Instructions

- This homework assignment is worth 73 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: December 2, 2022 by 11:59 pm.**

Exercise 1

(5 points) Why feature scaling is important for the k-means algorithm? Be specific.

Exercise 2

(5 points) How can clustering be used to improve the performance of a linear model?

(a) Creating different models for different cluster groups.

(b) Creating an input feature for cluster ids as dummy variables.

(c) Creating an input feature for cluster centroids as a continuous variable.

(d) Creating an input feature for cluster size as a continuous variable.

(e) All of the above.

(f) None of the above.

Exercise 3

(5 points) What are the risks of initial random cluster centroids assignments in k-means? Be specific.

Exercise 4

Consider the `Mall_Customers.csv` data file. This file contains the basic information (ID, age, gender, income, spending score) about a mall customers in the US. **In Python**, answer the following:

(a) (5 points) Using the pandas library, read the csv data file and create a data-frame called `customers`. Remove the observations with missing values (if there is missing values).

(b) (8 points) Using the appropriate Python commands, put `Gender`, `Age` and `Annual Income (k$)` in the same scale.

(c) (30 points) Because you are not familiar enough with buying patterns in malls, estimate the number of clusters for this dataset using the Calinski-Harabasz, Davies-Bouldin, and Silhouette scores. Do the following:

- Using `Gender`, `Age` and `Annual Income (k$)` cluster that data into clusters ($k = 2, 3, \ldots, 9, 10$). Use `n_init = 20`.
- For each clustering results, compute the Calinski-Harabasz, Davies-Bouldin, and Silhouette scores.
- Visualize the Calinski-Harabasz, Davies-Bouldin, and Silhouette scores.
- Estimate the number of clusters.

(d) (8 points) Using the results from part (c), cluster the data into that number of clusters (use `n_init = 20`).

(e) (7 points) Describe each of the clusters. Does the clustering results make sense? if not, suggest how would improve this analysis.