Assignment 1 Problems

Advanced Data Mining: Winter 1403: Dr. Minaei
Due Sunday, Esfand 27, 1403

Somayeh rezaie
Amirhossein Namazi

Problem 1 *(10 points)*

Give an example of how a banking research center, with access to all customer records, can use classification, regression, clustering and association analysis to help administrators make better decisions. For each technique, specify the goal (business question), the input and the output. You may assume you have access to any relevant data you need.

Problem 2 *(10 points)*

a. Define association rule mining and explain about its challenges and issues in transactional data.
b. In addition to traditional association rule mining, we have Fuzzy association rule mining. Explain the difference between traditional association rule mining and fuzzy association rule mining. Provide an example to illustrate the distinction.
c. Suggest two potential applications of fuzzy association rule mining.
d. Discuss two challenges commonly encountered in fuzzy association rule mining.

Problem 3 *(10 points)*

a. List and briefly explain at least three key steps in data preprocessing. Provide a real-world example for each step.
b. Define noise in data. How does noise differ from outliers? Provide an example of each.
c. Explain the difference between simple random sampling and stratified sampling. When is stratified sampling more effective?
d. What is the curse of dimensionality? How does PCA (Principal Component Analysis) address this issue?
e. Normalize the following dataset using min-max normalization (scale to [0, 1]): [12, 25, 30, 45, 60]

Problem 4 *(10 points)*

a. Compare the multi-tiered architecture of a data warehouse with traditional OLTP system architecture. Why is OLAP considered "subject-oriented" while OLTP is "application-oriented"? Provide examples.
b. What is a data cube? Describe its components (measures, dimensions) and provide a real-world example. A retail company wants to analyze sales performance across regions and product categories over time. Which schema (star, snowflake, or fact constellation) would you recommend? Justify your choice.
c. A healthcare provider wants to predict patient readmission rates using historical data stored in a data warehouse. How would OLAP and data mining collaborate to achieve this goal?

Problem 5 *(15 points)*

Please answer the following questions. You are welcome to refer to the paper if needed to support your responses (https://arxiv.org/pdf/2301.04819).

a. How can Large Language Models (LLMs) improve the efficiency and accuracy of data labeling in training datasets, particularly in semi-supervised or weakly-supervised learning environments?
b. What are the challenges and benefits of using LLMs for data preprocessing, such as cleaning, feature extraction, and transformation, especially when dealing with unstructured or noisy data?
c. In the context of data-centric AI, how can LLMs enhance data visualization by automatically generating insights or summaries, and what are the potential limitations of relying on LLMs for this task?

Problem 6 *(15 points)*

Here, we further explore the cosine and correlation measures.

a. What is the range of values that are possible for the cosine measure?

b. If two objects have a cosine measure of 1, are they identical? Explain.

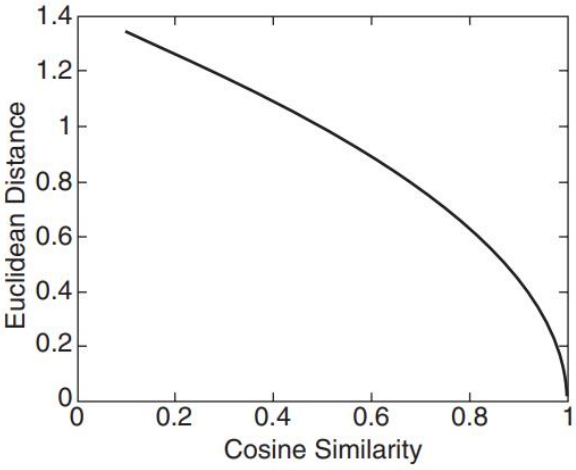
c. What is the relationship of the cosine measure to correlation, if any?
(Hint: Look at statistical measures such as mean and standard deviation in cases where cosine and correlation are the same and different.)

d. Figure 1(a) shows the relationship of the cosine measure to Euclidean distance for 100,000 randomly generated points that have been normalized to have an L2 length of 1. What general observation can you make about the relationship between Euclidean distance and cosine similarity when vectors have an L2 norm of 1?

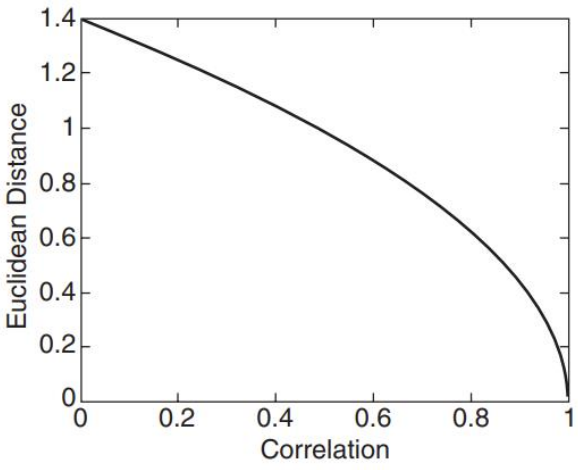
e. Figure 1(b) shows the relationship of correlation to Euclidean distance for 100,000 randomly generated points that have been standardized to have a mean of 0 and a standard deviation of 1. What general observa-tion can you make about the relationship between Euclidean distance and correlation when the vectors have been standardized to have a mean of 0 and a standard deviation of 1?

f. Derive the mathematical relationship between cosine similarity and Eu-clidean distance when each data object has an L_2 length of 1.

g. Derive the mathematical relationship between correlation and Euclidean distance when each data point has been been standardized by subtracting its mean and dividing by its standard deviation.



(a) Relationship between Euclidean distance and the cosine measure.



(b) Relationship between Euclidean distance and correlation.

Figure 1. Figures for Problem 4

Problem 7 *(10 points)*

Consider the problem of finding the K nearest neighbors of a data object. A programmer designs Algorithm 2.1 for this task.

Algorithm 2.1 Algorithm for finding K nearest neighbors.

1: **for** $i = 1$ to *number of data objects* **do**
2: Find the distances of the i^{th} object to all other objects.
3: Sort these distances in decreasing order.
(Keep track of which object is associated with each distance.)
4: **return** the objects associated with the first K distances of the sorted list
5: **end for**

a. Explain how this algorithm works.

b. Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same.

c. How would you fix this problem?

Problem 8 *(20 points)*

Python Data Exploration

a. Load the wine dataset from sklearn.datasets
   This dataset contains 4 variables. data – feature values for each wine object, feature_names – name of each feature, target – class of each wine object, and target_names – name of each class

b. Explore this dataset. Which feature has the smallest values? Which feature has the largest values? Which feature has the largest range?

c. Find the correlation matrix (correlation between all pairs of features). Which features have the strongest correlation?

d. Find the principal components on the unstandardized data. Plot the first two component, add color to the graph based on the target.

e. Standardize the dataset (subtract mean, divide by standard deviation). Find the principal components on this standardized dataset. Plot the first two component, add color to the graph based on the target.

For this exercise, your solution should include the commands that you executed in Python and the corresponding output/plot, any answers/explanation required.

Notes

- If you have any questions, feel free to ask. You can ask your questions in the Telegram group.
- Please upload your assignments as a zipped folder with all necessary components. Upload your file in HW1-ADM-YourStudentID-YourName.zip format.