École polytechnique
MAP 534: introduction to machine learning
A. Durmus alain.durmus@polytechnique.edu

3 Distance functions and k-nearest-neighbors

3.1 Exercises

In all this section, for $x \in \mathbb{R}^d$, x_i stands for the i-th component of x.

Exercise 3.1. (1) Show that the L^1 and L^∞ distance define for any $x, y \in \mathbb{R}^d$,

$$\mathbf{d}_1(x,y) = \|x-y\|_1 = \sum_{i=1}^d |x_i - y_i| \ , \quad \mathbf{d}_\infty(x,y) = \|x-y\|_\infty = \max_i |x_i - y_i| \ . \tag{12}$$

are distance functions.

(2) Show that L^1, L^2 and L^∞ distance are pairwise equivalent.

Exercise 3.2. A function $N : \mathbb{R}^d \to \mathbb{R}_+$ is referred to as a norm if for any $x, y \in \mathbb{R}^d$, $\alpha \in \mathbb{R}$, 1. $N(x) = 0$ if and only if $x = 0$; 2. $N(x+y) \leq N(x) + N(y)$; 3. $N(\alpha x) = |\alpha| N(x)$. Show that $\mathbf{d}_N(x,y) = N(x-y)$ is a distance function.

Exercise 3.3. We consider that $\mathsf{X} = \mathrm{C}([0,1], \mathbb{R})$ the space of continuous function from $[0,1]$ to $\mathbb{R}$.

(1) Show that

$$(f,g) \in \mathsf{X}^2 \mapsto \sup |f-g| \ , \tag{13}$$

and

$$(f,g) \in \mathsf{X}^2 \mapsto \int_0^1 |f-g| \, \mathrm{d}x \ , \tag{14}$$

are distance functions on X.

(2) Are they equivalent?

Exercise 3.4. (1) For which function $f : \mathbb{R} \to \mathbb{R}$, is $(x,y) \mapsto |f(x) - f(y)|$ a distance function on $\mathbb{R}$?

(2) Show that $(x,y) \mapsto |x^{-1} - y^{-1}|$ is a distance function on $\mathbb{R} \setminus \{0\}$.

Exercise 3.5. Show that the function

$$(x,y) \in \mathbb{S}^d \times \mathbb{S}^d \mapsto \langle x, y \rangle \ , \tag{15}$$

is a discerning similarity function on the d-dimensional sphere $\mathbb{S}^d = \{x \in \mathbb{R}^d \, : \, \|x\| = 1\}$.

3.2 Homework

Exercise 3.6 (Homework 1)**.**

- Implement a function that computes the edit distance using the dynamic programming approach.
- Estimate the random complexity time of your algorithm with respect to the maximal length of a words. To this end, for varying lengths n, you may generate between 10^5 and 10^6 random words of length n and compute the average computation time of your algorithm to compute the distance between pairs of your samples. Finally, you can plot your estimations as a function of n.

Exercise 3.7 (Homework optional).

- Given some data $\{x_i\} \in \mathbb{R}^d$. Implement the underlying $k-d$-tree structure and and the search trees seen in the course. Here we may use a number of maximal children equal to 4 or pass it as a parameter.
- Download the data on the moodle of the course and compare the average execution time of your algorithm with a naive approach to find the closest neighbour. To do so, you will sample uniformly at random, between 10^5 and 10^6 points and compute the average execution time associated with.