

Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 11 - Distances between distributions

Salvatore Ruggieri

Department of Computer Science

University of Pisa, Italy

salvatore.ruggieri@unipi.it

Distances and Metrics

A numerical measurement of how far apart two objects are.

Distances and Metrics

A distance over a set \mathcal{A} is a function $d : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}$ such that:

- $d(x, y) \geq 0$ iff $x = y$ *non-negativity*
- $d(x, y) = 0$ iff $x = y$ *identity of indiscernibles*
- $d(x, y) = d(y, x)$ *symmetry*

Moreover, d is called a metric if in addition:

- $d(x, z) \leq d(x, y) + d(y, z)$ *triangle inequality*

Examples over $\mathcal{A} = \mathbb{R}^n$:

- Manhattan or L_1 distance $d_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n |\mathbf{x}_i - \mathbf{y}_i|$
- Euclidian or L_2 distance $d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n (\mathbf{x}_i - \mathbf{y}_i)^2}$
- Chebyshev or L_∞ distance $d_\infty(\mathbf{x}, \mathbf{y}) = \max_{i=1}^n |\mathbf{x}_i - \mathbf{y}_i|$

We aim at defining **distances and metrics over probability distributions**, i.e., when

$$\mathcal{A} = \{F \mid F : \mathbb{R} \rightarrow [0, 1] \text{ is a CDF}\}$$

Distances over probability distributions

A numerical measurement of **how far apart two probability distributions are.**

- ML/DM models are supposed to be applied on the same distribution as the training set:
 - ▶ How far is the test data distribution from the one of the training data? *[Transfer learning]*
 - ▶ Is the data changing over time, thus my model is inadequate? *[Dataset shift]*
- ML/DM algorithms are supposed to choose the best hypothesis:
 - ▶ What is the split in a DT which best distinguish the distribution of classes?
 - ▶ Is my model separating positive and negatives as much as possible?
 - ▶ Is my clustering separating groups with different distributions?
- Data preprocessing looks at feature distribution:
 - ▶ Are these two features conveying the same information?
 - ▶ Can this feature be predictive to the class feature?
- ... and many other applications in Data Science

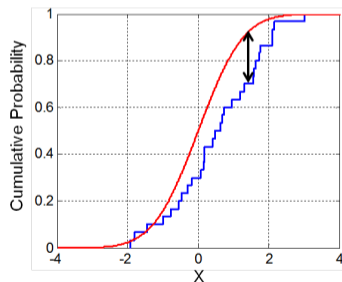
Total variation distance and KS distance

Let X, Y be random variables:

- Total Variation (TV) distance (discrete and continuous case):

$$d_{TV}(X, Y) = \frac{1}{2} \sum_i |p_X(a_i) - p_Y(a_i)| \qquad d_{TV}(X, Y) = \frac{1}{2} \int |f_X(x) - f_Y(x)| dx$$

- ▶ d_{TV} is a metric with $d_{TV}(X, Y) \in [0, 1]$
- Kolmogorov-Smirnov Distance (KS) distance:
$$d_{KS}(X, Y) = \sup_x |F_X(x) - F_Y(x)|$$
 - ▶ d_{KS} is a metric with $d_{KS}(X, Y) \in [0, 1]$
- d_{TV} and d_{KS} have no closed forms in general
- d_{KS} can be estimated from samples of the distributions



See R script

Entropy $H(X)$ of a random variable X

- The **Shannon's information entropy** is the average level of “information” (or “surprise”, “uncertainty”, “unpredictability”) inherent to the variable's possible outcomes

- ▶ Information is inversely proportional to probability

$$\frac{1}{p(a_i)}$$

- Highly likely/unlikely events carry less/more new information

- ▶ Information content $ic()$ of two independent events should sum up

$$\log \frac{1}{p(a_i)}$$

- $ic(p(A \cap B)) = ic(p(A)) + ic(p(B)) = ic(p(A)p(B))$

- $ic(p(\Omega)) = ic(1) = 0$

- $ic(p(A)) \geq 0$

- $H(X) = E[-\log p(X)]$ (*discrete*) $H(X) = E[-\log f(X)]$ (*continuous*)

$$H(X) = - \sum_i p(a_i) \log p(a_i)$$

$$H(X) = - \int_{-\infty}^{\infty} f(x) \log f(x) dx$$

- ▶ For X discrete, $H(X) \geq 0$ since $-\log p(X) = \log 1/p(X) \geq 0$

- zero reached when $p(a_1) = 1$ and $p(a_i) = 0$ for $i \neq 1$

- ▶ For $X \sim \text{Ber}(p)$, $H(X) = -p \log p - (1-p) \log (1-p)$

- for $X \sim \text{Ber}(0.5)$: $H(X) = -2 \cdot 1/2 \log 1/2 = 1$

[binary entropy function]
[unit of entropy is called a **bit**]

Entropy bounds

Corollary of Jensen's inequality [T, Ex. 8.11].

For a concave function g , namely $g''(x) \leq 0$: $g(E[X]) \geq E[g(X)]$

- $\log(x)$ is concave since $\log''(x) = -1/x^2 \leq 0$
- Let X be discrete with finite domain of n elements
 - ▶ By corollary above:

$$H(X) = E\left[\log \frac{1}{p(X)}\right] \leq \log E\left[\frac{1}{p(X)}\right]$$

- ▶ By change of variable:

$$E\left[\frac{1}{p(X)}\right] = \sum_i \frac{p(a_i)}{p(a_i)} = n$$

and then maximum entropy is:

$$H(X) \leq \log n$$

- ▶ E.g., $X \sim \text{Ber}(p)$, maximum entropy (uncertainty) for equiprobable events $p = 1/2$

See R script

Cross entropy

- X, Y discrete random variables with p.m.f. p_X and p_Y :
- Cross entropy of X w.r.t. Y : $H(X; Y) = E_X[-\log p(Y)]$

$$H(X; Y) = - \sum_i p_X(a_i) \log p_Y(a_i)$$

$$\text{with } p_X(a_i) \log p_Y(a_i) = \begin{cases} 0 & \text{if } p_X(a_i) = 0 \\ -\infty & \text{if } p_X(a_i) > 0 \wedge p_Y(a_i) = 0 \end{cases}$$

- $H(X; Y)$ is the “information” or “uncertainty” or “loss” when using Y to encode X
- The closer p_X and p_Y , the lower is $H(X; Y)$
- The lower bound is for $Y = X$, for which $H(X; Y) = H(X)$

Kullback-Leibler divergence

KL divergence

For X, Y discrete random variables with p.m.f. p_X and p_Y :

$$D_{KL}(X \parallel Y) = \sum_i p_X(a_i) \log \frac{p_X(a_i)}{p_Y(a_i)} = H(X; Y) - H(X)$$

- Measure how distribution of Y (model) can reconstruct the distribution of X (data)
 - ▶ Also called: *relative entropy* or *information gain* of X w.r.t. Y

- Properties

- ▶ $D_{KL}(X \parallel Y) \geq 0$
- ▶ $D_{KL}(X \parallel Y) = 0$ iff $F_X = F_Y$
- ▶ $D_{KL}(X \parallel Y) \neq D_{KL}(Y \parallel X)$

[Gibbs' inequality]

[not a distance!]

- For X, Y continuous: $D_{KL}(X \parallel Y) = \int_{-\infty}^{\infty} f_X(x) \log \frac{f_X(x)}{f_Y(x)} dx$

See R script

Joint entropy

- X, Y discrete random variables with p.m.f. p_X and p_Y :
- Joint p.m.f. p_{XY} . Joint entropy of (X, Y) :

$$H((X, Y)) = - \sum_{i,j} p_{XY}(a_i, a_j) \log p_{XY}(a_i, a_j)$$

- If $X \perp\!\!\!\perp Y$, then:

$$\begin{aligned} H((X, Y)) &= - \sum_{i,j} p_X(a_i) p_Y(a_j) (\log p_X(a_i) + \log p_Y(a_j)) = \\ &= - \left(\sum_i p_X(a_i) \right) \left(\sum_j p_Y(a_j) \log p_Y(a_j) \right) - \left(\sum_j p_Y(a_j) \right) \left(\sum_i p_X(a_i) \log p_X(a_i) \right) = H(X) + H(Y) \end{aligned}$$

Mutual information

Mutual information

For X, Y discrete random variables with p.m.f. p_X and p_Y and joint p.m.f. p_{XY} :

$$I(X, Y) = D_{KL}(p_{XY} \parallel p_X p_Y) = \sum_{i,j} p_{XY}(a_i, a_j) \log \frac{p_{XY}(a_i, a_j)}{p_X(a_i)p_Y(a_j)} = H(X) + H(Y) - H((X, Y))$$

- MI measures how dependent two distributions are
 - ▶ Measure how product of marginals can reconstruct the joint distribution

- Properties

- ▶ $I(X, Y) = I(Y, X)$, and $I(X, Y) \geq 0$
- ▶ $I(X, Y) = 0$ iff $X \perp\!\!\!\perp Y$
- ▶ $NMI = \frac{I(X, Y)}{\min\{H(X), H(Y)\}} \in [0, 1]$

[Normalized mutual information]

- For X, Y continuous: $I(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \log \frac{f_{XY}(x, y)}{f_X(x)f_Y(y)} dx dy$

See R script

The data processing inequality

- Let X be unknown, and assume to observe a noisy version Y of it
- Let $Z = f(Y)$ be a data processing to improve the “quality” of Y
- Z does not increase the information about X , i.e.:

[Data processing inequality]

$$I(X, Y) \geq I(X, Z)$$

- If $I(X, Y) = I(X, Z)$ and Z is a summary of Y , we call it a *sufficient statistics*
 - ▶ Let $X \sim \text{Ber}(\theta)$ and $Y = (Y_1, \dots, Y_n) \sim \text{Ber}(\theta)^n$ modelling i.i.d. observations
 - ▶ $Z = \sum_{i=1}^n Y_i \sim \text{Binom}(n, \theta)$ is a sufficient statistics
 - ▶ **Proof (sketch):** use $D_{KL}(p_{XY} \parallel p_X p_Y)$ and:

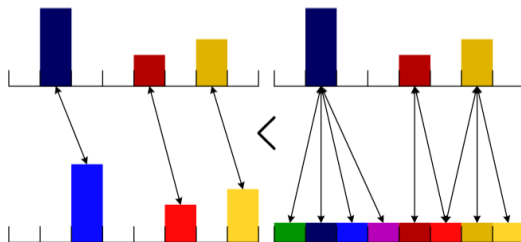
$$p(Y_1 = y_1, \dots, Y_n = y_n) = \prod_i \theta^{y_i} (1 - \theta)^{(1 - y_i)} = \theta^{\sum_i y_i} (1 - \theta)^{n - \sum_i y_i} = p(Z = \sum_i y_i)$$

Earth mover's distance / Wasserstein metric

- The minimum cost to transform one distribution to another
- Cost = amount of mass to move \times distance to move it
- X, Y discrete random variables:

$$EMD(X, Y) = \frac{\sum_{i,j} F_{i,j} \cdot |a_i - a_j|}{\sum_{i,j} F_{i,j}}$$

where F is the flow which minimizes the numerator (total cost) subject to **some constraints**.



Earth mover's distance / Wasserstein metric

- The minimum cost to transform one distribution to another
- Solution of the transportation problem (version from [Ramdas et al. 2015](#)):

$$EMD(X, Y) = \int_0^1 |F_X^{-1}(p) - F_Y^{-1}(p)| dp$$

and for X, Y univariate:

$$EMD(X, Y) = \sum_i |F_X(a_i) - F_Y(a_i)| \quad EMD(X, Y) = \int_{-\infty}^{\infty} |F_X(x) - F_Y(x)| dx$$

- For empirical distributions from **ordered** samples x_1, \dots, x_n and y_1, \dots, y_n :

$$EMD(X, Y) = \frac{1}{n} \sum_i |x_i - y_i|$$

See R script

Reference book chapter for this lesson



Kevin P. Murphy (2022)

Probabilistic Machine Learning: An Introduction

Chapter 6: Information Theory

[online book](#)