Data Science 677 Lecture 12

(April 17, 2019)

Unlabeled data set

---- use k-means to cluster your “weeks” and see the composition of the clusters

First Assignment:

1. You have a labeled data set (2017 + 2018)

Take (x1,y1), ……(x104, y104) - 2 years of data and cluster

Task 1: Use sklearn library , take 5 clusters and plot the results

(random initial condition, use the defaults….)

Task 2: for the same dataset, find out the best k k = [1,2, 3,4,5,6,7,8]. Plot the distortion vs k and find out the best k

Task 3: for this optimal k, examine your clusters:

Cluster # % red weeks % green labels

1 30 70

2 95 5

Task 4: implement k-means (assignment and update) for Euclidean, street and Minkowski distance

All k-means by Monday