IUST
Iran University of
Science and Technology

Assignment 4 Problems

Advanced Data Mining : Spring 1401 : Dr. Minaei
Due Tuesday, Tir 6, 1402

Iman Barati
Armin Tavakoli

Problem 1

Use k-means algorithm and Euclidean distance to cluster the following 8 samples (input points) into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7= (1,2), A8=(4,9).

Also consider the points A1, A3, A7 as the centers of the primary cluster (initial seeds).

a) Draw a 10 x 10 Cartesian coordinate and place the mentioned points in it.

b) Calculate and draw the distance matrix based on the Euclidean distance for the mentioned points.

c) Execute the k-means algorithm for one epoch and calculate and draw the new clusters and the centers of each cluster.

d) Repeat step c for a few more epochs until the algorithm converges; How many epochs are needed for convergence?

Problem 2

Use the Nearest Neighbor clustering algorithm and Euclidean distance to cluster the examples from the previous question: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

Calculate and draw the clusters of this algorithm as in the previous question. (Suppose that the threshold t is 4)

Problem 3

Use single and complete link agglomerative clustering to group the data described by the following distance matrix and show the dendrograms.

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Problem 4

Use single-link, complete-link, average-link agglomerative clustering as well as medoid and centroid to cluster the following 8 examples:

A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

write all the steps and draw the dendrograms diagram for each algorithm.

Problem 5

If Epsilon is 2 and minpoint is 2, what are the clusters that DBScan would discover with the following 8 examples: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

The distance matrix is the same as the one in Exercise 1. Draw the 10 by 10 space and illustrate the

discovered clusters. What if Epsilon is increased to radical 10?

Problem 6 (code)

In this section, we want to implement a number of clustering algorithms based on a number of artificial points and analyze the results. For this purpose, consider a number of artificial points created in the artificial_data.txt file as your data set. In this file, each line represents a point and its class (you don't need the class for clustering), each line of this file is x, y, class, where x and y are the coordinates of the points. Do the following in order:

a) First display the data set points and consider different colors for each class.

b) Implement the k-means clustering algorithm from **scratch** and run it on the said points for k from 2 to 10 and draw the created clusters with different colors for each k. What is the best k? Explain.

c) Choose an agglomerative clustering algorithm and implement it from **scratch** and find the best number of clusters based on the criteria explained in the class and draw the clusters for that.

d) According to the real classes, which clustering algorithm has performed better? Check and explain your method.

Bonus

Problem 7 (code)

As a data scientist in this assignment, you are tasked with performing customer segmentation on the "Mall Customers" dataset using the Gaussian Mixture Model (GMM). The goal is to identify different customer groups based on their characteristics. Customer segmentation can help businesses better understand their customers and tailor marketing strategies accordingly.

Dataset:

The "Mall Customers" dataset contains information about customers of a shopping mall, including their age, gender, annual income, and spending score. The spending score is a metric that represents how much a customer spends in the mall. By analyzing these features, we can gain insights into customer behavior and preferences.

Dataset Link: Mall Customers Dataset

Assignment Tasks:

1. Data Exploration:

- Load the "Mall Customers" dataset and perform exploratory data analysis.
- Understand the structure of the dataset, check for missing values, and analyze the distributions and relationships between different variables.

2. Data Preprocessing:

- Preprocess the dataset as needed.
- Handle missing values, convert categorical variables, and scale numerical features if necessary.

3. Gaussian Mixture Model:

- Implement the Gaussian Mixture Model (GMM) algorithm from **scratch** for customer segmentation.

- Tune the hyperparameters, such as the number of components in the mixture model, to optimize the clustering results.

4. Cluster Analysis:

- Analyze and interpret the obtained clusters.

- Examine the characteristics of each cluster and identify meaningful patterns or insights.

- Visualize the clusters using appropriate plots or charts to better understand the customer segments.

5. Marketing Strategy:

- Based on the customer segments identified, propose a marketing strategy for the shopping mall.

- Discuss how the different customer groups can be targeted with personalized marketing campaigns or promotions to maximize customer engagement and satisfaction.

Submission Guidelines:

Prepare a report documenting your findings and analysis. Include code snippets, visualizations, and interpretations. Explain the steps you took, the choices you made, and the rationale behind them. Submit the report along with the code used for the analysis.

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 HW4-ADM-YourStudentID-YourName.zip format.