Report

Tuğrul Tosun

22.12.20

1 Part 1: K-Nearest Neighbor

1.1 K-fold Cross-validation

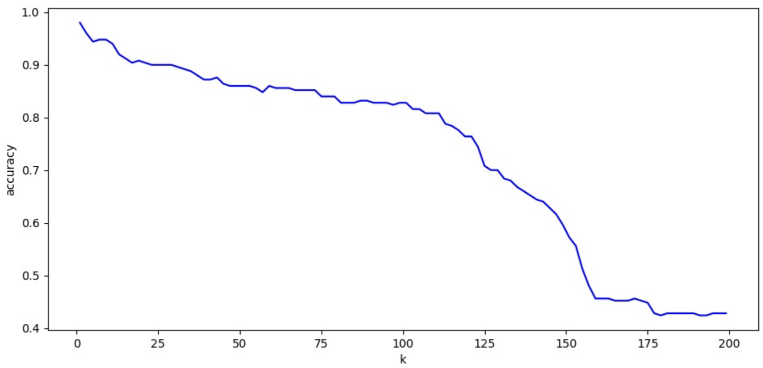


Figure 1: KNN.

1.2 Accuracy drops with very large k values

As k gets larger majority class of our train set will have impact on our testing results. When k gets larger our knn algorithm will calculate the most occurrence of neighbour for that k so when k get close to our training set size, our data will be classified as the label that occurs more in train set.

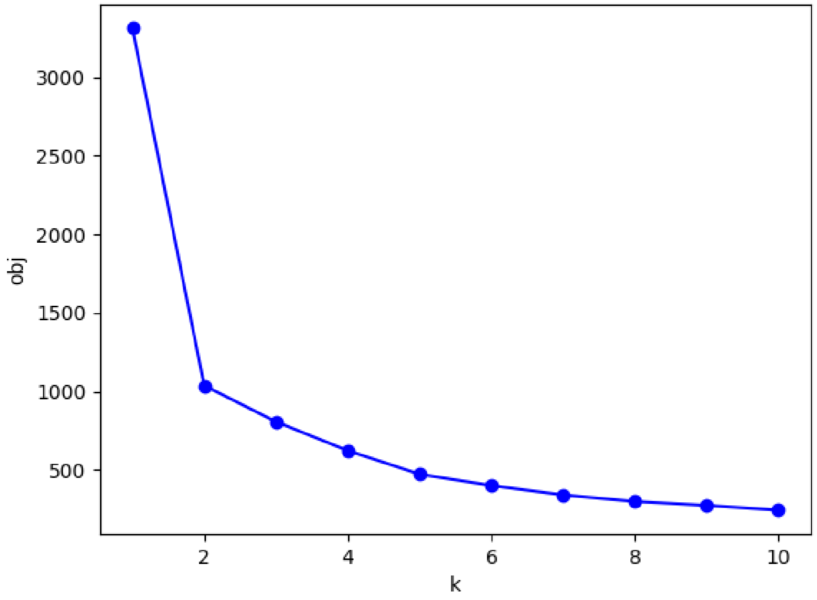
1.3 Accuracy on test set with the best k

My best k is 1 and accuracy for the test data for k=1 is 0.99(99%). I have also printed these values in my code in last few lines.

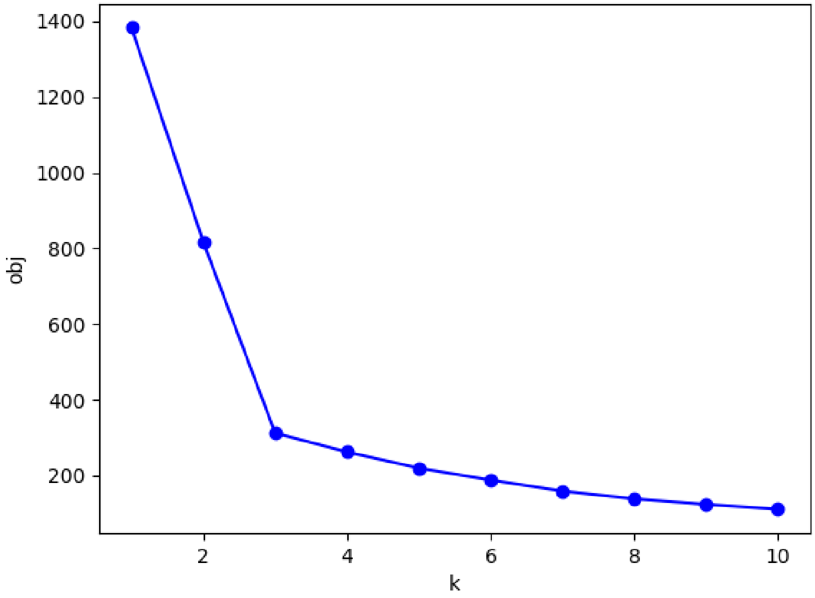
2 Part 2: K-means Clustering

2.1 Elbow method

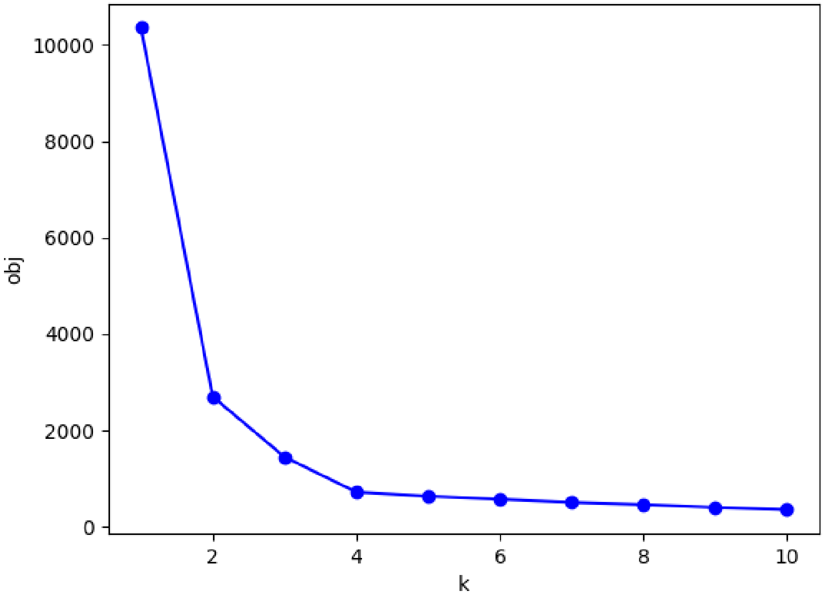
Plot k vs final objective values according to the description in the homework on each clustering data. Decide which k is suitable on each data.



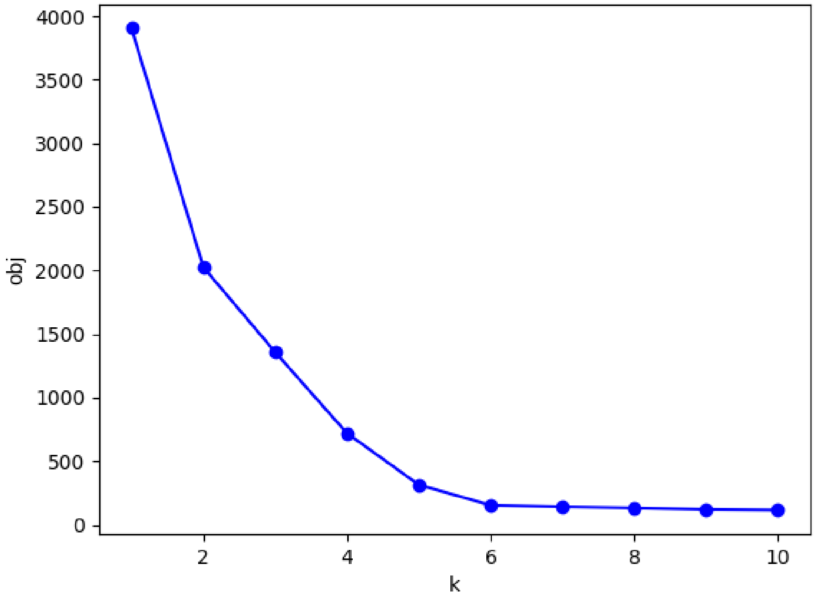
Clustering 1 data:choosing k=2 for clustering



Clustering 2 data:choosing k=3 for clustering



Clustering 3 data:choosing k=4 for clustering

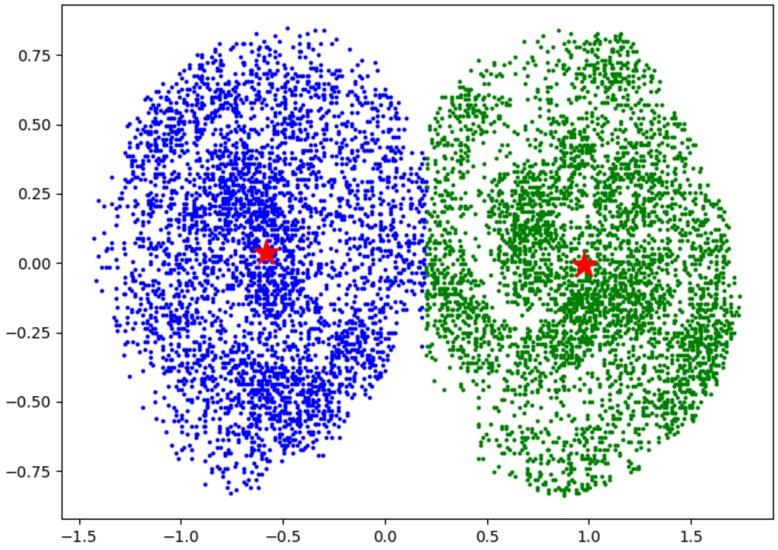


Clustering 4 data:choosing k=5 for clustering

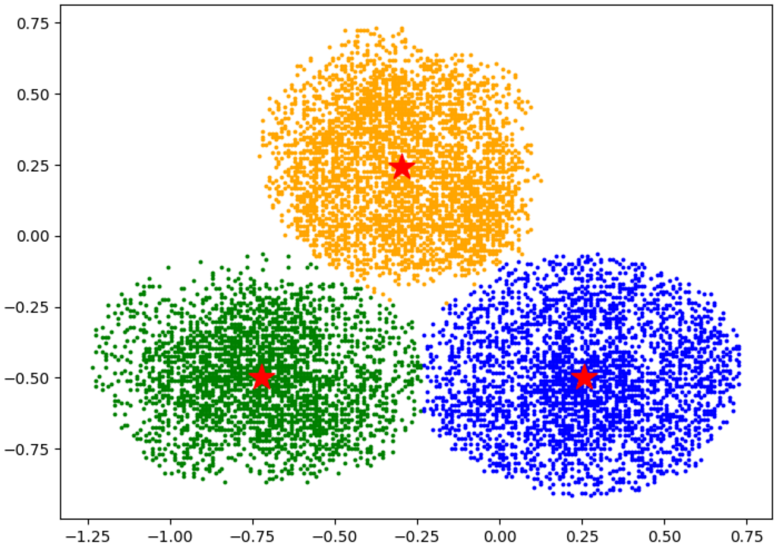
2.2 Resultant Clusters

Plot the final clusters in each data according to the description in the homework.

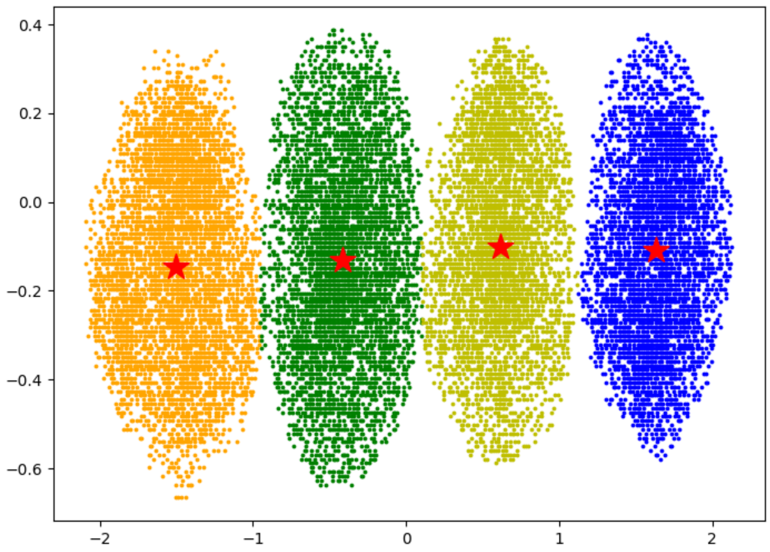
Red stars in figures represent cluster centroids.



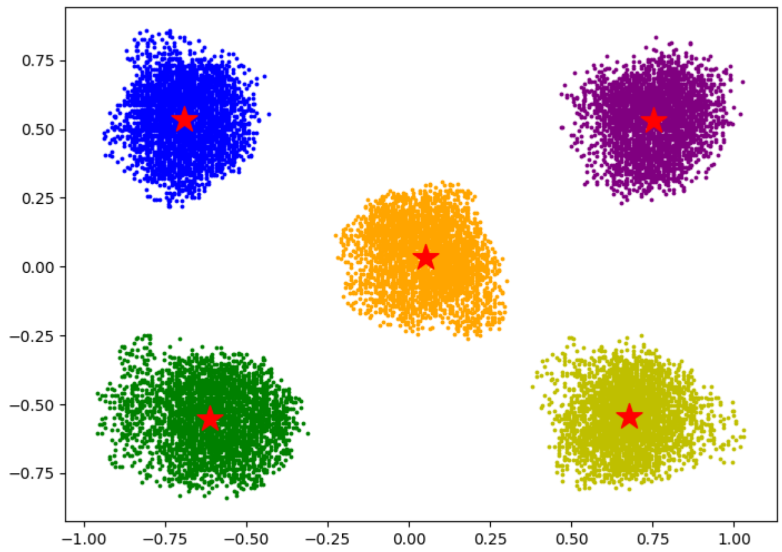
Clustering 1



Clustering 2



Clustering 3

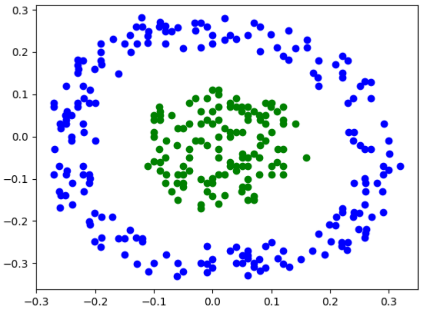


Clustering 4

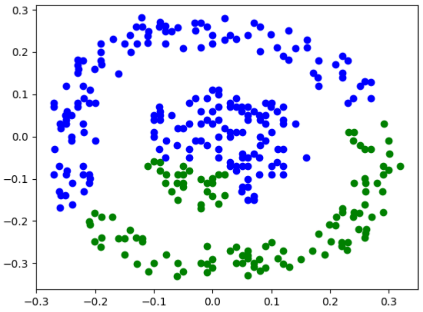
3 Part 3: Hierarchical Agglomerative Clustering

3.1 data1

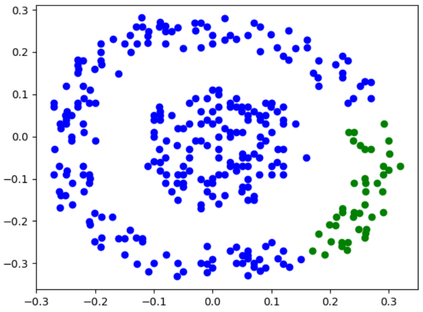
Since data clusters by formed by only min distances between clusters as you see single linkage performs better since inner and outer data members are more close to their inner and outer data clusters.



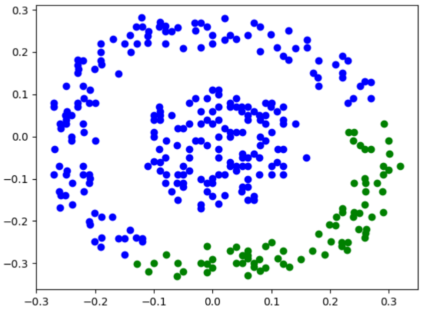
Data1: Single-Linkage



Data1: Complete-Linkage



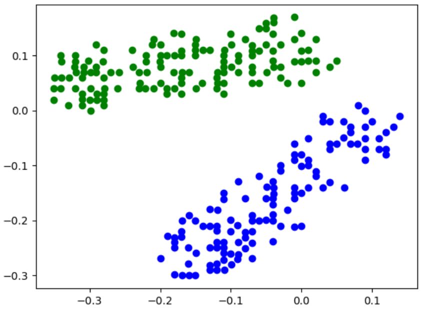
Data1: Average-Linkage



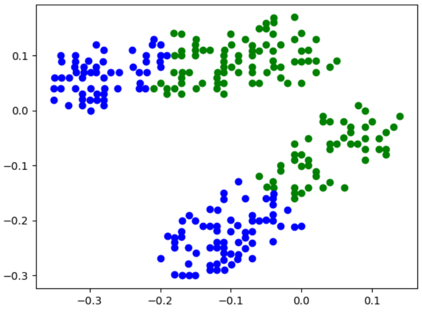
Data1: Centroid

3.2 data2

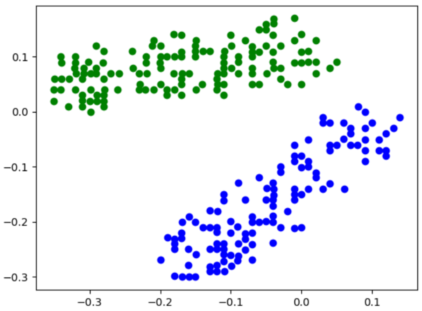
Since data again has two distant clusters single linkage performs well on this data. But this time also average link performs well which will mean that clusters were merged better when their average distance between each pair also looks like smallest distant between each cluster or so.



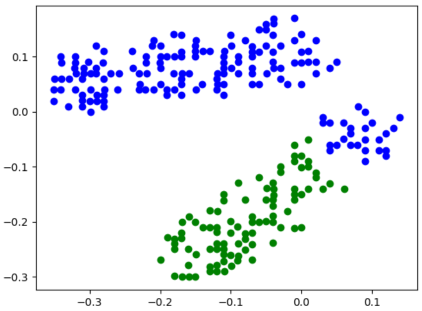
Data2: Single-Linkage



Data2: Complete-Linkage



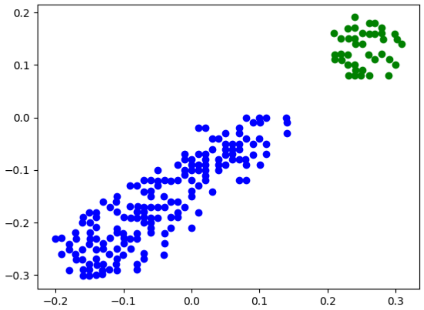
Data2: Average-Linkage



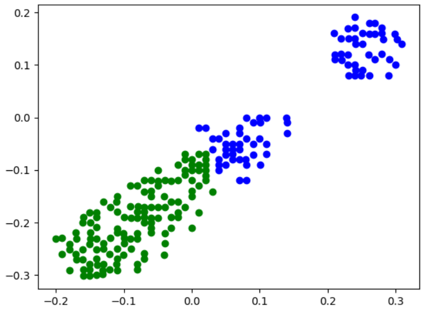
Data2: Centroid

3.3 data3

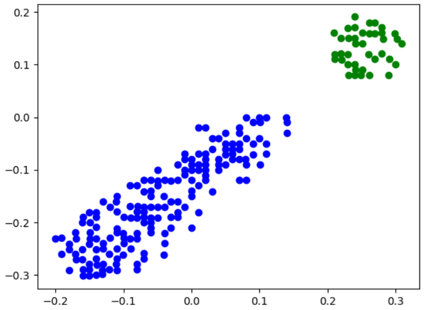
Complete linkage didn’t perform well on this data since cluster distances are taken as max distance between each clusters then middle side tend to join right side on our data graph as you see.



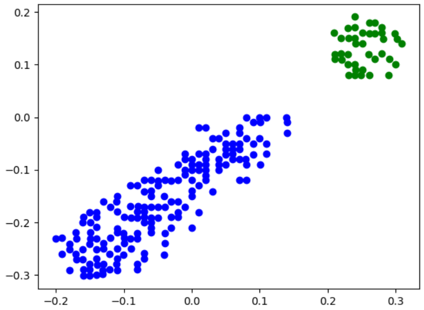
Data3: Single-Linkage



Data3: Complete-Linkage



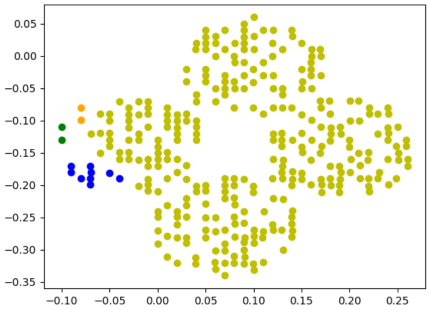
Data3: Average-Linkage



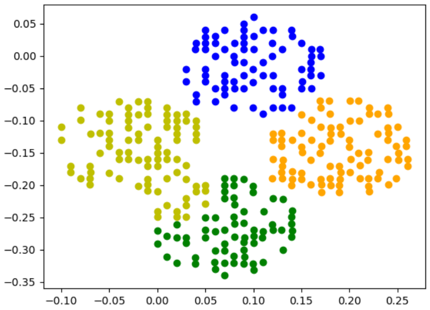
Data3: Centroid

3.4 data4

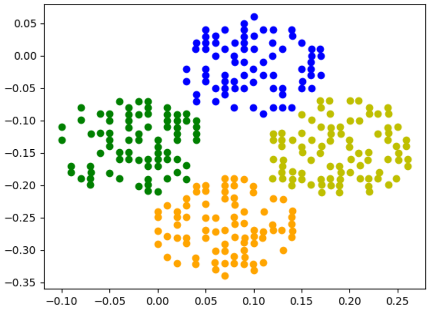
Average linkage and centroid criterion fit well on this data since each cluster seems samely distributed and have same size and shape almost, they tend to join clusters in their corresponding data places.



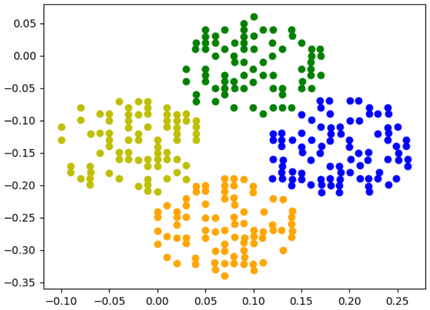
Data4: Single-Linkage



Data4: Complete-Linkage



Data4: Average-Linkage



Data4: Centroid