Data Mining Homework 01

Qiu Yihang

March 2023

1 Whether Distances are Metrics

1.1 Jaccard Distance is A Metric

Proof. Recall that Jaccard Distance is

$$d(C_1, C_2) = 1 - \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$$

Property 1. For any set A, B, $|A \cap B| \le |A \cup B|$, i.e. $\dfrac{|A \cap B|}{|A \cup B|} \le 1$. Thus, $d(A, B) \ge 0$.

Property 2. $d(A, B) = 0$ **iff.** $A = B$. The proof is as follows.

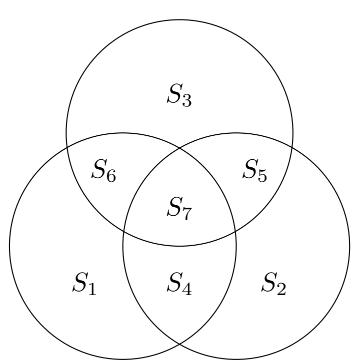
$$A = B \Longrightarrow |A \cup B| = |A \cap B| \Longrightarrow d(A, B) = 0.$$

$$d(A, B) = 0 \Longrightarrow |A \cup B| = |A \cap B| \Longrightarrow \begin{cases} \forall x \in A, x \in A \cup B \Rightarrow x \in A \cap B \Rightarrow x \in B \\ \forall x \in B, x \in A \cup B \Rightarrow x \in A \cap B \Rightarrow x \in A \end{cases}$$

$$\Longrightarrow A \subset B, B \subset A \Longrightarrow A = B.$$

Property 3. $d(A, B) = 1 - \dfrac{|A \cap B|}{|A \cup B|} = 1 - \dfrac{|B \cap A|}{|B \cup A|} = d(B, A)$.

Property 4. For any set A, B, C, $d(A, B) \le d(A, C) + d(C, B)$.



Let $A = \bigcup\{S_1, S_4, S_6, S_7\}, B = \bigcup\{S_3, S_5, S_6, S_7\}, C = \bigcup\{S_2, S_4, S_5, S_7\}$, where $\forall i, j \in \{1, 2, ..., 7\}, S_i \cap S_j = \varnothing$.

First we prove <u>the lemma</u> that for $A, B, C \subseteq X$, it holds that

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

The proof is as follows.

$$\begin{aligned}
& |A \cap C| \cdot |B \cup C| + |A \cup C| \cdot |B \cap C| \leq |C| \cdot (|A| + |B|) \\
\iff & (|S_6| + |S_7|) \cdot (|S_2| + |S_4| + |S_5| + |S_7| + |S_3| + |S_6|) + \\
& (|S_4| + |S_7|) \cdot (|S_1| + |S_4| + |S_6| + |S_7| + |S_3| + |S_5|) \\
& \leq (|S_3| + |S_5| + |S_6| + |S_7|) \cdot (|S_1| + |S_4| + |S_6| + |S_7| + |S_2| + |S_4| + |S_5| + |S_7|) \\
\iff & (|S_6| + |S_7|) \cdot |S_3| + (|S_4| + |S_7|) \cdot (|S_5| + |S_6|) \\
& \leq (|S_3| + |S_5|)(|S_2| + |S_4| + |S_5| + |S_7| + |S_6|) + (|S_6| + |S_7|)(|S_1| + |S_4| + |S_7|) \\
\iff & 0 \leq |S_3| \cdot |S_2| + |S_3| \cdot |S_4| + |S_3| \cdot |S_5| + |S_2| \cdot |S_5| + |S_5|^2 + \\
& |S_5| \cdot |S_6| + |S_6| \cdot |S_1| + |S_7| \cdot |S_1| + |S_7| \cdot |S_4| + |S_7|^2 \text{ (Always Holds.)}
\end{aligned}$$

Moreover, we have $|A \cup C| \cdot |B \cup C| \geq |A \cup B| \cdot |C|$. (The proof is as follows.)

$$\begin{aligned}
& |A \cup C| \cdot |B \cup C| \geq |A \cup B| \cdot |C| \\
\iff & (|S_1| + |S_4| + |S_6| + |S_7| + |S_3| + |S_5|) \cdot (|S_2| + |S_4| + |S_5| + |S_7| + |S_3| + |S_6|) \\
& \geq (|S_1| + |S_4| + |S_6| + |S_7| + |S_2| + |S_5|) \cdot (|S_3| + |S_5| + |S_6| + |S_7|) \\
\iff & |S_3| \cdot (|S_2| + |S_4| + |S_5| + |S_7| + |S_3| + |S_6|) + (|S_1| + |S_4| + |S_6| + |S_7| + |S_5|) \cdot (|S_2| + |S_4|) \\
& \geq |S_2| \cdot (|S_3| + |S_5| + |S_6| + |S_7|) \\
\iff & (|S_4| + |S_5| + |S_6| + |S_7|) \cdot (|S_3| + |S_4|) + |S_3|^2 + |S_2| \cdot (|S_1| + |S_4|) \geq 0
\end{aligned}$$

Meanwhile, we have

$$\begin{aligned}
|A| + |B| &= |S_1| + |S_4| + |S_6| + |S_7| + |S_2| + |S_5| \\
&= (|S_1| + |S_6| + |S_2| + |S_5|) + (|S_4| + |S_7|) = |A \cup B| + |A \cap B|
\end{aligned}$$

Then we know

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

i.e. $\quad \dfrac{|A \cap C|}{|A \cup C|} + \dfrac{|B \cap C|}{|B \cup C|} \leq \dfrac{|A \cap B|}{|A \cup B|} + 1 \iff 1 - \dfrac{|A \cap B|}{|A \cup B|} \leq 1 - \dfrac{|A \cap C|}{|A \cup C|} + 1 - \dfrac{|C \cap B|}{|C \cup B|}$

$\iff d(A, B) \leq d(A, C) + d(C, B)$

Therefore, Jaccard distance is a metric. ■

1.2 Cosine Distance is Not A Metric

Disproof. Recall that cosine distance for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ is

$$d(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2}\sqrt{\sum_{i=1}^{d} y_i^2}}$$

Obvious exists $\boldsymbol{x} = (1, 1, 1, 1)$ and $\boldsymbol{y} = (-1, -1, -1, -1)$ s.t. $d(\boldsymbol{x}, \boldsymbol{y}) = -1 < 0$.

Thus, cosine distance is not a metric. ■

1.3 Edit Distance is A Metric

Proof. Recall that edit distance (Levenshtein Distance) for two strings x and y is

$$\operatorname{lev}(x, y) = \begin{cases} |x|, & \text{if } |y| = 0 \\ |y|, & \text{if } |x| = 0 \\ \operatorname{lev}(x[1:], y[1:]), & \text{if } x[0] = y[0] \\ 1 + \min\{\operatorname{lev}(x[1:], y), \operatorname{lev}(x, y[1:]), \operatorname{lev}(x[1:], y[1:])\} & \text{otherwise} \end{cases}$$

where $s[k:]$ means the string of all but the first k characters of s.

Property 1. By the definition of $\operatorname{lev}(\cdot)$, obvious for any string x, y, $\operatorname{lev}(x, y) \geq 0$.

Property 2. For any string x, there is

$$\operatorname{lev}(x, x) = \operatorname{lev}(x[1:], x[1:]) = \operatorname{lev}(x[2:], x[2:]) = \cdots = \operatorname{lev}(x[|x|:], x[|x|:]) = 0.$$

Property 3. By the definition of edit distance, it is obvious that $\operatorname{lev}(x, y) = lev(y, x)$.

Property 4. For any string x, y and z, $\operatorname{lev}(x, y) \leq \operatorname{lev}(x, z) + \operatorname{lev}(z, y)$. The proof is as follows.

We prove Property 4 by contradiction.

Assume exist x, y, z s.t. $\operatorname{lev}(x, y) > \operatorname{lev}(x, z) + \operatorname{lev}(z, y)$.

Consider the arguments of $\operatorname{lev}(\cdot)$ when recursively calculate $\operatorname{lev}(x, y), \operatorname{lev}(x, z), \operatorname{lev}(z, y)$. Then we get three sequences, $x \to a_1 \to a_2 \to \cdots \to a_N \to y$, $x \to b_1 \to b_2 \to\to \ldots b_M \to z$, $z \to c_1 \to c_2 \to \cdots \to c_K \to y$, which in fact tell us how to change x to y, x to z, z to y respectively.

By the meaning of edit distance, we know $\operatorname{lev}(x, y), \operatorname{lev}(x, z), \operatorname{lev}(z, y)$ are costs when we change $x \to a_1 \to a_2 \to \cdots \to a_N \to y$, $x \to b_1 \to b_2 \to \cdots \to b_M \to z$, $z \to c_1 \to c_2 \to \cdots \to c_K \to y$ respectively. Moreover, $x \to a_1 \to a_2 \to \ldots a_N \to y$ is the way to change x to y with the minimal cost.

Meanwhile, there exists a sequence $x \to b_1 \to b_2 \to \cdots \to b_M \to z \to c_K \to \cdots \to c_2 \to c_1 \to y$ with cost $\operatorname{lev}(x, z) + \operatorname{lev}(z, y) < \operatorname{lev}(x, y)$. **Contradiction!**

In conclusion, edit distance is a metric. ■

1.4 Hamming Distance is A Metric

Proof. Recall that Hamming distance for two strings x and y with the same length L is

$$d(x,y) = \sum_{i=1}^{L} \mathbb{1}[x_i \neq y_i]$$

Property 1. Obvious for any string x, y, $d(x,y) \geq 0$.

Property 2. For any string x, $d(x,x) = \sum_{i=1}^{L} 0 = 0$.

Property 3. For any string x and y, $d(x,y) = \sum_{i=1}^{L} \mathbb{1}[x_i \neq y_i] = \sum_{i=1}^{L} \mathbb{1}[y_i \neq x_i] = d(y,x)$.

Property 4. For any string x, y and z, $d(x,y) \leq d(x,z) + d(z,y)$. The proof is as follows.

First we prove that $\mathbb{1}[a \neq b] \leq \mathbb{1}[a \neq c] + \mathbb{1}[c \neq b]$ for any char a, b, c by contradiction.

Assume exists a, b, c s.t. $\mathbb{1}[a \neq b] \leq \mathbb{1}[a \neq c] + \mathbb{1}[c \neq b]$.

The only possible case is that $\nVdash[a \neq b] = 1$ and $\mathbb{1}[a \neq c] = \mathbb{1}[c \neq b] = 0$.

Then we get $a \neq b$ and $a = c = b$. **Contradiction!**

Further, we have

$$\begin{aligned} d(x,y) = \sum_{i=1}^{L} \mathbb{1}[x_i \neq y_i] &\leq \sum_{i=1}^{L} \mathbb{1}[x_i \neq z_i] + \mathbb{1}[z_i \neq y_i] \\ &= \sum_{i=1}^{L} \mathbb{1}[x_i \neq z_i] + \sum_{i=1}^{L} \mathbb{1}[z_i \neq y_i] = d(x,z) + d(z,y) \end{aligned}$$

Therefore, Hamming Distance is a metric. ■

2 Average Distance Between A Pair Of Points

Proof. Let the line of length L be AB. Let the pair of points be M and N. Let $x \triangleq |AM|, y \triangleq |AN|$.

Obvious $x, y \sim \mathcal{U}(0, L)$.

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

Thus, the average distance between a pair of points is $\frac{1}{3}L$. ■

3 Eckart-Young-Mirsky Theorem

Let $\boldsymbol{A} = \boldsymbol{U\Sigma V}^\top$ and $\boldsymbol{B} = \boldsymbol{USV}^\top$ where $\boldsymbol{S}$ = diagonal $r \times r$ matrix with $s_i = \begin{cases} \sigma_i & \text{if } i = 1, 2, ..., k \\ 0 & \text{otherwise} \end{cases}$

Prove that $\boldsymbol{B}$ is one of the best k-rank approximations to $\boldsymbol{A}$ in terms of Frobenius norm error.

Proof. Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. To prove the proposition, we just need to prove that

$$\text{for any } \boldsymbol{C} \in \left\{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \mathtt{rank}(\boldsymbol{X}) = k\right\},\ \min_{\boldsymbol{C}} \|\boldsymbol{A} - \boldsymbol{C}\|_F = \|\boldsymbol{A} - \boldsymbol{B}\|_F.$$

First we prove the following lemma.

<u>Lemma.</u> For any matrix $\boldsymbol{A} \in \mathbb{R}^{m' \times n'}$ and $\boldsymbol{W} \in \mathbb{R}^{p' \times n'}$ s.t. $\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}$, $\|\boldsymbol{AW}^\top\|_F = \|\boldsymbol{A}\|_F$.

For any matrix $\boldsymbol{A} \in \mathbb{R}^{m' \times n'}$ and $\boldsymbol{V} \in \mathbb{R}^{m' \times p'}$ s.t. $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$, $\|\boldsymbol{VA}\|_F = \|\boldsymbol{A}\|_F$.

The proof for the lemma is as follows.

$$\begin{aligned}
\|\boldsymbol{AW}^\top\|_F &= \sqrt{\mathtt{tr}\left(\left(\boldsymbol{AW}^\top\right)^\top \boldsymbol{AW}^\top\right)} = \sqrt{\mathtt{tr}\left(\boldsymbol{WA}^\top\boldsymbol{AW}^\top\right)} = \sqrt{\mathtt{tr}\left(\boldsymbol{A}^\top\boldsymbol{AW}^\top\boldsymbol{W}\right)} \\
&= \sqrt{\mathtt{tr}\left(\boldsymbol{A}^\top\boldsymbol{AI}\right)} = \sqrt{\mathtt{tr}\left(\boldsymbol{A}^\top\boldsymbol{A}\right)} = \|\boldsymbol{A}\|_F \qquad \text{where } \boldsymbol{I} \text{ is identity matrix.} \\
\|\boldsymbol{VA}\|_F^2 &= \sqrt{\mathtt{tr}\left((\boldsymbol{VA})^\top \boldsymbol{VA}\right)} = \sqrt{\mathtt{tr}\left(\boldsymbol{A}^\top\boldsymbol{V}^\top\boldsymbol{VA}\right)} = \sqrt{\mathtt{tr}\left(\boldsymbol{A}^\top\boldsymbol{IA}\right)} \\
&= \sqrt{\mathtt{tr}\left(\boldsymbol{A}^\top\boldsymbol{A}\right)} = \|\boldsymbol{A}\|_F \qquad \text{where } \boldsymbol{I} \text{ is identity matrix.}
\end{aligned}$$

□

By the lemma, we know

$$\|\boldsymbol{A} - \boldsymbol{B}\|_F^2 = \|\boldsymbol{U}\left(\boldsymbol{\Sigma} - \boldsymbol{S}\right)\boldsymbol{V}^\top\|_F^2 = \|\boldsymbol{\Sigma} - \boldsymbol{S}\|_F^2 = \sum_{i=1}^{r}(\sigma_i - s_i)^2 = \sum_{i=k+1}^{r}\sigma_i^2.$$

Now we just need to prove that

$$\text{for any } \boldsymbol{C} \in \left\{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \mathtt{rank}(\boldsymbol{X}) = k\right\},\ \min_{\boldsymbol{C}} \|\boldsymbol{A} - \boldsymbol{C}\|_F^2 = \sum_{i=k+1}^{r}\sigma_i^2.$$

By SVD Theorem, we know there exist matrices $\widetilde{\boldsymbol{U}}, \widetilde{\boldsymbol{V}}$ and an $r \times r$ diagonal matrix $\boldsymbol{\Gamma} = \mathtt{diag}\left(\gamma_1, \gamma_2, ...\gamma_k, 0, ..., 0\right)$ s.t. $\boldsymbol{C} = \widetilde{\boldsymbol{U}}\boldsymbol{\Gamma}\widetilde{\boldsymbol{V}}^\top$. By the lemma, we know $\|\boldsymbol{C}\|_F = \sum_{i=1}^r \gamma_i^2$.

Let $\hat{\boldsymbol{U}} = \boldsymbol{U}^\top\widetilde{\boldsymbol{U}} = \left(\hat{\boldsymbol{u}}_1, \hat{\boldsymbol{u}}_2, ..., \hat{\boldsymbol{u}}_r\right)^\top, \hat{\boldsymbol{V}}^\top = \boldsymbol{V}^\top\widetilde{\boldsymbol{V}} = \left(\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2, ..., \hat{\boldsymbol{v}}_r\right)$.

Obvious $\hat{\boldsymbol{U}}$ is orthogonal since $\hat{\boldsymbol{U}}^\top\hat{\boldsymbol{U}} = \widetilde{\boldsymbol{U}}^\top\boldsymbol{UU}^\top\widetilde{\boldsymbol{U}} = \widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}} = \boldsymbol{I}$. Similarly, $\hat{\boldsymbol{V}}$ is orthogonal.

Then we have

$$\begin{aligned}
\langle \boldsymbol{A}, \boldsymbol{C} \rangle &= \mathtt{tr}\left(\boldsymbol{V\Sigma}^\top\boldsymbol{U}^\top\widetilde{\boldsymbol{U}}\boldsymbol{\Gamma}\widetilde{\boldsymbol{V}}^\top\right) = \mathtt{tr}\left(\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top\widetilde{\boldsymbol{U}}\boldsymbol{\Gamma}\widetilde{\boldsymbol{V}}^\top\boldsymbol{V}\right) = \mathtt{tr}\left(\boldsymbol{\Sigma}^\top\hat{\boldsymbol{U}}\boldsymbol{\Gamma}\hat{\boldsymbol{V}}^\top\right) \\
&= \sum_{i=1}^{k}\sigma_i\gamma_i\hat{\boldsymbol{u}}_i^\top\hat{\boldsymbol{v}}_i \le \sum_{i=1}^{k}\sigma_i\gamma_i\|\hat{\boldsymbol{u}}_i\|\|\hat{\boldsymbol{v}}_i\| = \sum_{i=1}^{k}\sigma_i\gamma_i.
\end{aligned}$$

Thus, we have

$$
\begin{aligned}
\|\boldsymbol{A}-\boldsymbol{C}\|_F^2 &= \mathtt{tr}\left(\left(\boldsymbol{A}-\boldsymbol{C}\right)^\top\left(\boldsymbol{A}-\boldsymbol{C}\right)\right) = \mathtt{tr}\left(\boldsymbol{A}^\top\boldsymbol{A}-\boldsymbol{A}^\top\boldsymbol{C}-\boldsymbol{C}^\top\boldsymbol{A}+\boldsymbol{C}^\top\boldsymbol{C}\right) \\
&= \mathtt{tr}\left(\boldsymbol{A}^\top\boldsymbol{A}\right) - 2\langle\boldsymbol{A},\boldsymbol{C}\rangle + \mathtt{tr}\left(\boldsymbol{C}^\top\boldsymbol{C}\right) = \|\boldsymbol{A}\|_F - 2\langle\boldsymbol{A},\boldsymbol{C}\rangle + \|\boldsymbol{C}\|_F \\
&\geq \sum_{i=1}^{r}\sigma_i^2 - 2\sum_{i=1}^{k}\sigma_i\gamma_i + \sum_{i=1}^{k}\gamma_i^2 = \sum_{i=k+1}^{r}\sigma_i^2 + \sum_{i=1}^{k}\left(\sigma_i-\gamma_i\right)^2 \\
&\geq \sum_{i=k+1}^{r}\sigma_i^2.
\end{aligned}
$$

Therefore, $\boldsymbol{B}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ is one of the best k-rank approximations to $\boldsymbol{A}$ in terms of Frobenius norm error. ■

4 Average Jaccard Similarity of Two Randomly-Sampled Sets

Solution. Let $J(A,B)$ be the Jaccard similarity of set A and B.

Totally, there are $\binom{n}{m}\cdot\binom{n}{m}$ possible pairs of (S,T).

It is obvious that $\max\{0,2m-n\}\leq|S\cap T|\leq m$.

When $|S\cap T|=k$, there are $\binom{n}{k}\cdot\binom{n-k}{m-k}\cdot\binom{n-m}{m-k}$ possible pairs of (S,T).

In this case, the Jaccard similarity of S and T is $\dfrac{k}{2m-k}$.

Thus,

$$
\mathbb{E}\left[J(S,T)\right] = \sum_{k=\max\{0,2m-n\}}^{m}\frac{\binom{n}{k}\cdot\binom{n-k}{m-k}\cdot\binom{n-m}{m-k}}{\binom{n}{m}^2}\cdot\frac{k}{2m-k}.
$$

■