Project 2: K-means, K-center, and KNN

The aim of this project is to practice the k-means algorithm and the k-nearest neighbor algorithm. You are strongly encouraged to use datasets related to your research. You only need to choose one dataset to work on. Six datasets are available at
`http://www.stat.psu.edu/˜jiali/stat557/material.html`.

Requirements for the project are listed below.

1. Apply the k-means algorithm to your dataset. Study how the error rates within the training set and on the testing set change with the number of prototypes.

2. Use cross-validation to choose the number of prototypes.

3. Explore whether dimension reduction can improve classification by the k-means algorithm.

4. Apply the k-nearest neighbor algorithm to your dataset. Study how the error rates within the training set and on the testing set change with k.

5. Unsupervised clustering: Ignore the class labels. Use the first two principle components of the data. Apply k-center and k-means clustering to two dimensional data. Compare the results using scatter plot. Try several different numbers of clusters.

6. Write a report. In the report, you are required to explain the contribution of each individual group member.