# Bài tập Lý thuyết môn Khai phá Dữ liệu

**Họ và tên:** Huỳnh Nguyễn Thế Dân

**MSSV:** 21110256

**Lớp:** 21TTH1

## Questions

## 1. Compute the Lp-norm between (1, 2) and (3, 4) for p = 1, 2, ∞.

For two points (x1, y1) and (x2, y2), the Lp-norm is defined as:

$$\left(\left|x_1 - x_2\right|^p + \left|y_1 - y_2\right|^p\right)^{\frac{1}{p}}$$

For p = 1:

$$\left(\left|1-3\right| + \left|2-4\right|\right)^{\frac{1}{1}} = \left(2+2\right)^{\frac{1}{1}} = 4$$

For p = 2 (Euclidean distance):

$$\left(\left|1-3\right|^2 + \left|2-4\right|^2\right)^{\frac{1}{2}} = \left(4+4\right)^{\frac{1}{2}} = \sqrt{8}$$

For p = ∞:

$$\max\left(\left|1-3\right|, \left|2-4\right|\right) = \max\left(2,2\right) = 2$$

So,

The L1-norm is 4

The L2-norm is $\sqrt{8}$

The L∞-norm is 2

## 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.

Let's denote the Mahalanobis distance between two points $x$ and $y$ as $D_M(x, y)$. It is defined as:

$$D_M(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$

Where $\Sigma$ is the covariance matrix of the data.

Now, let's perform Principal Component Analysis (PCA) on the data to get the principal components $v_1, v_2, \ldots, v_n$ and the corresponding standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_n$. Then, the transformed data points $x'$ and $y'$ can be obtained by projecting the original data points onto these principal components and dividing by the standard deviation of each component.

Now, the Mahalanobis distance becomes:

$$D_M(x', y') = \sqrt{(x' - y')^T \Sigma'^{-1} (x' - y')}$$

Where $\Sigma'$ is the covariance matrix of the transformed data.

Since the principal components are orthogonal, $\Sigma'$ is diagonal with entries $\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \ldots, \frac{1}{\sigma_n^2}$.

Thus, $\Sigma'^{-1}$ is also diagonal with entries $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$.

Now, $(x' - y')$ is simply the difference between the transformed points $x'$ and $y'$.

So, $(x' - y')^T \Sigma'^{-1} (x' - y')$ is the sum of squares of the components of $(x' - y')$ weighted by $\sigma_i^2$, which is precisely the Euclidean distance between $x'$ and $y'$.

Thus, $D_M(x', y') = \|x' - y'\|$, which is the Euclidean distance between the transformed points.

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

Match-based similarity: The match-based similarity is the ratio of the number of common elements to the total number of distinct elements in both sets.

$$\text{Common elements} = \{A, C\}$$

$$\text{Total distinct elements} = \{A, B, C, D, E\}$$

$$\text{Match-based similarity} = \frac{|\text{Common elements}|}{|\text{Total distinct elements}|} = \frac{2}{5}$$

Cosine similarity: Cosine similarity measures the cosine of the angle between two vectors representing the sets in a high-dimensional space.

$$\text{Set 1 vector} = (1, 1, 1, 0, 0)$$

$$\text{Set 2 vector} = (1, 0, 1, 1, 1)$$

$$\text{Cosine similarity} = \frac{\text{Set 1 vector} \cdot \text{Set 2 vector}}{||\text{Set 1 vector}|| \cdot ||\text{Set 2 vector}||} = \frac{2}{\sqrt{3} \cdot \sqrt{4}} = \frac{2}{2\sqrt{3}} = \frac{1}{\sqrt{3}}$$

Jaccard coefficient: Jaccard coefficient is the ratio of the size of the intersection of the sets to the size of their union.

$$\text{Intersection} = \{A, C\}$$

$$\text{Union} = \{A, B, C, D, E\}$$

$$\text{Jaccard coefficient} = \frac{|\text{Intersection}|}{|\text{Union}|} = \frac{2}{5}$$

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

$$\cos(\mathbf{X}, \mathbf{Y}) = \frac{||\mathbf{X}||^2 + ||\mathbf{Y}||^2 - ||\mathbf{X} - \mathbf{Y}||^2}{2||\mathbf{X}||||\mathbf{Y}||}$$

Proof: Using the cosine similarity formula:

$$\text{cosine}(X, Y) = \frac{X \cdot Y}{||X||||Y||}$$

Expanding the dot product:

$$X \cdot Y = ||X||||Y|| \cdot \cos(\theta)$$

Where $\theta$ is the angle between the vectors X and Y.

Now, solve for $\cos(\theta)$:

$$\cos(\theta) = \frac{X \cdot Y}{||X||||Y||} = \frac{||X||||Y|| \cdot \cos(\theta)}{||X||||Y||}$$

$$\Rightarrow \cos(\theta) = \frac{X \cdot Y}{\|X\|\|Y\|}$$

Now, square both sides:

$$\cos^2(\theta) = \frac{(X \cdot Y)^2}{\|X\|^2\|Y\|^2}$$

$$\cos^2(\theta) = \frac{(X \cdot Y)^2}{\|X\|^2\|Y\|^2} = \frac{(X \cdot Y)^2}{(\|X\|^2)(\|Y\|^2)}$$

$$(\cos^2(\theta))(\|X\|^2)(\|Y\|^2) = (X \cdot Y)^2$$

$$\Rightarrow \cos^2(\theta)(\|X\|^2)(\|Y\|^2) = (X \cdot Y)^2$$

Now, solve for $\cos(\theta)$:

$$\cos(\theta) = \frac{X \cdot Y}{\|X\|\|Y\|} = \pm\frac{(X \cdot Y)}{\|X\|\|Y\|} = \pm\frac{\|X\|\|Y\| \cdot \cos(\theta)}{\|X\|\|Y\|}$$

$$\Rightarrow \cos(\theta) = \pm\frac{X \cdot Y}{\|X\|\|Y\|}$$

Thus, the cosine similarity is given by:

$$\text{cosine}(X, Y) = \pm\frac{X \cdot Y}{\|X\|\|Y\|}$$

And since the cosine function is even (i.e., $\cos(-\theta) = \cos(\theta)$), the positive sign can be used.

$$\text{cosine}(X, Y) = \frac{X \cdot Y}{\|X\|\|Y\|}$$

This completes the proof.

Hết.