Yeabin Moon, Ph.D.
BUS 212A-2 Spring 2023
January 5, 2023

Homework 6

Submission Instructions

1) You have to use Juppyter Notebook
2) Click the Save button at the top of the Jupyter Notebook.
3) Select Cell → All Output → Clear. This will clear all the outputs from all cells (but will keep the content of all cells).
4) Select Cell → Run All. This will run all the cells in order, and will take several minutes.
5) Once you've rerun everything, select File → Download as → PDF via LaTeX (If you have trouble using "PDF via LaTex", you can also save the webpage as pdf. Make sure all your solutions especially the coding parts are displayed in the pdf, it's okay if the provided codes get cut off because lines are not wrapped in code cells).
6) Look at the PDF file and make sure all your solutions are there, displayed correctly. The PDF is the only thing your graders will see!
7) Submit your PDF on Latte.

Question 1. What are some of the main applications of clustering algorithms?

Question 2. Describe two techniques to select the right number of clusters when using k-means.

Question 3. In this problem, you will perform $K-$means clustering manually, with $K = 2$, on a small example with $n = 6$ observations and $p = 2$ features. The observations are as follows.

```
import pandas as pd
df = pd.DataFrame({'x1':[1,1,0,5,6,4],'x2':[4,3,4,1,2,0]})
```

1) Plot the observations.
2) Randomly assign a cluster label to each observation. Report the cluster labels for each observation.
3) Compute the centroid for each cluster.
4) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.
5) Repeat (3) and (4) until the answers obtained stop changing.
6) In your plot from (1), color the observations according to the cluster labels obtained.

Question 4. The classic Olivetti faces dataset contains 400 grayscale 64 × 64–pixel images of faces. Each image is flattened to a 1D vector of size 4,096. Forty different people were photographed (10 times each), and the usual task is to train a model that can predict which person is represented in each picture. Load the dataset using the $sklearn.datasets.fetch\_olivetti\_faces()$ function, then split it into a training set, a validation set, and a test set (note that the dataset is already scaled between 0 and 1). Since the dataset is quite small, you will probably want to use stratified sampling to ensure that there are the same number of images per person in each set. Next, cluster the images using k-means, and ensure that you have a good number of clusters (using one of the techniques discussed in this chapter). Visualize the clusters: do you see similar faces in each cluster?