BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN ĐHQG-HCM
KHOA TOÁN – TIN HỌC
---o0o---

BTLT TUẦN 4: KHAI PHÁ DỮ LIỆU

Giáo viên hướng dẫn: Nguyễn Thanh Bình

Sinh viên thực hiện: Nguyễn Công Hoài Nam

Mã số sinh viên: 21280099

Tp. Hồ Chí Minh, ngày 6 tháng 4 năm 2024

1. Compute the L_p – norm between (1, 2) and (3, 4) for p = 1, 2, ∞

Ta có với hai điểm dữ liệu $\bar{X} = (x_1, \ldots, x_n)$ và $\bar{Y} = (y_1, \ldots, y_n)$ chuẩn L_p giữa hai điểm là

$$Dist(\bar{X}, \bar{Y}) = \|\bar{X} - \bar{Y}\|_p = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

Với X(1,2) và Y(3,4) ta có

- Với p = 1, $\|X - Y\|_1 = |1 - 3| + |2 - 4| = 2 + 2 = 2$
- Với p = 2, $\|X - Y\|_2 = (|1 - 3|^2 + |2 - 4|^2)^{1/2} = \sqrt{8} = 2\sqrt{2}$
- Với p = ∞, $\|X - Y\|_\infty = \max\limits_{1 \le i \le n} |x_i - y_i| = max(|1 - 3|, |2 - 4|) = 2$

Giải thích trường hợp $p = \infty$, đặt Z = X - Y ta có:
Vì $|z_i| \le \max\limits_{1 \le i \le n} |z_i|$ nên:

$$\|Z\|_p = \left(\sum_{i=1}^{n} |z_i|^p \right)^{1/p} \ge \max_{1 \le i \le n} |z_i| \tag{1}$$

Và,

$$\|Z\|_p \le \left(\sum_{i=1}^{n} \max_{1 \le i \le n} |z_i|^p \right)^{1/p} \le n^{1/p} \max_{1 \le i \le n} |z_i| \tag{2}$$

Từ (1) và (2) ta có:

$$\max_{1 \le i \le n} |z_i| \le \|Z\|_p \le n^{1/p} \max_{1 \le i \le n} |z_i| \tag{3}$$

Thế $p = \infty$ vào (3):

$$\begin{aligned} &\max_{1 \le i \le n} |z_i| \le \|Z\|_\infty \le n^{1/\infty} \max_{1 \le i \le n} |z_i| \\ \Leftrightarrow &\max_{1 \le i \le n} |z_i| \le \|Z\|_\infty \le 1 \max_{1 \le i \le n} |z_i| \end{aligned}$$

Vì $\|Z\|_\infty$ kẹp giữa hai $\max\limits_{1 \le i \le n} |z_i|$ nên:

$$\|Z\|_\infty = \max_{1 \le i \le n} |z_i|$$

Hay

$$\|X - Y\|_\infty = \max_{1 \le i \le n} |x_i - y_i|$$

2. Show that the Mahalanobis distance between two data points is equivalent to the Euclidean distance on a transformed data set, where the transformation is performed by representing the data along the principal components, and dividing by the standard deviation of each component.

Ta đã biết với Σ là ma trận hiệp phương sai của tập dữ liệu. Với hai điểm $\bar{X}$ và $\bar{Y}$, khoảng cách Mahalanobis $Maha(\bar{X}, \bar{Y})$ được cho bởi công thức:

$$Maha(X, Y) = (X - Y) \cdot \Sigma^{-1} \cdot (X - Y)^T \tag{4}$$

Ta tiến hành chéo hoá ma trận Σ

$$\begin{aligned} &\Sigma = P \Lambda P^T \\ \Rightarrow &\Sigma^{-1} = (P \Lambda P^T)^{-1} = (P^T)^{-1} \Lambda^{-1} P^{-1} = P \Lambda^{-1} P^T \\ \Rightarrow &\Sigma^{-1} = P \Lambda^{-1} P^T \end{aligned} \tag{5}$$

Lúc này ma trận Σ được phân rã thành:

- P là ma trận gồm các vector riêng

- Λ là ma trận chéo chứa các giá trị riêng ứng với vector riêng (khi nay là phương sai ứng với thành phần chính)
- P^T là chuyển vị của P

Khi đó Λ^{-1} là nghịch đảo của Λ tức là nghịch đảo của phương sai của thành phần chính, và phương sai bằng bình phương độ lệch chuẩn.
Thế (5) vào (4) ta được:

$$
\begin{aligned}
Maha(X,Y) &= (X-Y)\cdot\Sigma^{-1}\cdot(X-Y)^T \\
Maha(X,Y) &= (X-Y)\cdot P\Lambda^{-1}P^T\cdot(X-Y)^T \\
Maha(X,Y) &= [(X-Y)P]\Lambda^{-1}[(X-Y)P]^T \\
Maha(X,Y) &= (\bar{X}-\bar{Y})\Lambda^{-1}(\bar{X}-\bar{Y})^T \\
&\text{Với } \bar{X}=XP \text{ và } \bar{Y}=YP
\end{aligned}
$$

Khi đó $\bar{X}$ và $\bar{Y}$ là các điểm dữ liệu trong tập dữ liệu đã được biến đổi.
Như đã nói ở trên Λ^{-1} là ma trận chéo chứa các nghịch đảo phương sai nên khoảng cách Mahalanobis bằng với khoảng cách Euclid khi bộ dữ liệu đã được biến đổi theo thành phần chính và chia cho độ lệch chuẩn (căn bậc 2 phương sai) với mỗi thành phần

3. Compute the match-based similarity, cosine similarity and the Jaccard coefficient, between the two sets A, B, C and A, C, D, E

Đặt $X=\{A,B,C\}$ và $Y=\{A,C,D,E\}$
Ta có bảng nhị phân

	A	B	C	D	E
X	1	1	1	0	0
Y	1	0	1	1	1

a) Match-based similarity

Độ tương đồng Match-based:

$$Match-based(X,Y)=\frac{f_{11}}{f_{11}+f_{10}+f_{01}}$$

Với

$$
\begin{aligned}
f_{11} &:= X=1, Y=1 \Rightarrow f_11=\{A,C\}=2 \\
f_{10} &:= X=1, Y=0 \Rightarrow f_10=\{B\}=1 \\
f_{01} &:= X=0, Y=1 \Rightarrow f_01=\{D,E\}=2
\end{aligned}
$$

Suy ra

$$Match-based(X,Y)=\frac{f_{11}}{f_{11}+f_{10}+f_{01}}=\frac{2}{1+2+2}=\frac{2}{5}=0,4$$

b) Cosine similarity

Độ tương đồng Consine là:

$$Cosine(X,Y)=\frac{X\cdot Y}{\|X\|\cdot\|Y\|}=\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

Mà

$$X \cdot Y = \sum_i x_i y_i = (1 \cdot 1) + (1 \cdot 0) + (1 \cdot 1) + (0 \cdot 1) = 2$$

$$\|X\| = \sqrt{\sum_i x_i^2} = 1^2 + 1^2 + 1^2 + 0^2 + 0^2 = \sqrt{3}$$

$$\|Y\| = \sqrt{\sum_i x_i^2} = 1^2 + 0^2 + 1^2 + 1^2 + 1^2 = \sqrt{4}$$

$$\Rightarrow \|X\| \cdot \|Y\| = \sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2} = \sqrt{3} \cdot \sqrt{4} = \sqrt{12}$$

Suy ra

$$Cosine(X, Y) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} = \frac{2}{\sqrt{12}} = \frac{1}{\sqrt{3}}$$

c) Jaccard coefficient

Hệ số Jaccard:

$$Jaccard(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Mà

$$X \cap Y = \{A, C\} \Rightarrow |X \cap Y| = 2$$
$$X \cup Y = \{A, B, C, D, E\} \Rightarrow |X \cup Y| = 5$$

Suy ra

$$Jaccard(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{2}{5} = 0,4$$

4. Let X and Y be two data points. Show that the cosine angle between the vectors X and Y is given by:

$$Cosine(\bar{X}, \bar{Y}) = \frac{\|\bar{X}\|^2 + \|\bar{Y}\|^2 - \|\bar{X} - \bar{Y}\|^2}{2\|\bar{X}\|\|\bar{Y}\|} \tag{6}$$

Từ bài trước ta có công thức hệ số Cosine:

$$\text{Cosine}(\bar{X}, \bar{Y}) = \frac{\bar{X} \cdot \bar{Y}}{|\bar{X}| \cdot |\bar{Y}|} = \frac{\sum_i \bar{x}_i \bar{y} i}{\sqrt{\sum_i \bar{x} i^2}\sqrt{\sum_i \bar{y}_i^2}} \tag{7}$$

Ta lại có,

$$|\bar{X} - \bar{Y}| = (\bar{X} - \bar{Y}) \cdot (\bar{X} - \bar{Y}) = \bar{X} \cdot \bar{X} - 2(\bar{X} \cdot \bar{Y}) + \bar{Y} \cdot \bar{Y} \text{ (Vì } \bar{X} \cdot \bar{Y} = \bar{Y} \cdot \bar{X}) \tag{8}$$

Thế (8) vào (6) ta được:

$$\begin{aligned} Cosine(\bar{X}, \bar{Y}) &= \frac{\|\bar{X}\|^2 + \|\bar{Y}\|^2 - X \cdot X - 2(X \cdot Y) + Y \cdot Y}{2\|\bar{X}\|\|\bar{Y}\|} \\ &= \frac{\|\bar{X}\|^2 + \|\bar{Y}\|^2 - \|\bar{X}\|^2 + 2(X \cdot Y) - \|\bar{Y}\|^2}{2\|\bar{X}\|\|\bar{Y}\|} \\ &= \frac{2 \cdot \bar{X} \cdot \bar{Y}}{2 \cdot \|\bar{X}\| \cdot \|\bar{Y}\|} \\ &= \frac{\bar{X} \cdot \bar{Y}}{\|\bar{X}\| \cdot \|\bar{Y}\|} = (7) \\ &\Rightarrow \textbf{Điều phải chứng minh} \end{aligned}$$