Instructions

- This homework assignment is worth 70 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: October 20, 2023 by 11:59 pm.**

Exercise 1

(4 points) What is "k" in the k-means algorithm? Be specific.

Exercise 2

(4 points) What is the minimum number of variables/features required to perform clustering?

(a) 0

(b) 1

(c) 2

(d) 3

(e) 4

Exercise 3

(4 points) Why is important to standardize the data before we run k-means algorithm?

Exercise 4

(4 points) Which of the following can act as possible termination conditions of the k-means algorithm?

(a) Fixed number of iterations.

(b) Assignment of observations to clusters does not change between iterations.

(c) Centroids do not change between iterations.

(d) (a) and (b)

(e) (a) and (c)

(f) (b) and (c)

(g) (a), (b) and (c)

(h) None of the above

Exercise 5

Consider the `customers.csv` datafile. This file contains information related to customers' activity on a company website. Below are the description of the variables.

- `ID:` customer ID
- `Visit_Time:` The number of visits to the company's website in a given month.
- `Average_Expense:` The average amount of money that the customer has spend.
- `Sex:` gender of the customer (0: female, 1: male).
- `Age:` age of the customer.

In Python, answer the following:

(a) (3 points) Using the `pandas` library, read the csv file and create a data-frame called `customers`.

(b) (3 points) Using the appropriate Python commands, remove the `ID` variable.

(c) (6 points) Using the appropriate standardization formula, put all the variables on the same scale. *Hint:* Notice that `Sex` is a 0-1 variable.

(d) (5 points) Cluster the data using hierarchical clustering. Use the average linkage. How many cluster do you see in the dendrogram?

(e) (6 points) Using the `KMeans` function from the `sklearn.cluster` library, cluster the customers into four clusters. Make sure you use standardized variables as the inputs in the k-means algorithms, append the cluster labels to the `customers` data-frame, and use `n_init = 20` in the `KMeans` function.

(f) (5 points) Describe each of the clusters from part (e).

Exercise 6

Imagine that you work for the marketing department of a company that sells different types of wine to customers. Your marketing team launched 32 initiatives over the past one year to increase the sales of wine (data for which is presented in the `offer_info.csv` file). Your team also acquired data that tells you which customers have responded to which of the 32 marketing initiatives recently (this data is presented in the `customer_offers.csv` file). Your marketing team now wants to begin targeting their initiatives more precisely, so they can provide offers customized to groups that tend to respond to similar offers. Your task is to use k-means clustering to discover a few groups of customers and explore what those groupings are and the types of offers that customers in those groups tend to respond to. **In Python**, answer the following:

(a) (3 points) Using the `pandas` library, read the csv file and create a data-frame called `customer_offers`.

(b) (4 points) Using the appropriate commands, report a quick summary of each of the variables in `customer_offers`.

(c) (3 points) Remove the `customer_name` variable.

(d) (5 points) Let's assume that you are planning to use all the 32 variables in your clustering analysis. Do we need to standardize the data? Explain. If so, use the min-max transformation.

(e) (6 points) Using the `KMeans` function from the `sklearn.cluster` library, cluster the data into 3 clusters (append the cluster membership to the data). Make sure you use `n_init = 20` in the `KMeans` function.

(f) (5 points) Describe each of the clusters from part (e).