {R_+}^{n \times m} \mid P\mathbf{1}_m = a,\ P^\top \mathbf{1}_m = b\}$.
Write $C_{xy} = [h(y_j - x_i]_{ij}$ When h is supposed convex, we get a closed form solution using quantile functions:

$$OT_h(\xi, \nu) = \int_0^1 h(Q_\nu(u) - Q_\xi(u))\, du \tag{2}$$

- point 1
- point 2

North-west corner solution of (1)

We find P_* the optimal solution in in $n+m$ operations.
$(\sigma,\,\tau)$ the sorting permutations of $\mathbf{x}$ and $\mathbf{y}$ respectively.

Proposition

The north-west corner solution $\mathrm{N}_{\sigma^{-1},\tau^{-1}}$ is optimal for (1).
$(P_*)_{\sigma,\tau}$ runs from the north-west to the bottom right corner for the allocations to deduce a feasible solution.

When $n=m$; $a=b=\mathbf{1}_n/n)$, $N_{\sigma^{-1},\tau^{-1}}$ is a permutation matrix divided by n and equals to 0 everywhere except for its entries indexed by $(i,\tau(\sigma^{-1}))_i$ equal to $1/n$. It means we assign the i-th smallest entry of x to the i-th smallest entry in y.

North-west corner solution of (1)

We find P_* the optimal solution in in $n+m$ operations.
(σ, τ) the sorting permutations of $\mathbf{x}$ and $\mathbf{y}$ respectively.

Proposition

The north-west corner solution $\mathrm{N}_{\sigma^{-1},\tau^{-1}}$ is optimal for (1).
$(P_*)_{\sigma,\tau}$ runs from the north-west to the bottom right corner for the allocations to deduce a feasible solution.

When $n=m$; $a=b=\mathbf{1}_n/n)$, $N_{\sigma^{-1},\tau^{-1}}$ is a permutation matrix divided by n and equals to 0 everywhere except for its entries indexed by $(i,\tau(\sigma^{-1}))_i$ equal to $1/n$. It means we assign the i-th smallest entry of x to the i-th smallest entry in y.

- point 1
- point 2

Generalizing sorting, CDFs and quantiles using optimal transport

Assume $\mathbf{y}$ is sorted, $y_1 \leq \ldots \leq y_m$.
Then $\tau = Id$ and if $n = m$ then the $i - th$ smallest value of $\mathbf{x}$ is assigned to ${\sigma_i}^{-1} \implies R$ and S are written using P_*.

Proposition

Let $n = m$ and $n = m$ and $\mathbf{a} = \mathbf{b} = \mathbf{1}_n/n$. Then for all strictly convex functions h and $\mathbf{y} \in \mathbb{O}_n$, if P_* is an optimal solution to 1, then:

$$R(\mathbf{x}) = n^2 P_\star \overline{\mathbf{b}} = n P_\star \begin{bmatrix} 1 \\ \vdots \\ n \end{bmatrix} = n F_\xi(\mathbf{x})$$

$$S(\mathbf{x}) = n P_\star^T \mathbf{x} = Q_\xi(\overline{\mathbf{b}}) \in \mathbb{O}_n$$

Generalizing sorting, CDFs and quantiles using optimal transport

Assume $\mathbf{y}$ is sorted, $y_1 \leq \dots \leq y_m$.
Then $\tau = Id$ and if $n = m$ then the $i-th$ smallest value of $\mathbf{x}$ is assigned to ${\sigma_i}^{-1} \implies R$ and S are written using P_*.

Proposition

Let $n = m$ and $n = m$ and $\mathbf{a} = \mathbf{b} = \mathbf{1}_n/n$. Then for all strictly convex functions h and $\mathbf{y} \in \mathbb{O}_n$, if P_* is an optimal solution to 1, then:

$$R(\mathbf{x}) = n^2 P_\star \overline{\mathbf{b}} = nP_\star \begin{bmatrix} 1 \\ \vdots \\ n \end{bmatrix} = nF_\xi(\mathbf{x})$$

$$S(\mathbf{x}) = nP_\star^T \mathbf{x} = Q_\xi(\overline{\mathbf{b}}) \in \mathbb{O}_n$$

- We can understand this expression as n times CDF of x which are the quantiles of ξ at levels $\bar{b}$. The proposition is only valid when $_i$ and ν are uniform of same size support. The paper now focuses on more general cases where $m = size(y \leq n$; $\mathbf{a}$ and $\mathbf{b}$ are no longer uniform.
- point 2

Kantorovich-ranks and sorts: compare discrete measures of several sizes and weights

Idea: Split the weight a_i of x_i to assign it to several y_i ($\Leftrightarrow$ using b_i) $\implies$ the i-th line (or j-th column) of a solution $P_* \in {R_+}^{n*m}$ often has several positive entries.

K-ranking operator $\xrightarrow{computes}$ convex combinations of rank values;

K-sorting operator $\xrightarrow{computes}$ convex combinations of values contained in x directly.

Convex combinations of ranks/values in Euclidean geometry.

Future works $\rightarrow$ alternative geometries (KL, hyperbolic, etc) on ranks/values.

Pointwise quantities depending on the ordering of **a**, **x**, **b**, **y**.

- point 1
- point 2

Kantorovich-ranks and sorts: compare discrete measures of several sizes and weights

Idea: Split the weight a_i of x_i to assign it to several y_j ($\Leftrightarrow$ using b_j) $\implies$ the i-th line (or j-th column) of a solution $P_* \in R_+^{n*m}$ often has several positive entries.
K-ranking operator $\xrightarrow{computes}$ convex combinations of rank values;
K-sorting operator $\xrightarrow{computes}$ convex combinations of values contained in x directly.
Convex combinations of ranks/values in Euclidean geometry.
Future works $\rightarrow$ alternative geometries (KL, hyperbolic, etc) on ranks/values.
Pointwise quantities depending on the ordering of **a**, **x**, **b**, **y**.

K-ranks & K-sorts

$\forall\ (\mathbf{x}, \mathbf{a}, \mathbf{y}, \mathbf{b}) \in \mathbb{R}^n \times \Sigma_n \times \mathbb{O}_m \times \Sigma_m$, let $P_\star \in U(\mathbf{a}, \mathbf{b})$ be an optimal solution for (1) with a given convex function h.
The K-ranks and K-sorts of $\mathbf{x}$ $w.r.t$ $\mathbf{a}$ evaluated using $(\mathbf{b},\ \mathbf{y})$ are respectively:

$$\widetilde{R}(\mathbf{a}, \mathbf{x}; \mathbf{b}, \mathbf{y}) := n\mathbf{a}^{-1} \circ \left(P_\star \overline{\mathbf{b}}\right) \in [0, n]^n$$
$$\widetilde{S}(\mathbf{a}, \mathbf{x}; \mathbf{b}, \mathbf{y}) := \mathbf{b}^{-1} \circ \left(P_\star^T \mathbf{x}\right) \in \mathbb{O}_m$$

$\tilde{R} \xrightarrow{outputs}$ vector of size n containing a continuous rank $\forall\ x_i$ which can be seen as n times an empirical CDF value in [0, 1] view as a convex mixture of the CDF values b_j of the y_j onto which each x_i is transported. $\tilde{S} \xrightarrow{outputs}$ split-quantile operator outputting m increasing barycenters of some of the entries in x.

K-ranks & K-sorts

$\forall\ (\mathbf{x}, \mathbf{a}, \mathbf{y}, \mathbf{b}) \in \mathbb{R}^n \times \Sigma_n \times \mathbb{O}_m \times \Sigma_m$, let $P_\star \in U(\mathbf{a}, \mathbf{b})$ be an optimal solution for (1) with a given convex function h.
The K-ranks and K-sorts of $\mathbf{x}$ $w.r.t$ $\mathbf{a}$ evaluated using ($\mathbf{b}$, $\mathbf{y}$) are respectively:

$$\widetilde{R}(\mathbf{a}, \mathbf{x}; \mathbf{b}, \mathbf{y}) := n\mathbf{a}^{-1} \circ \left(P_\star \overline{\mathbf{b}}\right) \in [0, n]^n$$
$$\widetilde{S}(\mathbf{a}, \mathbf{x}; \mathbf{b}, \mathbf{y}) := \mathbf{b}^{-1} \circ \left(P_\star^T \mathbf{x}\right) \in \mathbb{O}_m$$

$\tilde{R} \xrightarrow{outputs}$ vector of size n containing a continuous rank $\forall\ x_i$ which can be seen as n times an empirical CDF value in [0, 1] view as a convex mixture of the CDF values b_j of the y_j onto which each x_i is transported. $\tilde{S} \xrightarrow{outputs}$ split-quantile operator outputting m increasing barycenters of some of the entries in x.

Computations and Non-differentiability

Small practical applications

Complexity of $O(nm(n+m))$ and non differentiable aspect of those operators.

Worse: $\frac{\partial P_*}{\partial x} = 0$ a.e.
Solution: use regularized OT

$\tilde{R}$, $\tilde{S}$ expressed using P_* not differentiable w.r.t inputs
Solution: use a differentiable alternative to OT using entropic regularization.

Small practical applications
Complexity of $O(nm(n+m))$ and non differentiable aspect of those operators.

Worse: $\frac{\partial P_*}{\partial x} = 0$ a.e.
Solution: use regularized OT

$\tilde{R}$, $\tilde{S}$ expressed using P_* not differentiable w.r.t inputs
Solution: use a differentiable alternative to OT using entropic regularization.

f Small practical applications of the previous operators due to the complexity of $O(nm(n+m))$ to solve an OT problem far most costly than usual sorting and the non differentiable aspect of those operators.
Even worse: Jacobian $\frac{\partial P_*}{\partial x}$ is alike R, null almost everywhere. $\implies$ use regularized OT.

$\tilde{R}$ and $\tilde{S}$ are expressed using the optimal solution P_* to the linear program in (1). But P_* is not differentiable w.r.t inputs, $\mathbf{x}$ nor parameters $\mathbf{b}$, $\mathbf{y}$. Instead, we can use a differentiable alternative to OT using entropic regularization. The optimal regularized transport plan is a dense matrix, that ensures differentiability everywhere w.r.t. both a and x.

Entropic regularization of the OT problem

$$P^{\varepsilon}_{\star} := \underset{P \in U(\mathbf{a},\mathbf{b})}{\text{argmin}} \langle P, C_{\mathbf{xy}} \rangle - \varepsilon H(P), \quad H(P) = -\sum_{i,j} P_{ij} (\log P_{ij} - 1)$$

Sinkhorn Rank and Sort

$$\widetilde{R}_{\varepsilon}(\mathbf{a}, \mathbf{x}; \mathbf{b}, \mathbf{y}) := n\mathbf{a}^{-1} \circ \mathbf{u} \circ K(\mathbf{v} \circ \overline{\mathbf{b}}) \in [0, n]^n$$
$$\widetilde{S}_{\varepsilon}(\mathbf{a}, \mathbf{x}; \mathbf{b}, \mathbf{y}) := \mathbf{b}^{-1} \circ \mathbf{v} \circ K^T(\mathbf{u} \circ \mathbf{x}) \in \mathbb{R}^m$$

Inputs: $\mathbf{a}, \mathbf{b}, \mathbf{x}, \mathbf{y}, \epsilon, h, \eta$
$C_{\mathbf{xy}} \leftarrow [h(y_j - x_i)]_{i,j}$
$K \leftarrow e^{-C_{\mathbf{xy}}/\epsilon}$, $\mathbf{u} = \mathbf{1}_n$;
while $\Delta(\mathbf{v} \circ K^T\mathbf{u}, \mathbf{b}) < \eta$ **do**
    $\mathbf{v} \leftarrow \mathbf{v}/K^T\mathbf{u}, \mathbf{u} \leftarrow \mathbf{a}/K^T\mathbf{v}$
end while
    return $\mathbf{u}$, $\mathbf{v}$, K

Use case: learning with Smoothed Ranks and Sorts

Labels $1, ..., L$, Set of points Ω. $f_\theta : \Omega \to \mathbb{R}^L$. The function selects the class attributed to ω by taking $l^\star = \text{argmax}_l [f_\theta(\omega)]_l$.
To train this model, classical approach:

$$\min_\theta \ \Sigma_i \ \mathbf{1}_L^T logf_\theta(\omega_i) - [f_\theta(\omega_i)]_{l_i}$$

which writes:

$$\mathcal{L}_{0/1}\left(f_\theta(\omega), l\right) = H\left(L - \left[R\left(f_\theta(\omega)\right]_l\right)\right.$$

with: $H(u) = \mathbf{1}_{\{u<0\}}$
Differentiable approximation:

$$\widetilde{\mathcal{L}}_{k,\varepsilon}\left(f_\theta(\omega), l\right) = J_k\left(L - \left[\widetilde{R}_\varepsilon\left(\frac{\mathbf{1}_L}{L}, f_\theta(\omega); \frac{\mathbf{1}_L}{L}, \frac{\overline{\mathbf{1}}_L}{L}, h\right)\right]_l\right)$$