Csci 543: Data Mining - Assignment 1

Due Date: 9:59 AM CDT, Monday, September 16, 2024

Question 1. Distance and Similarity Measures

This assignment explores various measures of distance and similarity, applying them to sets of high-dimensional data points. You will calculate Euclidean distance, cosine similarity, and discuss each measure's applicability and insights.

1.1: Data Points in Low Dimensionality

This example helps understand the basic mechanics of both Euclidean distance and Cosine similarity.

Data Points

- $P = (1, 2)$
- $Q = (2, 4)$
- $R = (1, 0)$

Calculate the Euclidean distances and cosine similarities between pairs of points P and Q, and P and R. Discuss how the dimensionality affects the calculation and interpretation of distances.

Answer

The Euclidean distance between P and Q is 2.236, and the cosine similarity between P and Q is 1. Additionally, the Euclidean distance between P and R is 2, while the cosine similarity between P and R is approximately 0.447.

Euclidean distance measures the straight-line distance between two points, making it more intuitive in low-dimensional spaces where both the magnitude and position of vectors are important for determining similarity. In contrast, cosine similarity focuses only on the angle between vectors, ignoring their magnitudes, which can make it harder to interpret in low dimensions.

For instance, even if two vectors are far apart, they could still have a high cosine similarity if they point in the same direction. In such cases, Euclidean

distance provides a clearer understanding by accounting for both direction and magnitude.

1.2: Data Points in High Dimensionality

This example will illustrate the impact of high dimensionality, showing the advantage of using Cosine similarity.

Data Points

- $S = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$
- $T = (10, 10, 10, 10, 10, 10, 10, 10, 10, 10)$
- $U = (1, 0, 1, 0, 1, 0, 1, 0, 1, 0)$

Calculate the Euclidean distances and cosine similarities between pairs of points S and T, and S and U. Discuss the implications of high dimensionality.

Answer

The Euclidean distance between S and T is 28.46, and the cosine similarity between S and T is 1. Additionally, the Euclidean distance between S and U is 2.24, while the cosine similarity between S and U is approximately 0.707.

Although S and T have very different magnitudes, they share the same orientation, which is why their cosine similarity is 1. In contrast, the Euclidean distance between S and T is 28.46, showing that the large difference in their magnitudes affects this measure, even though they point in the same direction.

This highlights a key limitation of Euclidean distance in high-dimensional spaces: it becomes less meaningful as the number of dimensions increases—a phenomenon known as the curse of dimensionality. As dimensionality increases, the space expands exponentially, causing data points that might seem close in lower dimensions to become far apart in higher dimensions. This makes it difficult to draw meaningful conclusions using Euclidean distance alone.

In high-dimensional data, comparing vectors based on their direction rather than their size often provides more useful insights. Cosine similarity measures the angle between vectors, helping determine if they point in the same or similar directions, which is often more relevant in high-dimensional spaces.

For example, when measuring the distance between S and U, which are not aligned, the Euclidean distance shows a small value, but the cosine similarity—because of the different directions—results in a value other than 1. This shows that in high-dimensional spaces, Euclidean distance can be misleading. Therefore, when working with high-dimensional data, cosine similarity, which focuses on vector direction, is a more accurate measure.

Question 2. Similarity and Correlation Measures

2.1: Calculations with Provided Data Points

Calculate the Cosine Similarity and Pearson Correlation for the following vectors:

- $X = (1, 2, 3)$
- $Y = (2, 3, 4)$
- **Cosine Similarity**:

$$\text{Cosine Similarity}(\mathbf{X}, \mathbf{Y}) = \frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\|\|\mathbf{Y}\|}$$

- **Pearson Correlation Coefficient**:

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Answer

The cosine similarity between X and Y is 0.993 and the pearson correlation coefficient is 1.

2.2: Creating Your Own Example

Provide an example where Cosine Similarity and Pearson Correlation give significantly different results, or where one measure is calculable and the other is not. Explain why these differences occur.

- Briefly describe the two vectors you choose.
- Calculate both Cosine Similarity and Pearson Correlation for your vectors, if applicable.
- Discuss why one measure might be undefined or significantly different from the other.
- Reflect on how these differences impact their applicability in real-world data analysis.

Answer

- $E = (100, 200, 300)$
- $F = (1, 2, 1)$

The cosine similarity between E and F is 0.873, while the Pearson correlation coefficient is nearly 0. This indicates that while the two vectors are aligned in a similar direction, there is no linear relationship between their values.

To create an example where Cosine Similarity and Pearson Correlation yield significantly different results, it's effective to use vectors with the same orientation but drastically different magnitudes across dimensions. This is because cosine similarity calculates the angle between the vectors (ignoring magnitude), while Pearson correlation measures how the values across dimensions correlate linearly.

Both cosine similarity and Pearson correlation range between -1 and 1. A value closer to 1 indicates that the vectors have either the same direction (cosine) or a positive linear relationship (Pearson). A value closer to -1 indicates the opposite: the vectors either point in opposite directions (cosine) or have a negative linear relationship (Pearson). A value of 0 implies no relation between them. In the case of E and F, the cosine similarity of 0.873 shows that the vectors are in a similar orientation, but the Pearson correlation being near 0 reflects that there is no linear correlation between their values.

When values follow a consistent pattern of increase or decrease, Pearson correlation tends to show a strong relationship. For E, the values increase steadily across dimensions, but for F, the values increase in the first dimension and decrease in the second dimension, leading to a lack of correlation between the two.

Cosine similarity is widely used in Natural Language Processing (NLP) to measure text similarity, where the magnitude of the vectors is irrelevant, and only the direction matters. It computes the cosine of the angle between two TF-IDF vectors, which are numerical representations of document content. On the other hand, Pearson correlation is commonly used for time-series analysis, as it effectively measures how two variables increase or decrease in relation to each other over time, making it ideal for analyzing trends or patterns in sequential data.

Question 3. Programming

Implement a decision tree classifier in Python. Your implementation should use the Gini index to measure impurity and select the median of feature values as the threshold for binary splits.

File Naming and Submission:

Please name your Python script as `Assignment_1_\[Your_Student_ID\].py`, replacing `\[Your_Student_ID\]` with your actual student ID. This naming convention is crucial for submission tracking and grading purposes.

Requirements:

1. Your decision tree should handle binary classification tasks.

2. Use the Gini index as the criterion for measuring the impurity of a node. The Gini impurity is calculated as:

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

where p_i is the proportion of class i instances among the training instances in the dataset D.

3. For each split, use the median of the feature values to determine the threshold.

4. Provide functions to fit the tree to a given dataset and to predict new instances.