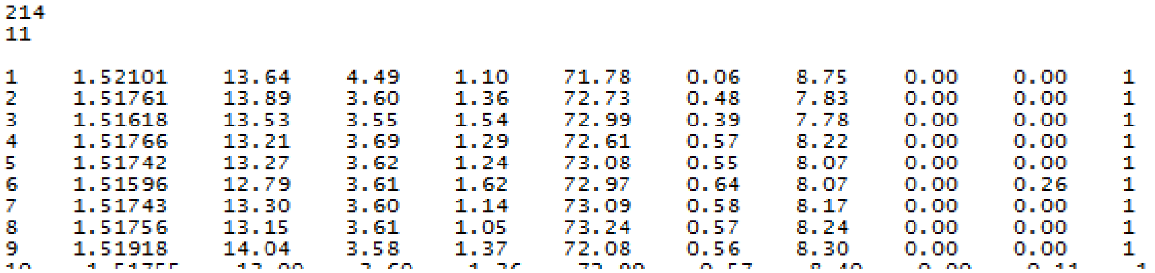
CS112 Object-Oriented Programming
Section H
Project Statement
Group Size: 2 persons maximum
Deadline: 10th May 2024

Background:

Clustering in machine learning is a technique used to group a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups. It is an unsupervised learning technique, meaning that the algorithm learns the patterns in the data without being explicitly told how to group the data. Clustering algorithms aim to partition the data into clusters such that the intra-cluster similarity is maximized and the inter-cluster similarity is minimized. Some common clustering algorithms include K-means, hierarchical clustering, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Clustering is used in various applications such as customer segmentation, anomaly detection, image segmentation, and more, where identifying natural groupings or patterns in data is useful.

Statement

In this project you have to implement a clustering techniques. Your solution must use appropriate OOP concepts. Format of the dataset should be as follows.



```
214
11

1    1.52101    13.64    4.49    1.10    71.78    0.06    8.75    0.00    0.00    1
2    1.51761    13.89    3.60    1.36    72.73    0.48    7.83    0.00    0.00    1
3    1.51618    13.53    3.55    1.54    72.99    0.39    7.78    0.00    0.00    1
4    1.51766    13.21    3.69    1.29    72.61    0.57    8.22    0.00    0.00    1
5    1.51742    13.27    3.62    1.24    73.08    0.55    8.07    0.00    0.00    1
6    1.51596    12.79    3.61    1.62    72.97    0.64    8.07    0.00    0.26    1
7    1.51743    13.30    3.60    1.14    73.09    0.58    8.17    0.00    0.00    1
8    1.51756    13.15    3.61    1.05    73.24    0.57    8.24    0.00    0.00    1
9    1.51918    14.04    3.58    1.37    72.08    0.56    8.30    0.00    0.00    1
```

- The digit 214 in first row is the number of rows
- Digit 11 in second row is the number of columns
- Third row is an empty one
- Rest is a grid of data.

Input data sets:

Download the following datasets and transform them in the above mentioned format.

- http://archive.ics.uci.edu/ml/datasets/Iris
- http://archive.ics.uci.edu/ml/datasets/Wine

Note: I may use any of the above or may be some third dataset for testing purpose during demo.

Write an application, to apply the following tasks to the input datasets.

Task 1:

Calculation Correlation Matrix:

- Create a correlation matrix from the data matrix using Pearson’s correlation coefficient
- The correlation matrix will be a NXN matrix (where N is number of records in your input dataset) containing Pearson’s correlation coefficient between each of the row in data matrix
- Pearson’s correlation coefficient formula:

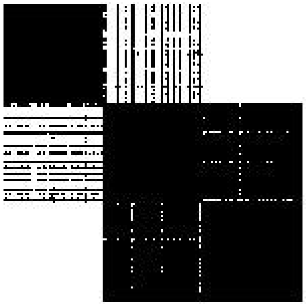
$$\sum(x-\bar{x})(y-\bar{y}) / \sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}$$

Discretize:

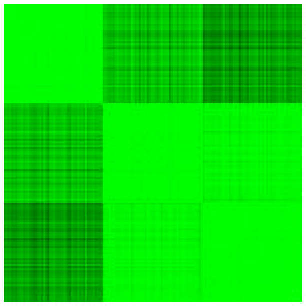
- Calculate median/mean of each column of the correlation matrix and set all the values in that column that are above the calculated median/mean to 1 and rest to 0

Visualize:

- Convert the discretized matrix into bitmap. Sample image follow.



- Provide functionality for zooming.
- Display the color coded image of similarly matrix. Follow the following steps to display color coded image
  - For each column in matrix (adjacency matrix of graph), find max value.
  - Divide each value in column by max value and multiply it with 255.
  - Resulting values will be in range 0 to 255.
  - Use this value for applying green shade to pixel.
  - Sample image follow

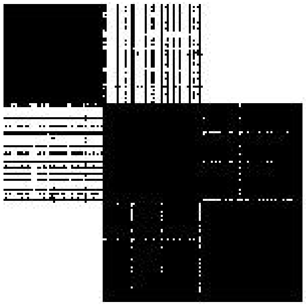


Task 2:

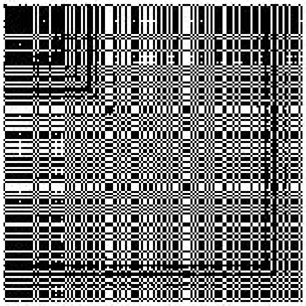
- Permute the Data Matrix
  - Do this by shuffling the individual rows in the dataset.
- Display color coded image of permuted Data Matrix

- Recover the image clusters using Signature technique. The method to generate the signature is as under
    - Sum all the values in a row
    - Calculate mean of the row
    - Multiply the Sum of the row with its Mean
    - The above three step produces a signature for a row
- Rearrange (sort) the Similarity Matrix by signature value of each row.
- Apply Task1 on the rearranged matrix
- Display the color coded image

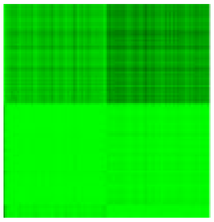
Screenshots



Bitmap of iris data Before Permutation



Bitmap of iris data After Permutation



Bitmap of iris data Correlation Matrix After Signature Generation and Arrangement

Note: Results may vary due to permutation.

Task 3:

- Also submit a write-up with following
    - Explaining each step
    - Screenshots

- Work distribution among groups members.

Grading policy

- Work completion
- Individual work, if I find a group or its member with contribution below a certain threshold ZERO credit will be awarded.

Submission Process:

- Zero credit for late or submission
- Name your submission file *<yourRoll1_ yourRoll2>_Prj [5 points will be deducted for NOT following the format]*
- Submit your project at MS teams