Homework 01

2253323 陶子芾

Q1. Distance calculation and similarity analysis

1） Euclidean Distance:

$$d(A,B)=\sqrt{\sum(A_i-B_i)^2}$$

$$\begin{aligned} d(A,B)&=\sqrt{(3-1)^2+(5-9)^2+(2-4)^2+(7-4)^2}\\ &=\sqrt{4+16+4+9}=\sqrt{33}\approx 5.74 \end{aligned}$$

$$\begin{aligned} d(A,C)&=\sqrt{(3-6)^2+(5-2)^2+(2-8)^2+(7-3)^2}\\ &=\sqrt{9+9+36+16}=\sqrt{70}\approx 8.37 \end{aligned}$$

$$\begin{aligned} d(B,C)&=\sqrt{(1-6)^2+(9-2)^2+(4-8)^2+(4-3)^2}\\ &=\sqrt{25+49+16+1}=\sqrt{91}\approx 9.54 \end{aligned}$$

Manhattan Distance：

$$d(A,B)=\sum|A_i-B_i|$$

$$\begin{aligned} d(A,B)&=|3-1|+|5-9|+|2-4|+|7-4|\\ &=2+4+2+3=11 \end{aligned}$$

$$\begin{aligned} d(A,C)&=|3-6|+|5-2|+|2-8|+|7-3|\\ &=3+3+6+4=16 \end{aligned}$$

$$\begin{aligned} d(B,C)&=|1-6|+|9-2|+|4-8|+|4-3|\\ &=5+7+4+1=17 \end{aligned}$$

Mahalanobis Distance：

$$\begin{aligned} \mu=\frac{A+B+C}{3}&=\left(\frac{3+1+6}{3},\frac{5+9+2}{3},\frac{2+4+8}{3},\frac{7+4+3}{3}\right)\\ &=\left(\frac{10}{3},\frac{16}{3},\frac{14}{3},\frac{14}{3}\right) \end{aligned}$$

Calculate covariance matrix S：

$$S=\frac{1}{n-1}X^TX$$

$$X = \begin{bmatrix} -\frac{1}{3} & -\frac{1}{3} & -\frac{8}{3} & \frac{7}{3} \\ -\frac{7}{3} & \frac{11}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{8}{3} & -\frac{10}{3} & \frac{10}{3} & -\frac{5}{3} \end{bmatrix}$$

$$S = \begin{bmatrix} 6.33 & -8.67 & 5.67 & -1.83 \\ -8.67 & 12.33 & -6.33 & -1.16 \\ 5.67 & -6.33 & 9.33 & -5.67 \\ -1.83 & 1.17 & -5.67 & 4.33 \end{bmatrix}$$

$$S^{-1} = \begin{bmatrix} 0.015 & -0.027 & -0.004 & 0.012 \\ -0.0267 & 0.052 & 0.023 & -0.036 \\ -0.004 & 0.023 & 0.05 & -0.049 \\ 0.012 & -0.036 & -0.049 & 0.053 \end{bmatrix}$$

$$d(A,B) = \sqrt{(A-B)^T S^{-1}(A-B)} \approx 2.00$$

$$d(A,C) = \sqrt{(A-C)^T S^{-1}(A-C)} \approx 2.00$$

$$d(B,C) = \sqrt{(B-C)^T S^{-1}(B-C)} \approx 2.00$$

Cosine Similarity：

$$\cos(A,B) = \frac{A \cdot B}{||A|| \times ||B||}$$

$$\cos(A,B) = \frac{84}{(9.33 \times 10.68)} \approx 0.843$$

$$\cos(A,C) = \frac{65}{(9.33 \times 10.63)} \approx 0.655$$

$$\cos(B,C) = \frac{68}{(10.68 \times 10.63)} \approx 0.598$$

2）Applicability analysis

1. Euclidean distance is applicable to all numerical data, but for features with different dimensions (such as height [cm] vs weight [kg]), standardization is required before use.

2. Manhattan distance is suitable for high-dimensional data, with less influence from outliers, and is suitable for situations where features are independent and have no strong correlation.

3. Mahalanobis distance is suitable for situations where features have correlation and requires calculating the covariance matrix.

4. Cosine similarity is suitable for text data or high-dimensional sparse data, while for low dimensional numerical data (such as this question), it may not be as intuitive as Euclidean

distance.

3） **Question Analysis:**

Euclidean distance is calculated based on the absolute differences between features. When the scales of features differ significantly, certain features (e.g., weight) may dominate the distance calculation, causing other features (e.g., height) to have negligible influence. In such cases, directly using Euclidean distance is not reasonable because the units and scales of different features vary, and the distance calculation will be biased toward features with larger scales.

Improvement Methods:

To address the issue of significant differences in feature scales, the following methods can be applied:

1. Standardization:

Transform the values of each feature to a standard normal distribution with a mean of 0 and a standard deviation of 1. Formula:

$$z = \frac{x - \mu}{\sigma}$$

xx: Original feature value,
μμ: Mean of the feature,
σσ: Standard deviation of the feature.

After standardization, all features are on the same scale, preventing features with larger scales from dominating the distance calculation.

2. Normalization:

Scale the values of each feature to a fixed range (e.g., [0, 1]). Formula:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

xmin$x$min: Minimum value of the feature,

xmax$x$max: Maximum value of the feature.

After normalization, all feature values are within the same range, eliminating the influence of scale differences.

3. Mahalanobis Distance:

Mahalanobis distance takes into account the covariance between features, automatically handling differences in feature scales and correlations between features. Formula:

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

Σ: Covariance matrix of the features.

Mahalanobis distance is particularly useful when there are correlations between features, as it better reflects the true distance between samples.

Conclusion:

When feature scales differ significantly, directly using Euclidean distance is not reasonable. To address this issue:

1. Apply standardization or normalization to eliminate the influence of feature scales, or
2. Use Mahalanobis distance to handle both scale differences and correlations between features.

These methods ensure that each feature contributes appropriately to the distance calculation, avoiding bias toward features with larger scales.

4） Is Cosine Similarity Suitable for This Dataset?

Whether cosine similarity is suitable for this dataset depends on the characteristics of the data:

1. Applicability of Cosine Similarity:

   Cosine similarity is typically used for text data or high-dimensional sparse data because it focuses on the direction of vectors rather than their magnitude (i.e., scale).

   In text analysis, documents are often represented as term frequency vectors, and cosine similarity can effectively measure the similarity between documents without being affected by document length.

2. Characteristics of This Dataset:

   The samples in the dataset are numerical vectors (e.g., A=(3,5,2,7)*A*=(3,5,2,7)), and the scales of the features may differ.

   If the scales of the features vary significantly, cosine similarity may not

accurately reflect the similarity between samples because it ignores the magnitude of the vectors (i.e., the specific numerical values of the features).

3. Suitability:

   If only the directional similarity (i.e., the proportional relationship between features) is of interest, cosine similarity is suitable.

   However, if the scales of the features differ significantly, and the magnitude of the values is important for similarity, cosine similarity may not be suitable.

Improvement Methods:

If cosine similarity is not suitable for this dataset, consider the following alternatives:

1. Standardize the Data and Use Cosine Similarity:

   Standardize the data (e.g., scale each feature to the same range) before calculating cosine similarity. This eliminates the influence of scale differences.

2. Use Euclidean Distance or Other Distance Metrics:

   If the magnitude of the feature values is important for similarity, use Euclidean distance, Manhattan distance, or other distance metrics based on numerical differences.

Conclusion:

1. If only the directional similarity (proportional relationship between features) is of interest, cosine similarity is suitable.
2. If the scales of the features differ significantly, and the magnitude of the values is important for similarity, cosine similarity may not be suitable.
3. In such cases, standardize the data before using cosine similarity or choose a more appropriate distance metric (e.g., Euclidean distance).