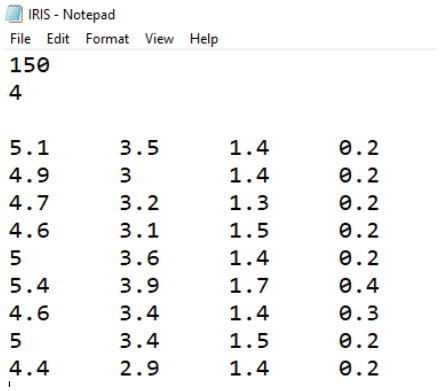
CS351-B Spring 2023
Assignment 2
Group Size 2 persons
Deadline: 17th April 2023, 11:59 p.m.

Statement

In this assignment, you have to implement a set of clustering techniques using suitable data structures. The format of the dataset should be as follows.



- The digit 150 in the first row is the number of rows
- Digit 4 in the second row is the number of columns
- The 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, in any programming language, 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:

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

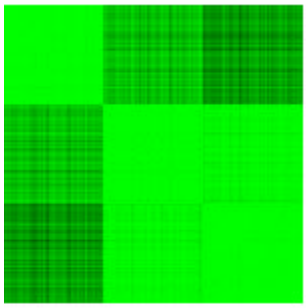
Discretize:

- 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

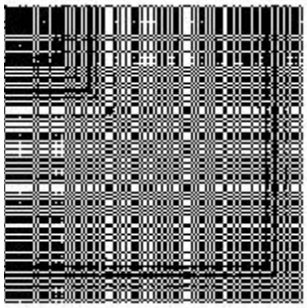


Task 2:

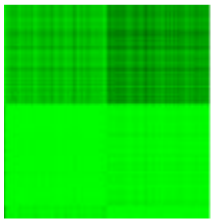
- Permute the Data Matrix
    - Do this by shuffling the individual rows in the dataset.
- Display color coded image of permuted Data Matrix
- Recover the image clusters using Signature technique. The method to generate the signature is as under
    - Sum all the values in a row
    - Calculate mean 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:

- Create a weighted graph for the permuted data set
    - Calculate correlation matrix and consider it as a graph saved in a 2D array.
    - Remove the edges having weights below certain threshold, provide input option.
    - Create a weighted graph where each node has a certain weight. The weight of the node (in this case) is the sum of weights of all the edges connected to it.
    - After that you find the node with the highest weight and get its neighbor and this becomes your one cluster
    - Then again find weights for each node and calculate the node with the highest weight.
    - The process is repeated until we are left with no clusters.

- Visualize each of the extracted cluster.

Task 4:

- Also submit a write-up with following
  - Explaining each step
  - Screenshots
  - Comparison of task 2 and 3
  - Work distribution among group members.