

CSE 422 Autumn 26 HW 2

Note: For all problems, you will be expected to explain your reasoning. Providing an expression without reasoning will not receive any partial credit.

Problem 1: Not quite a metric of success...

Recall metric spaces are pairs (X, d) satisfying:

1. $d(x, y) = 0$ if and only if $x = y$.
2. $d(x, y) = d(y, x)$ for all $x, y \in X$.
3. $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$.

The following four examples are NOT metric spaces. For each one, give a counter-example explaining why some property is violated.

- (a) [5 points] (ℓ_p for $p < 1$.) Let $p \in (0, 1)$, let $X = \mathbb{R}^k$, and

$$d(x, y) = \left(\sum_{i=1}^k |x_i - y_i|^p \right)^{1/p}.$$

- (b) [5 points] Let X be the space of all Riemann integrable functions over $[0, 1]$, and let

$$d(f, g) = \int_0^1 |f(x) - g(x)| dx.$$

- (c) [5 points] Let $X = \mathbb{R}^k$, and let $d(x, y) = |||x||_2 - ||y||_2|$.

- (d) [5 points] (Minimum of two metrics.) Let d_1, d_2 be two valid distance functions over a shared space X , and let

$$d(x, y) = \min(d_1(x, y), d_2(x, y)).$$

Problem 2: Jaccard Similarity.

Let X denote the set of all finite subsets of the integers, and for any two $A, B \in X$, define their *Jaccard similarity* to be

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

For the special case where $A, B = \emptyset$, we will let $J(\emptyset, \emptyset) = 1$. Our goal in this problem is to show that $d(A, B) = 1 - J(A, B)$ is a metric on X .

(a) [5 points] First, verify that d satisfies the first two properties of distance functions (i.e. not triangle inequality).

So, the main remaining difficulty is to prove the triangle inequality. Fix sets $A, B, C \in X$. Our goal is to show that $d(A, C) \leq d(A, B) + d(B, C)$.

To do so, there's actually a really cute *probabilistic method* proof. Consider the following random process: choose a *uniformly random* permutation π of the elements in $A \cup B \cup C$, and for any set S , let $h_\pi(S)$ denote the smallest element of S under this permutation:

$$h_\pi(S) = \arg \min_{i \in S} \pi(i).$$

(b) [6 points] Can you derive an expression for $\Pr_\pi[h_\pi(A) \neq h_\pi(B)]$ in terms of the Jaccard similarity?

(c) [9 points] Complete the derivation!

Hint: Let $S_{A,B}$ denote the event that $h_\pi(A) \neq h_\pi(B)$, and similarly define $S_{A,C}$ and $S_{B,C}$. Think about how these sets relate to each other.

Problem 3: Lower bounds for JL.

Recall the Johnson-Lindenstrauss guarantee we showed in class: if $x_1, \dots, x_n \in \mathbb{R}^k$ is any dataset, then if we let $G : \mathbb{R}^k \rightarrow \mathbb{R}^\ell$ be a (suitably scaled) random Gaussian matrix, where $\ell = O(\log n / \varepsilon^2)$, then with high probability, we have that

$$(1 - \varepsilon) \|Gx_i - Gx_j\|_2 \leq \|x_i - x_j\|_2 \leq (1 + \varepsilon) \|Gx_i - Gx_j\|_2, \text{ for all } i, j. \quad (1)$$

In this problem, your goal is to show that many (in fact, it turns out all) aspects of this theorem are essentially necessary.

You may assume without justification that the volume of a (open or closed) radius r ball in ℓ_2 in d dimensions is $c_d r^d$, for some constant c_d that depends only on the dimension.

(a) [7 points] Let $y_1, \dots, y_n \in \mathbb{R}^\ell$ satisfy that, for all $i \neq j$, we have that

$$\|y_i - y_j\|_2 \in [1, 2],$$

that is, all of their pairwise distances are between 1 and 2. Show that $\ell \geq \log_5 n$.

Hint: consider the volumes of balls of various radii centered at the y_i .

(b) [7 points] Use this to show the following claim: there exists a set of points $x_1, \dots, x_n \in \mathbb{R}^n$ so that if $\ell < \log_5 n$, there is some constant $\varepsilon > 0$ so that there does not exist a set of points $y_1, \dots, y_n \in \mathbb{R}^\ell$ satisfying

$$(1 - \varepsilon) \|y_i - y_j\|_2 \leq \|x_i - x_j\|_2 \leq (1 + \varepsilon) \|y_i - y_j\|_2.$$

Note that this shows that the $\log n$ is necessary!

(c) [6 points] Now, we're going to show that the randomness is in some sense, necessary. In particular, show that for any *fixed* matrix $G : \mathbb{R}^{\ell \times k}$ for $\ell < k$, there exists a set $x_1, \dots, x_{\ell+1}$ of $\ell + 1$ points in $\mathbb{R}^k$ so that Equation 1 is violated, for all $\varepsilon < 1$.