Data Science Laboratory 1

Lab 10 Practical Problems

The goal here is to perform your own cluster analysis on the Iris data set (see `https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html` and `https://en.wikipedia.org/wiki/Iris_flower_data_set`).

1. Reacquaint yourself with the Iris data set and import the following libraries: numpy, pandas, sklearn.cluster (specifically import KMeans) and matplotlib.pyplot.

   Talking Points:

   - What attributes are in our data set?

   - What value of k would we expect to take?

2. Perform K-means clustering for $k = 3$.

   Talking Points:

   - Try repeating this for different values of k. What do you notice?

3. Use the elbow method to determine the optimum k for the k-means clustering algorithm.

   Talking Points:

   - How does this match what you saw in Q2?

   - What would happen if you instead used the bisecting K-means algorithm?