CS 165B – Machine Learning, Summer 2020

Homework 3

Due *Thursday, September 3, 2020 11:59pm*

Write a Python3 program called `hw3.py` that creates implements the k-means algorithm, then plot the choice of k against the within-cluster variance.

1 Details

The program should be able to read in the contents of files, cluster the data using the k-means algorithm, and calculate the within-cluster sum of squares. The clusters should be initialized randomly. You are provided a skeleton code for this assignment. Avoid changing the signature of these functions in the skeleton code as they will be called to grade your code: read_data(), init_centers_random(), train_kmean(), sum_of_within_cluster_ss()

In addition to implementing the algorithm, you also need to analyze the data set `wine.txt`. Show how the within-cluster sum of squares changes when k goes from 2 to 10. You can either use matplotlib or plot elsewhere. In either case, the image of the plot needs to be submitted along with the code.

A data set named `simple.txt` is provided to help you test your code. It contains three distinct clusters, and when correctly clustered, the within-cluster sum of squares at k=3 is about 417. Note that you do *not* need to plot this data set.

The program must be able to run on CSIL and it must finish running within 1 minute. You are not allowed to use any third-party libraries or frameworks (not including the standard library) for this homework except those declared in the skeleton code.

2 Grading

Your submission will be graded on both the correct implementation of k-means, as well as the plot. The code must work on CSIL or you will receive 0%.

Grade Breakdown:

- 40% for correctly implementing k-means algorithm
- 30% for correctly calculating the within-cluster sum of squares
- 30% for plotting the within-cluster sum of squares

3 Submission

Submit *only* `hw3.py` and the plot to GauchoSpace. Submit them individually (not in a folder). Do not zip the files.