AI3602 Data Mining: Homework 1

Xiangyuan Xue (521030910387)

1. (1) Jaccard distance is defined as the complement of Jaccard similarity, namely

$$d_{\text{Jaccard}}(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Notice that $|A \cap B| \leq |A \cup B|$, which implies that $d_{\text{Jaccard}}(A,B) \geq 0$. According to the reflexivity and symmetry of set intersection and union, it is obvious that $d_{\text{Jaccard}}(A,A) = 0$ and $d_{\text{Jaccard}}(A,B) = d_{\text{Jaccard}}(B,A)$. Hence, the non-negativity, identity of indiscernibles and symmetry of Jaccard distance are verified.

Before proving the triangle inequality, we introduce the following lemma and corollary.

Lemma. For any sets A, B, C, it holds that

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

Proof. Notice that

$$\begin{aligned} |A \cap C| \cdot |B \cup C| &= |A \cap C| \cdot (|B| + |C| - |B \cap C|) \\ &= |C| \cdot |A \cap C| + |A \cap C| \cdot (|B| - |B \cap C|) \\ &\leq |C| \cdot |A \cap C| + |C| \cdot (|B| - |B \cap C|) \\ &\leq |C| \cdot (|B| + |A \cap C| - |B \cap C|) \end{aligned}$$

By swapping A and B, we can obtain that

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

Adding these two inequalities yields the desired inequality that

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

Corollary. For any sets A, B, it holds that

$$|A \cap B| \cdot |A \cup B| \leq |A| \cdot |B|$$

Proof. Apply the lemma by substituting A, B with S and C with T, which yields

$$|S \cap T| \cdot |S \cup T| \leq |S| \cdot |T|$$

This is exactly the desired inequality, which proves the corollary.

Now we are ready to prove the triangle inequality. Notice that

$$\begin{aligned}
\frac{|A\cap C|}{|A\cup C|}+\frac{|B\cap C|}{|B\cup C|} &= \frac{|A\cap C|\cdot|B\cup C|+|A\cup C|\cdot|B\cap C|}{|A\cup C|\cdot|B\cup C|} \\
&\le \frac{|C|\cdot(|A|+|B|)}{|A\cup C|\cdot|B\cup C|} \\
&\le \frac{|C|\cdot(|A|+|B|)}{|(A\cup C)\cap(B\cup C)|\cdot|(A\cup C)\cup(B\cup C)|} \\
&= \frac{|C|}{|(A\cap B)\cup(A\cap C)\cup(B\cap C)\cup C|}\cdot\frac{|A|+|B|}{|A\cup B\cup C|} \\
&\le \frac{|A|+|B|}{|A\cup B|} \\
&= 1+\frac{|A\cap B|}{|A\cup B|}
\end{aligned}$$

Rearrange the terms and we can obtain that

$$1-\frac{|A\cap B|}{|A\cup B|}\le\left(1-\frac{|A\cap C|}{|A\cup C|}\right)+\left(1-\frac{|B\cap C|}{|B\cup C|}\right)$$

which indicates that $d_{\text{Jaccard}}(A,B)\le d_{\text{Jaccard}}(A,C)+d_{\text{Jaccard}}(C,B)$. Therefore, the triangle inequality of Jaccard distance is verified. Since all the properties are satisfied, we claim that Jaccard distance is a metric.

(2) Cosine distance is defined as the complement of cosine similarity, namely

$$d_{\text{Cosine}}(\boldsymbol{x},\boldsymbol{y})=1-\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|}=1-\frac{\sum_{i=1}^n x_iy_i}{\sqrt{\sum_{i=1}^n x_i^2}\cdot\sqrt{\sum_{i=1}^n y_i^2}}$$

Assume that $\boldsymbol{x}=(1,0)$, $\boldsymbol{y}=(0,1)$ and $\boldsymbol{z}=(1,1)$. By definition, we have

$$\begin{aligned}
d_{\text{Cosine}}(\boldsymbol{x},\boldsymbol{y})&=1\\
d_{\text{Cosine}}(\boldsymbol{x},\boldsymbol{z})&=1-\frac{\sqrt{2}}{2}\\
d_{\text{Cosine}}(\boldsymbol{z},\boldsymbol{y})&=1-\frac{\sqrt{2}}{2}
\end{aligned}$$

which implies that

$$d_{\text{Cosine}}(\boldsymbol{x},\boldsymbol{y})=1>2-\sqrt{2}=d_{\text{Cosine}}(\boldsymbol{x},\boldsymbol{z})+d_{\text{Cosine}}(\boldsymbol{z},\boldsymbol{y})$$

Therefore, cosine distance does not satisfy the triangle inequality, which indicates that cosine distance is not a metric.

(3) Edit distance is defined as the minimum number of operations required to transform one string into another. We denote the edit distance between S and T by $d_{\text{Edit}}(S,T)$. It is obvious that $d_{\text{Edit}}(S,T)\ge 0$ since the number of operations is non-negative. In addition, $d_{\text{Edit}}(S,S)=0$ since no operation is required to transform S into itself. Hence, the non-negativity and identity of indiscernibles of edit distance are verified.

Notice that the transformation from S to T is invertible with insertion replaced by deletion and vice versa, so the transformation from T to S requires the exactly same number of operations as the transformation from S to T, namely $d_{\text{Edit}}(S, T) = d_{\text{Edit}}(T, S)$. Hence, the symmetry of edit distance is verified.

Now we prove the triangle inequality by contradiction. Assume that there exist three strings S, T and R violating the triangle inequality, namely

$$d_{\text{Edit}}(S, T) > d_{\text{Edit}}(S, R) + d_{\text{Edit}}(R, T)$$

This means we can transform S into T by first transforming S into R and then transforming R into T with less operations than transforming S directly into T, which is contradictory to the definition of edit distance. Hence, it holds that $d_{\text{Edit}}(S, T) \leq d_{\text{Edit}}(S, R) + d_{\text{Edit}}(R, T)$, which verifies the triangle inequality of edit distance. Since all the properties are satisfied, we claim that edit distance is a metric.

(4) Hamming distance between two strings of equal length is defined as the number of positions at which the corresponding symbols are different, namely

$$d_{\text{Hamming}}(S, T) = \sum_{i=1}^{n} \mathbb{1}(S_i \neq T_i)$$

Since the indicator function $\mathbb{1}(\cdot) \geq 0$, it holds that $d_{\text{Hamming}}(S, T) \geq 0$. In addition, $d_{\text{Hamming}}(S, S) = 0$ since no position is different, which verifies the non-negativity and identity of indiscernibles of Hamming distance.

By the symmetry of inequality relation, it holds that

$$d_{\text{Hamming}}(S, T) = \sum_{i=1}^{n} \mathbb{1}(S_i \neq T_i) = \sum_{i=1}^{n} \mathbb{1}(T_i \neq S_i) = d_{\text{Hamming}}(T, S)$$

which verifies the symmetry of Hamming distance.

By the properties of indicator function $\mathbb{1}(\cdot)$, it holds that

$$\mathbb{1}(x \neq y) \leq \mathbb{1}(x \neq z) + \mathbb{1}(z \neq y)$$

We can verify this inequality by considering the possible cases. If $x = y$, we have $\mathbb{1}(x \neq y) = 0$ and the inequality holds. If $x \neq y$, we have $\mathbb{1}(x \neq y) = 1$. At the same time, at least one of $x \neq z$ and $z \neq y$ holds, which implies that the right-hand side is at least 1 and the inequality holds. Therefore, it holds that

$$\begin{aligned} d_{\text{Hamming}}(S, T) &= \sum_{i=1}^{n} \mathbb{1}(S_i \neq T_i) \\ &\leq \sum_{i=1}^{n} \mathbb{1}(S_i \neq R_i) + \sum_{i=1}^{n} \mathbb{1}(R_i \neq T_i) \\ &= d_{\text{Hamming}}(S, R) + d_{\text{Hamming}}(R, T) \end{aligned}$$

which verifies the triangle inequality of Hamming distance. Since all the properties are satisfied, we claim that Hamming distance is a metric.

2. (1) Since x and y are uniformly distributed in $[0, L]$, the probability density functions of x and y are given by

$$f_X(x) = \begin{cases} \dfrac{1}{L}, & x \in [0, L] \\ 0, & \text{otherwise} \end{cases}, \quad f_Y(y) = \begin{cases} \dfrac{1}{L}, & y \in [0, L] \\ 0, & \text{otherwise} \end{cases}$$

Since x and y are independent, the joint probability density function is given by

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y) = \begin{cases} \dfrac{1}{L^2}, & x, y \in [0, L] \\ 0, & \text{otherwise} \end{cases}$$

Therefore, the average distance between x and y is given by

$$\begin{aligned} \mathbb{E}\left[|x - y|\right] &= \int_0^L \int_0^L f_{X,Y}(x, y) \cdot |x - y| \, \mathrm{d}x \, \mathrm{d}y \\ &= \frac{1}{L^2} \int_0^L \int_0^L |x - y| \, \mathrm{d}x \, \mathrm{d}y \\ &= \frac{1}{L^2} \int_0^L \mathrm{d}x \int_0^x (x - y) \, \mathrm{d}y + \frac{1}{L^2} \int_0^L \mathrm{d}y \int_0^y (y - x) \, \mathrm{d}x \\ &= \frac{1}{L^2} \cdot \frac{L^3}{6} + \frac{1}{L^2} \cdot \frac{L^3}{6} \\ &= \frac{L}{3} \end{aligned}$$

which proves that the average distance between x and y is $\frac{L}{3}$.

(2) Since $\boldsymbol{x}$ and $\boldsymbol{y}$ are uniformly distributed in $[0, L]^d$, the squared distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ can be specified as

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \sum_{i=1}^d (x_i - y_i)^2$$

Notice that each dimension is independent, which can be computed separately. Therefore, the average squared distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$\begin{aligned} \mathbb{E}\left[\|\boldsymbol{x} - \boldsymbol{y}\|^2\right] &= \sum_{i=1}^d \mathbb{E}\left[(x_i - y_i)^2\right] \\ &= \sum_{i=1}^d \int_0^L \int_0^L f_{X,Y}(x, y) \cdot (x - y)^2 \, \mathrm{d}x \, \mathrm{d}y \\ &= \frac{d}{L^2} \int_0^L \int_0^L (x - y)^2 \, \mathrm{d}x \, \mathrm{d}y \\ &= \frac{d}{L^2} \cdot \frac{L^4}{6} \\ &= \frac{L^2 d}{6} \end{aligned}$$

which indicates that the average squared distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is $\frac{L^2 d}{6}$.

(3) Suppose that $x_1, x_2, \dots, x_n$ are independently and uniformly distributed in $[0, L]$. Let random variable y denote the minimum distance between any two of them, namely

$$y = \min_{1 \le i < j \le n} |x_i - x_j|$$

Note that y is no more than $\frac{L}{n-1}$, which corresponds to the case that $x_1, x_2, \dots, x_n$ distribute uniformly in $[0, L]$ with equal distance, thus $y \in [0, \frac{L}{n-1}]$.

Now we should compute the cumulative distribution function of y. By symmetry, we might as well consider the case that $x_1 \le x_2 \le \dots \le x_n$, which indicates

$$\begin{aligned}
\mathbb{P}\left[y \ge t\right] &= n! \cdot \mathbb{P}\left[y \ge t \land x_1 \le x_2 \le \dots \le x_n\right] \\
&= n! \cdot \mathbb{P}\left[x_i + t \le x_{i+1}, i = 1, 2, \dots, n-1\right] \\
&= \frac{n!}{L^n} \int_0^{L-(n-1)t} \int_{x_1+t}^{L-(n-2)t} \cdots \int_{x_{n-1}+t}^{L} \mathrm{d}x_1 \, \mathrm{d}x_2 \dots \mathrm{d}x_n
\end{aligned}$$

This integral can be computed recursively. For the innermost integral, we have

$$\int_{x_{n-1}+t}^{L} \mathrm{d}x_n = L - x_{n-1} - t$$

For the second innermost integral, we have

$$\int_{x_{n-2}+t}^{L-t} (L - x_{n-1} - t) \, \mathrm{d}x_{n-1} = \frac{1}{2}(L - x_{n-2} - 2t)^2$$

By induction, we have

$$\int_0^{L-(n-1)t} \frac{1}{(n-1)!}(L - x_1 - (n-1)t)^{n-1} \, \mathrm{d}x_1 = \frac{1}{n!}(L - (n-1)t)^n$$

Then the required probability $\mathbb{P}\left[y \ge t\right]$ is given by

$$\mathbb{P}\left[y \ge t\right] = \frac{n!}{L^n} \cdot \frac{1}{n!}(L - (n-1)t)^n = \left(1 - (n-1)\frac{t}{L}\right)^n$$

Therefore, the expected minimum distance between any two points is given by

$$\begin{aligned}
\mathbb{E}\left[\min_{1 \le i < j \le n} |x_i - x_j|\right] &= \int_0^{\frac{L}{n-1}} \mathbb{P}\left[y \ge t\right] \mathrm{d}t \\
&= \int_0^{\frac{L}{n-1}} \left(1 - (n-1)\frac{t}{L}\right)^n \mathrm{d}t \\
&= -\frac{L}{n^2-1} \left(1 - (n-1)\frac{t}{L}\right)^{n-1} \Bigg|_0^{\frac{L}{n-1}} \\
&= \frac{L}{n^2-1}
\end{aligned}$$

which proves that the expected minimum distance between any two points is $\frac{L}{n^2-1}$.

(4) As the dimension d of the space increases, the average distance between two points in the space increases without bound. Almost all the pairs of points are at about the same distance, which brings great difficulty to the clustering algorithm. As the number of points n increases, the average minimum distance between any two points decreases, which benefits the clustering algorithm to some extent.