Statistical Distances

References:

- JMLR - A kernel two-sample test ⭐
- Blog - Maximum Mean Discrepancy (MMD) in Machine Learning ⭐
- Wasserstein GAN and the Kantorovich-Rubinstein Duality - Vincent Herrmann ⭐
- 科学空间 - 两个多元正态分布的KL散度、巴氏距离和W距离 ⭐
- 知乎 - EMD(earth mover's distances)距离
- 知乎 - KL散度衡量的是两个概率分布的距离吗？ - 刘斯坦的回答
- 知乎 - 深度理解什么是RKHS(再生希尔伯特空间)
- What is an RKHS?

KL Divergence

Definition

Kullback-Leibler divergence (KL divergence) is a measure of the distance between two probability distributions p and q. It is defined as:

$$D_{\mathrm{KL}}\Big(q(\mathbf{x})\|p(\mathbf{x})\Big) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} = \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})} \log \frac{q(\mathbf{x})}{p(\mathbf{x})}$$

Interpretation

Consider a random variable X with distribution p. By Huffman encoding theorem, the encoding length of a sample $\mathbf{x}$ is given by $\log(1/p(\mathbf{x}))$. So the average encoding length is of X is

$$\mathrm{L1} = E_{\mathbf{x}\sim p(\mathbf{x})} \log \frac{1}{p(\mathbf{x})} = \int p(\mathbf{x}) \log \frac{1}{p(\mathbf{x})} d\mathbf{x}$$

If we mistakenly encode X using the distribution q instead, the encoding length of a sample $\mathbf{x}$ is given by $\log(1/q(\mathbf{x}))$, and the average encoding length is

$$\mathrm{L2} = E_{\mathbf{x}\sim p(\mathbf{x})} \log \frac{1}{q(\mathbf{x})} = \int p(\mathbf{x}) \log \frac{1}{q(\mathbf{x})} d\mathbf{x}$$

The excess encoding length is given by

$$\mathrm{L2} - \mathrm{L1} = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} =: D_{\mathrm{KL}}(p\|q)$$

This is the average encoding loss, defined as **KL divergence**.

Properties

- The KL divergence is always non-negative, because

$$
\begin{aligned}
0 &= \log\left[\int p(\mathbf{x})d\mathbf{x}\right] \\
&= \log\left[\int q(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}d\mathbf{x}\right] \\
&\geq \int q(\mathbf{x})\log\frac{p(\mathbf{x})}{q(\mathbf{x})}d\mathbf{x} \qquad \text{(Jensen's Inequality)}
\end{aligned}
$$

- KL divergence between 1-d Gaussians: Given $p = \mathcal{N}(\mu_1, \sigma_1^2)$, $q = \mathcal{N}(\mu_2, \sigma_2^2)$, we have:

$$D_{\text{KL}}(p\|q) = \ln\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

- KL divergence between multivariate Gaussians: Given $p = \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$, $q = \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$, we have:

$$D_{\text{KL}}(p\|q) = \frac{1}{2}\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \log\det(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1) + \operatorname{Tr}\left(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1\right) - n\right]$$

Wasserstein Distance

Definition

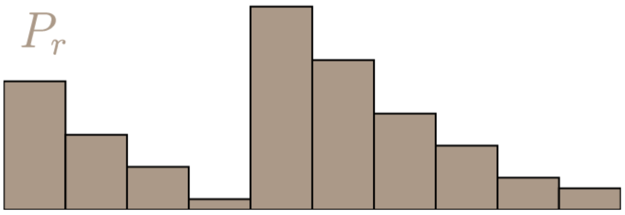
Given two probability measures p and q on a topological space $\mathcal{X}$, the Wasserstein distance between them is defined as

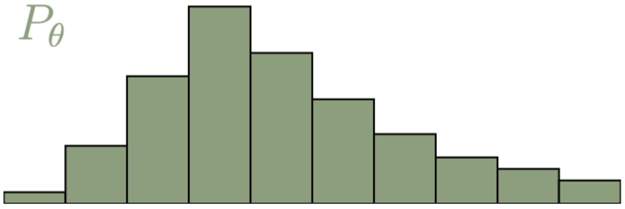
$$\mathcal{C}[p, q] = \inf_{\gamma\in\Pi[p,q]}\iint \gamma(\boldsymbol{x}, \boldsymbol{y})c(\boldsymbol{x}, \boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}$$

- $\Pi[p, q]$ is the set of all joint distributions of p and q on $\mathcal{X}$.
- $c(\boldsymbol{x}, \boldsymbol{y})$ is the cost function that measures the discrepancy between the pair of points $\boldsymbol{x}$ and $\boldsymbol{y}$. It is usually chosen to be the Euclidean distance: $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$.
- inf means infimum (greatest lower bound).

Discrete Case: The Earth Mover's Distance

For discrete probability distributions, the Wasserstein distance is also descriptively called the **earth mover's distance (EMD)**. Consider two one-dimensional discrete probability distributions p and q with l possible values. If we imagine the distributions as different heaps of a certain amount of earth, then the EMD is the minimal total amount of work it takes to transform one heap into the other. Here "work" is defined as the amount of earth in a chunk times the distance it is moved.





We call the transport plan $\gamma(x, y)$, which is the amount of earth moved from one place x to another place y. To be a valid transport plan, it must satisfy the constraints $\sum_x \gamma(x, y) = p(y)$ and $\sum_y \gamma(x, y) = q(x)$. Equivalently, we can call γ a joint probability distribution whose marginals are p and q, respectively.

Calculating the EMD is an optimization problem:

$$\mathrm{EMD}(p,q)=\inf_{\gamma\in\Pi[p,q]}\sum_x\sum_y\|x-y\|\gamma(x,y)=\inf_{\gamma\in\Pi[p,q]}\mathbb{E}_{(x,y)\sim\gamma}\|x-y\|$$

Denote $\mathbf{\Gamma}=\gamma(x,y)$, $\mathbf{D}=\|x-y\|$, with $\mathbf{\Gamma},\mathbf{D}\in\mathbb{R}^{l\times l}$, the problem can be written as

$$\mathrm{EMD}(p,q)=\inf_{\gamma\in\Pi}\langle\mathbf{D},\mathbf{\Gamma}\rangle_{\mathrm{F}}$$

where $\langle\cdot,\cdot\rangle_{\mathrm{F}}$ is the Frobenius inner product (sum of all the element-wise products).

> Γ is usually sparse, i.e. most of its entries are zero. This is because we only need to transport a small amount of earth from one heap to the other.

The problem can be solved using linear programming. For details, see Wasserstein GAN and the Kantorovich-Rubinstein Duality - Vincent Herrmann.

Wasserstein Distance between Gaussians

Choose the distance as the ρ-exponential of the Euclidean distance, we can define the Wasserstein-ρ distance:

$$\begin{aligned}\mathcal{W}_\rho[p,q]=(\mathcal{C}[p,q])^{1/\rho}&=\left(\inf_{\gamma\in\Pi[p,q]}\iint\gamma(\boldsymbol{x},\boldsymbol{y})\|\boldsymbol{x}-\boldsymbol{y}\|^\rho d\boldsymbol{x}d\boldsymbol{y}\right)^{1/\rho}\\&=\left(\inf_{\gamma\in\Pi[p,q]}\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\gamma(\boldsymbol{x},\boldsymbol{y})}\left[\|\boldsymbol{x}-\boldsymbol{y}\|^\rho\right]\right)^{1/\rho}\end{aligned}$$

Consider two Gaussian distributions: $p=\mathcal{N}(\boldsymbol{\mu}_p,\mathbf{\Sigma}_p),\quad q=\mathcal{N}(\boldsymbol{\mu}_q,\mathbf{\Sigma}_q)$. In terms of Wasserstein-ρ distance, there is only analytical solution when $\rho=2$. Denote $\mathcal{W}_2^2[p,q]=(\mathcal{W}_2[p,q])^2$, we have

$$\mathcal{W}_2^2[p,q]=\|\boldsymbol{\mu}_p-\boldsymbol{\mu}_q\|^2+\mathrm{Tr}(\mathbf{\Sigma}_p)+\mathrm{Tr}(\mathbf{\Sigma}_q)-2\mathrm{Tr}((\mathbf{\Sigma}_p\mathbf{\Sigma}_q)^{1/2})$$

Specifically, when $\mathbf{\Sigma}_p$ and $\mathbf{\Sigma}_q$ are commutable, the expression simplifies to

$$\mathcal{W}_2^2[p,q]=\|\boldsymbol{\mu}_p-\boldsymbol{\mu}_q\|^2+\|\mathbf{\Sigma}_p^{1/2}-\mathbf{\Sigma}_q^{1/2}\|_F^2$$

MMD

Problem Formulation

Problem: Let x and y be random variables defined on a topological space $\mathcal{X}$, with respective Borel probability measures p and q. Given observations $X:=\{x_1,\dots,x_m\}$ and $Y:=\{y_1,\dots,y_n\}$, independently and identically distributed from p and q, respectively, can we decide whether $p\neq q$?

> In other words, given the samples of two distributions, p and q, how can we decide whether p and q are different?

To start with, we wish to determine a criterion that, in the population setting, takes
on a unique and distinctive value only when $p=q$.

Definition 2: Let $\mathcal{F}$ be a class of functions $f:\mathcal{X}\to\mathbb{R}$ and let p,q,x,y,X,Y be defined as above. We define the maximum mean discrepancy (MMD) as

$$\mathrm{MMD}[\mathcal{F},p,q]:=\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{x\sim p}[f(x)]-\mathbb{E}_{y\sim q}[f(y)]\right)\tag{1}$$

A biased empirical estimate of the MMD is obtained by replacing the population expectations with empirical expectations computed on the samples X and Y,

$$\mathrm{MMD}_b[\mathcal{F}, X, Y] := \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(y_i) \right) \tag{2}$$

We must therefore identify a function class that is rich enough to uniquely identify whether $p = q$, yet restrictive enough to provide useful finite sample estimates.

A Quick Review of Reproducing Kernel Hilbert Spaces

- **Evaluation functional**: Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$. For a fixed $x \in \mathcal{X}$, the map $\delta_x : \mathcal{H} \to \mathbb{R},\ f \mapsto f(x)$ is called the **Dirac evaluation functional** at x.
- **Reproducing kernel Hilbert space**: A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if δ_x is continuous $\forall x \in \mathcal{X}$.
- **Reproducing kernel**: Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued functions defined on a non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **reproducing kernel** of $\mathcal{H}$ if it satisfies
    - $\forall x \in \mathcal{X},\ k(\cdot, x) \in \mathcal{H}$,
    - $\forall x \in \mathcal{X},\ \forall f \in \mathcal{H},\ f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$ (the reproducing property).

The MMD in Reproducing Kernel Hilbert Spaces

Given a reproducing kernel Hilbert space $\mathcal{H}$ and its reproducing kernel k, we define a feature mapping $\phi : \mathcal{X} \to \mathcal{H}$, where $\phi(x) = k(\cdot, x)$. In particular, we have $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = k(x, y)$.

We next extend the notion of feature map to the embedding of a probability distribution: define an element $\mu_p \in \mathcal{H}$ such that $\mathbb{E}_{x\sim p} f(x) = \langle f, \mu_p \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$, which we call the **mean embedding** of p. Similarly, we can define an element $\mu_q \in \mathcal{H}$ such that $\mathbb{E}_{y\sim q} f(y) = \langle f, \mu_q \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. So the MMD can be expressed as:

$$\begin{aligned}
\mathrm{MMD}^2[\mathcal{F}, p, q] &= \left[\sup_{\|f\|_{\mathcal{H}} \le 1} \Big(\mathbb{E}_{x\sim p} f(x) - \mathbb{E}_{y\sim q} f(y) \Big) \right]^2 \\
&= \left[\sup_{\|f\|_{\mathcal{H}} \le 1} \Big(\langle f, \mu_p \rangle_{\mathcal{H}} - \langle f, \mu_q \rangle_{\mathcal{H}} \Big) \right]^2 \\
&= \left[\sup_{\|f\|_{\mathcal{H}} \le 1} \langle \mu_p - \mu_q, f \rangle_{\mathcal{H}} \right]^2 \\
&= \|\mu_p - \mu_q\|_{\mathcal{H}}^2
\end{aligned}$$

The MMD for Finite Samples

Consider a random variable x with distribution p, and a random variable y with distribution q. Denote x', y' as the independent copies of x, y. The squared population MMD is

$$\begin{aligned}
\mathrm{MMD}^2[\mathcal{F}, p, q] &= \|\mu_p - \mu_q\|_{\mathcal{H}}^2 \\
&= \langle \mu_p - \mu_q, \mu_p - \mu_q \rangle_{\mathcal{H}} \\
&= \langle \mu_p, \mu_p \rangle_{\mathcal{H}} + \langle \mu_q, \mu_q \rangle_{\mathcal{H}} - 2\langle \mu_p, \mu_q \rangle_{\mathcal{H}} \\
&= k(x, x') + k(y, y') - 2k(x, y) \\
&= \mathbb{E}_{x\sim p, x'\sim p}[k(x, x')] + \mathbb{E}_{y\sim q, y'\sim q}[k(y, y')] - 2\mathbb{E}_{x\sim p, y\sim q}[k(x, y)]
\end{aligned}$$

Given two groups of samples, $(x_1, x_2, \ldots, x_m)$ and $(y_1, y_2, \ldots, y_n)$, an unbiased empirical estimate is a sum of two U-statistics and a sample average:

$$\begin{aligned}
\mathrm{MMD}_u^2[\mathcal{F}, X, Y] = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j\neq i}^{m} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i}^{n} k(y_i, y_j) \\
- \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j)
\end{aligned} \tag{4}$$

When $m = n$, a slightly simpler empirical estimate may be used. Let $Z := (z_1, \ldots, z_m)$ be m i.i.d. random variables, where $z := (x, y) \sim p \times q$ (i.e., x and y are independent). An unbiased estimate of MMD^2 is

$$\mathrm{MMD}_u^2[\mathcal{F}, X, Y] = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} h(z_i, z_j)$$

which is a one-sample U-statistic with

$$h(z_i, z_j) := k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)$$

Note that MMD_u^2 may be negative, since it is an unbiased estimator of $\mathrm{MMD}[\mathcal{F}, p, q])^2$. The only terms missing to ensure nonnegativity are $h(z_i, z_i)$, which were removed to remove spurious correlations between observations. Consequently we have the bound

$$\mathrm{MMD}_u^2 + \frac{1}{m(m-1)} \sum_{i=1}^{m} \Big[k(x_i, x_i) + k(y_i, y_i) - 2k(x_i, y_i) \Big] \geq 0$$