Practice Questions: Set Distances

CS768 Learning With Graphs Prof. Abir De

Date: February 9, 2026

1) Jaccard distance: identities and metric properties

Setup. Let U be a finite universe and let $\mathcal{S} := \{A \subseteq U : A \neq \varnothing\}$. For $A, B \in \mathcal{S}$ define the **Jaccard distance**

$$d_J(A,B) \ := \ 1 - \frac{|A \cap B|}{|A \cup B|}.$$

(a) Prove the identity $d_J(A,B) = \dfrac{|A \triangle B|}{|A \cup B|}$**.**

(b) Prove identity: $d_J(A,B) = 0 \iff A = B$**.**

(c) Prove triangle inequality: for all $A, B, C \in \mathcal{S}$**,**

$$d_J(A,C) \ \leq \ d_J(A,B) + d_J(B,C).$$

2) Expected intersection and union under an independent multi-label model

Setup. Let $U = \{1, \ldots, m\}$. Conditioned on x, assume the label-set $Y \subseteq U$ has independent membership:

$$\mathbf{1}_{\{i \in Y\}} \mid x \sim \text{Bernoulli}(p_i), \qquad \text{independent across } i.$$

For a threshold $t \in [0,1]$, define the deterministic prediction

$$\hat{Y}_t := \{i \in U : \ p_i \geq t\}.$$

(a) Prove $\mathbb{E}[|Y \cap \hat{Y}_t| \mid x] = \sum_{i \in \hat{Y}_t} p_i$**.**

(b) Prove $\mathbb{E}[|Y \cup \hat{Y}_t| \mid x] = |\hat{Y}_t| + \sum_{i \notin \hat{Y}_t} p_i$**.**

(c) Provide a concrete example showing $\mathbb{E}\big[\frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|}\big] \neq \frac{\mathbb{E}[|Y \cap \hat{Y}|]}{\mathbb{E}[|Y \cup \hat{Y}|]}$**.**

(Hint: Take $m = 2$, $p_1 = p_2 = \frac{1}{2}$, and choose $\hat{Y} = \{1\}$ (e.g., any $t \in (\frac{1}{2}, 1]$ would give $\hat{Y} = \varnothing$, but here we explicitly fix $\hat{Y}$ as deterministic). Enumerate $Y \in \{\varnothing, \{1\}, \{2\}, \{1,2\}\}$, each with probability 1/4.)

3) Symmetric difference loss: Bayes-optimal set predictor

Setup. Let $U = \{1, \ldots, m\}$. For a deterministic prediction $\hat{Y} \subseteq U$ and random $Y \subseteq U$, define

$$\ell_{\triangle}(\hat{Y}, Y) := |\hat{Y} \triangle Y|.$$

Let $p_i := \Pr(i \in Y \mid x)$.

(a) Prove:

$$\mathbb{E}[\ell_{\triangle}(\hat{Y}, Y) \mid x] = \sum_{i \in \hat{Y}} (1 - p_i) + \sum_{i \notin \hat{Y}} p_i.$$

(b) Deduce and prove the Bayes-optimal predictor:

$$\hat{Y}^{\star}(x) = \{i : \ p_i \geq \tfrac{1}{2}\}.$$

4) MinHash: exact collision probability and concentration

Setup. Let U be finite and $A, B \subseteq U$ be nonempty. Let π be a uniformly random permutation of U. Define MinHash

$$h_\pi(A) := \arg\min_{a \in A} \pi(a),$$

i.e. the element of A with minimum rank under π (ties impossible under a permutation).

(a) Prove: $\Pr[h_\pi(A) = h_\pi(B)] = \dfrac{|A \cap B|}{|A \cup B|}$.

(b) Unbiased estimator and variance. Let $\pi_1, \ldots, \pi_k$ be i.i.d. uniform permutations and define

$$\widehat{J} := \frac{1}{k} \sum_{r=1}^{k} \mathbf{1}\{h_{\pi_r}(A) = h_{\pi_r}(B)\}.$$

Let $J := \frac{|A \cap B|}{|A \cup B|}$. By (a), each indicator is Bernoulli(J) and i.i.d. Show that $\hat{J}$ is an unbiased estimator of J and find its variance.

5) W_1 on $\mathbb{R}$: monotone coupling theorem and application

Setup. Let $A = \{a_1, \ldots, a_n\} \subset \mathbb{R}$ and $B = \{b_1, \ldots, b_n\} \subset \mathbb{R}$, and define empirical measures

$$\mu_A = \frac{1}{n} \sum_{i=1}^{n} \delta_{a_i}, \qquad \mu_B = \frac{1}{n} \sum_{i=1}^{n} \delta_{b_i}.$$

Define the 1-Wasserstein distance (with ground cost $c(x, y) = |x - y|$) by

$$W_1(\mu_A, \mu_B) \ := \ \min_{\pi \in \Pi(\mu_A, \mu_B)} \int_{\mathbb{R} \times \mathbb{R}} |x - y| \, d\pi(x, y),$$

where $\Pi(\mu_A, \mu_B)$ denotes the set of couplings of μ_A and μ_B, i.e.

$$\Pi(\mu_A, \mu_B) := \{\pi \text{ probability measure on } \mathbb{R} \times \mathbb{R} : (\mathrm{proj}_1)_\# \pi = \mu_A, \ (\mathrm{proj}_2)_\# \pi = \mu_B\}.$$

(Equivalently, one may restrict to discrete couplings $\pi = \frac{1}{n} \sum_{i,j} p_{ij} \, \delta_{(a_i, b_j)}$ with $p_{ij} \geq 0$, $\sum_j p_{ij} = 1$, and $\sum_i p_{ij} = 1$.)

(a) Prove (monotone coupling for empirical measures on $\mathbb{R}$):

$$W_1(\mu_A, \mu_B) \ = \ \frac{1}{n} \sum_{i=1}^{n} |a_{(i)} - b_{(i)}|,$$

where $a_{(1)} \leq \cdots \leq a_{(n)}$ and $b_{(1)} \leq \cdots \leq b_{(n)}$ are the sorted sequences.

(b) Compute W_1 for $A = \{0, 2\}$ and $B = \{1, 3\}$.