DM Assignment1 Report

520030910342 Jiyu Liu

1. Concept Questions

1.1. Question 1

Please prove or disprove that the following distances are metrics. a. Jaccard distance b. cosine distance c. edit distance d. hamming distance:

a. Jaccard distance:

1.

$\because |x \cup y| \geq |x \cap y|$

$\therefore d(x, y) = 1 - \frac{|x \cap y|}{|x \cup y|} \geq 0.$

2.

First, assume that d(x, y) = 0,

$\because d(x, y) = 1 - \frac{|x \cap y|}{|x \cup y|} = 0$

$\therefore |x \cap y| = |x \cup y|$

$\therefore x = y$

Then, assume that x = y,

$\because x = y$

$\therefore |x \cap y| = |x \cup y|$

Q.E.D

3.

$d(x, y) = 1 - \frac{|x \cap y|}{|x \cup y|} = 1 - \frac{|y \cap x|}{|y \cup x|} = d(y, x)$

4.

First, we are going to prove the lemma below:

For any set X and its subsets A, B, C, i.e. $A, B, C \subseteq X$, it holds that

$|A \cap C| \cdot |B \cup C| + |A \cup C| \cdot |B \cap C| \leq |C| \cdot (|A| + |B|)$

Proof:

First, we have $|A \cap C| \cdot |B \cup C| = |A \cap C| \cdot (|B| + |C| - |B \cap C|) = |A \cap C| \cdot (|B| - |B \cap C|) + |A \cap C| \cdot |C| \leq |C| \cdot (|B| - |B \cap C| + |A \cap C|)$.

Similarly, we have $|A \cup C| \cdot |B \cap C| \leq |C| \cdot (|A| - |A \cap C| + |B \cap C|)$.

Thus, $|A \cap C| \cdot |B \cup C| + |A \cup C| \cdot |B \cap C| \leq |C| \cdot (|B| - |B \cap C| + |A \cap C|) + |C| \cdot (|A| - |A \cap C| + |B \cap C|) = |C| \cdot (|A| + |B|)$.

Then, we are going to prove the triangle inequality of Jaccarf distance:

$d(x, z) + d(z, y) = 1 - \frac{|x \cap z|}{|x \cup z|} + 1 - \frac{|z \cap y|}{|z \cup y|} = 2 - \frac{|x \cap z||z \cup y| + |z \cap y||x \cup z|}{|x \cup z||z \cup y|} \geq 2 - \frac{|z| \cdot (|x| + |y|)}{|x \cup z||z \cup y|} \geq 2 - \frac{|z| \cdot (|x| + |y|)}{|(x \cup z) \cap (y \cup z)| \cdot |x \cup z \cup y \cup z|} \geq 2 - \frac{|z|}{|(x \cap y) \cup z)|} \cdot \frac{|x| + |y|}{|x \cup y \cup z|} \geq 2 - \frac{|x| + |y|}{|x \cup y|} = 1 - \frac{|x \cap y|}{|x \cup y|} = d(x, y)$

Therefore, Jaccard distance is a metric.

b.cosine distance:

Assuming that $\mathbf{x} = (1, 2)$ and $\mathbf{y} = (2, 4)$

Then, $d(x, y) = 1 - s(x, y) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = 0$

However, obviously $x \neq y$.

Therefore, cosine distance is not a metric.

c.edit distance:

1.

Because that every edit operation has positive cost, so the edit distance $d(x, y)$ is always larger than or equal to 0.

2.

According to the definition, $d(x, y) = 0$ if and only if $x = y$.

3.

Because the cost of a certain operation and its inverse are the same, we can conclude that $d(x, y) = d(y, x)$.

4.

According to the definition below,

$$d(x,y) = d_{ij} =$$

$$\begin{cases} d_{i-1,j-1} & \text{for } a_i = b_j \\ \min \begin{cases} d_{i-1,j} + w_{\text{del}}\,(a_i) \\ d_{i,j-1} + w_{\text{ins}}\ (b_j) \\ d_{i-1,j-1} + w_{\text{sub}}\ (a_i, b_j) \end{cases} & \text{for } a_i \neq b_j \end{cases}$$

We can indicate that d(x,y) is the minimal distance between x and y, which takes the least number of edit, so $d(x,y) \leq d(x,z) + d(z,y)$ for any z.
Therefore, edit distance is a metric.

d.hamming distance:

1.
Because every different position of the two strings has positive cost, so d(x,y) is always larger than or equal to 0.
2.
If d(x,y) = 0, then every position of x and y is the same, thus x = y. If x = y, apparently d(x,y) = 0.
3.
Obviously, $d(x,y) = d(y,x)$.
4.
There are five types of position in the distance calculation of x, y and z:
1. x and y are the same, but z is not.
2. x and z are the same, but y is not.
3. y and z are the same, but x is not.
4. x, y, z are all the same.
5. x, y, z are different from each other.

we denote the distance in this position as d_p.
In case 1: $d_p(x,y) = 0 < d_p(x,z) + d_p(y,z) = 2$.
In case 2 and 3: $d_p(x,y) = 1 = d_p(x,z) + d_p(y,z)$.
In case 4: $d_p(x,y) = 0 = d_p(x,z) + d_p(y,z)$.
In case 5: $d_p(x,y) = 1 < d_p(x,z) + d_p(y,z) = 2$.
Therefore, we have $d_p(x,y) \leq d_p(x,z) + d_p(y,z)$ in every position.
So the triangle inequility holds.

1.2. Question 2

Prove the average distance between a pair of points on a line of length L is L/3.

Denote the probability density function of uniformly distributed points as f, then we have
$f(x) = \begin{cases} \frac{1}{L}, x \in [0, L] \\ 0, otherwise \end{cases}$
Randomly pick two points independently, denoting them by X_1 and X_2.
Then, we denote between them by $Y = |X_1 - X_2|$.
Therefore, we have
$E(Y) = E(|X_1 - X_2|)$
$= \int_0^L \int_0^L |x_1 - x_2| f(x_1) f(x_2) \mathrm{d}x_1 \mathrm{d}x_2$
$= \frac{1}{L^2}(\int_0^L \int_0^{x_1} (x_1 - x_2)\mathrm{d}x_2 \mathrm{d}x_1 + \int_0^L \int_{x_1}^L (x_2 - x_1)\mathrm{d}x_2 \mathrm{d}x_1)$
$= \frac{1}{L^2}(\int_0^L (x_1^2 - \frac{x_1^2}{2})\mathrm{d}x_1 + \int_0^L (\frac{L^2}{2} - \frac{x_1^2}{2} + x_1^2 - L \cdot x_1)\mathrm{d}x_1)$
$= \frac{L}{3}$

1.3. Question 3

Let $A = U\Sigma V^T$ and $B = USV^T$ where $S =$ diagonal $r \times r$ matrix with $s_i = \sigma_i (i = 1...k)$, and $s_i = 0$ otherwise. Please prove B is a best k-rank approximation to A in terms of Frobenius norm error.

Denote a random k-rank approximation to A as A_k.
To prove that B is the best approximation, we are going the prove the lemma below:
If $A, B \in R^{m \times n}$ and $rank(B) = k$, then $\sigma_{k+i}(A) \leq \sigma_i(A - B)$ for all i.

proof:
$\sigma_i(A - B) = \sigma_i(A - B) + \sigma_1(B - B_k)$
$= \sigma_1((A - B) - (A - B)_{i-1} + \sigma_1(B - B_k))$
$\geq \sigma_1((A - B) - (A - B)_{i-1} + B - B_k)$

$= \sigma_1(A - (A-B)_{i-1} - B_k)$
$\geq \sigma_1(A - A_{k+i-1})$
$= \sigma_{k+i}(A)$

Then, we are going to prove that B is a best k-rank approximation to A in terms of Frobenius norm error.
$\because |A - B| = diag(0, 0, 0, ..., \sigma_{k+1}, ..., \sigma_n)$
$\therefore \|A - B\|_F = \| \sum_{i=1}^{n} \sigma_i u_i v_i - \sum_{i=1}^{k} \sigma_i u_i v_i \|_F$
$= \| \sum_{i=k+1}^{n} \sigma_i u_i v_i \|_F$
$= \sqrt{\sum_{i=k+1}^{n} \sigma_i^2}$
$\therefore \|A - B\|_F^2 = \sum_{i=k+1}^{n} \sigma_i(A)^2 \leq \sum_{i=1}^{n-k} \sigma_i(A - A_k)^2 \leq \sum_{i=1}^{n} \sigma_i(A - A_k)^2 = \|A - A_k\|_F^2$

Q.E.D

1.4. Question 4

Suppose we have a universal set U of n elements, and we choose two subsets S and T at random, each with m of the n elements. What is the expected value of the Jaccard similarity of S and T ? A sum expression of the expectation is acceptable if you can't simplify it.

$P(|S \cap T| = k) = \frac{\binom{m}{k}\binom{n-m}{m-k}}{\binom{n}{m}}$
$E(d(S,T)) = E(1 - \frac{|S \cap T|}{|S \cup T|}) = 1 - E(\frac{S \cap T}{S \cup T}) = 1 - E(\frac{|S \cap T|}{|S| + |T| - |S \cap T|}) = 1 - \sum_{k=0}^{m} P(|S \cap T| = k) \cdot \frac{k}{2m-k} = 1 - \sum_{k=0}^{m} \frac{\binom{m}{k}\binom{n-m}{m-k}}{\binom{n}{m}} \cdot \frac{k}{2m-k}$