Homework 5 (Due Monday 4/29)
CS534 Machine Learning, Spring 2019

This homework will explore unsupervised learning. You will implement the K-means algorithm and the gap statistic to evaluate selection of K in a $p = 32$-dimensional dataset.

Problem 1 - K-Means (20 points)

Implement the K-means algorithm from Algorithm Table 14.1 in the textbook. Generate a toy dataset in 2-dimensions and verify that your implementation is working. Display the final cluster assignments of this data in a scatter plot.

Problem 2 - Gap Statistic (20 points)

Refer to the gap statistic paper in this problem.

Implement the gap statistic and use your implementation to determine the optimal number of clusters for the dataset provided in the file *HW5.mat*. Evaluate $K = \{1, 2, \ldots, 20\}$ and use $B = 100$ reference (null) datasets. The reference datasets should be generated using the principal components method described on page 414 of the paper. In this method, the bounding rectangle is rotated to align with the data and the data is generated uniformly within this rotated rectangle (you can practice this on your 2D dataset to verify correctness of your implementation).

Plot the gap statistic as a function of K and indicate the best choice K^*. Include error bars generated from the clustering the B reference datasets for each K. Clearly indicate the values of K where $Gap(K) \geq Gap(K+1) - s_{K+1}$, and also the gap-optimal choice of K, $K^* = \text{argmin}_K\{Gap(K) \geq Gap(K+1) - s_{k+1}\}$

*Note: If you want to accelerate the K-means algorithm, you can reduce the number of comparisons between the K means and the data points using the triangle inequality.