Homework 2 Sample Solutions

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

(a) For $k \geq 2$, let $x = 0$, $y = e_1$, and $z = e_1 + e_2$, where e_1, e_2 are standard basis vectors. Then

$$d(x, y) = d(y, z) = 1, \qquad d(x, z) = 2^{1/p} > 2 = d(x, y) + d(y, z),$$

since $0 < p < 1$. Thus the triangle inequality fails. (For $k = 1$, the formula is just $d(x, y) = |x - y|$, which *is* a metric)

(b) Take $f(x) = 0$ for every $x \in [0, 1]$ and

$$g(x) = \begin{cases} 1, & x = 1/2, \\ 0, & x \neq 1/2. \end{cases}$$

Both functions are Riemann integrable. Moreover, changing a bounded function at a single point does not change its integral. Hence

$$d(f, g) = \int_0^1 |f(x) - g(x)|\, dx = 0 \quad \text{but} \quad f \neq g.$$

This violates the first property: distance zero must imply equality.

Remark. For anyone familiar with real analysis, the set of bounded measurable functions usually *is* considered a metric space, however, formally, we need to identify functions f with the set of all functions that agree with f except on a measure zero set. Then one can verify this is a valid metric space under the metric above.

(c) Let $x = e_1$ and $y = -e_1$. These are distinct vectors with the same Euclidean norm, so

$$d(x, y) = \big|\|e_1\|_2 - \| - e_1\|_2\big| = |1 - 1| = 0.$$

Again, distance zero does not imply equality.

(d) First, observe that for any $c_1, c_2 > 0$, the following is a valid distance function on $\mathbb{R}^2$:

$$d_{c_1,c_2}((x_1, y_1), (x_2, y_2)) = c_1|x_1 - y_1| + c_2|x_2 - y_2|\ .$$

This is because any positive scaling of a distance is a valid distance, and a sum of two metrics is also a valid distance.

So consider $d = \min(d_{1,2}, d_{2,1})$, let $a = (0, 0)$, let $b = (1, 0)$, and let $c = (1, 1)$. Then $d(a, c) = 3$, but $d(a, b) = d(b, c) = 1$, and so $d(a, c) > d(a, b) + d(b, c)$, violating triangle inequality.

Problem 2: Jaccard Similarity

(a) If $A \cup B \neq \emptyset$, then $0 \leq |A \cap B| \leq |A \cup B|$, so $d(A, B) \in [0, 1]$. Also,

$$d(A, B) = 0 \iff |A \cap B| = |A \cup B| \iff A \cap B = A \cup B \iff A = B.$$

Here the middle equivalence uses $A \cap B \subseteq A \cup B$ and finiteness. The convention $J(\emptyset, \emptyset) = 1$ gives $d(\emptyset, \emptyset) = 0$ as well. Finally, intersection and union are symmetric in A, B, so

$$\boxed{d(A, B) = 0 \iff A = B, \qquad d(A, B) = d(B, A).}$$

(b) We claim that

$$\Pr_{\pi}[h_\pi(A) \neq h_\pi(B)] = 1 - J(A, B) = d(A, B).$$

Let u be the first element of $A \cup B$ under π. The relative order induced by a uniformly random permutation of $A \cup B \cup C$ is uniform on $A \cup B$, so u is uniform over $A \cup B$. Moreover,

$$h_\pi(A) = h_\pi(B) \iff u \in A \cap B.$$

Indeed, if u belongs to both sets, it is first in both. Otherwise it is first in exactly one set, and their hash values differ. Thus

$$\Pr_{\pi}[h_\pi(A) = h_\pi(B)] = \frac{|A \cap B|}{|A \cup B|} = J(A, B),$$

which is equivalent to the desired claim.

(c) Use the *same* random permutation for all three sets. For every fixed π, if $h_\pi(A) = h_\pi(B)$ and $h_\pi(B) = h_\pi(C)$, then $h_\pi(A) = h_\pi(C)$. Taking the contrapositive,

$$S_{A,C} \subseteq S_{A,B} \cup S_{B,C}.$$

The union bound and part (b) now give

$$\begin{aligned} d(A, C) &= \Pr_{\pi}(S_{A,C}) \\ &\leq \Pr_{\pi}(S_{A,B}) + \Pr_{\pi}(S_{B,C}) \\ &= d(A, B) + d(B, C). \end{aligned}$$

Together with part (a), this proves that $d = 1 - J$ is a metric on all finite subsets of the integers, including the empty set.

Problem 3: Lower bounds for JL

(a) Consider the open balls $B_i = B(y_i, 1/2) \subseteq \mathbb{R}^\ell$. They are pairwise disjoint: if $z \in B_i \cap B_j$ for $i \neq j$, then the Euclidean triangle inequality would give

$$\|y_i - y_j\|_2 \leq \|y_i - z\|_2 + \|z - y_j\|_2 < 1,$$

contrary to the assumption. Every B_i is contained in $B(y_1, 5/2)$, because

$$\|z - y_1\|_2 \leq \|z - y_i\|_2 + \|y_i - y_1\|_2 < \tfrac{1}{2} + 2 = \tfrac{5}{2}.$$

Therefore, we have a collection of n disjoint balls of radius $1/2$, all fitting within a ball of radius $5/2$. By considering their volume, we obtain:

$$nc_\ell \left(\frac{1}{2}\right)^\ell \leq c_\ell \left(\frac{5}{2}\right)^\ell, \qquad \text{so} \qquad \boxed{n \leq 5^\ell \quad \Longrightarrow \quad \ell \geq \log_5 n.}$$

(b) Choose $x_i = e_i/\sqrt{2} \in \mathbb{R}^n$ for $i = 1, \ldots, n$. Then $\|x_i - x_j\|_2 = 1$ whenever $i \neq j$. Take the constant $\varepsilon = 1/3$. If points $y_1, \ldots, y_n \in \mathbb{R}^\ell$ satisfied the stated inequalities, they would obey

$$\frac{1}{1+\varepsilon} \leq \|y_i - y_j\|_2 \leq \frac{1}{1-\varepsilon}, \qquad \text{i.e.} \qquad \frac{3}{4} \leq \|y_i - y_j\|_2 \leq \frac{3}{2}.$$

Rescale by setting $z_i = \frac{4}{3} y_i$. Their pairwise distances lie in $[1, 2]$, so part (a) implies $\ell \geq \log_5 n$. Therefore

$$\boxed{\ell < \log_5 n \quad \Longrightarrow \quad \text{no such embedding exists for } \varepsilon = 1/3.}$$

This holds even for arbitrary embeddings, without assuming linearity or randomness, and proves that the $\log n$ dependence is necessary.

(c) In fact, you can do this even with two points. Fix $G \in \mathbb{R}^{\ell \times k}$ with $1 \leq \ell < k$. By rank-nullity,

$$\dim \ker G = k - \operatorname{rank}(G) \geq k - \ell > 0.$$

Choose a nonzero $v \in \ker G$, and consider a dataset containing v and $u = -v$. Then $\|v - u\|_2 = 2\|v\|_2 > 0$, but $\|G(v - u)\|_2 = 0$. Therefore there is no $\varepsilon > 0$ for which the JL condition can hold with G, for these two vectors.