Lachlan Sinclair

3/1/2020

Week 8 Assignment 1

Reflections

a. For the k-means method all data points must calculate their distance to all of the "exemplars", they are then categorized by whichever one they are closest to. Once they have been categorized the exemplar then moves to the average of all values that were assigned to its cluster. These two steps are then repeated until the iterations no longer produce significant changes to the exemplars locations. This process can occur relatively quickly, and according to the lectures is mathematically provable to not require a large amount of iterations, making it suitable to use on large datasets. For k-nearest neighbors every feature vector must be compared to every other feature vector to using the distance function to determine the closest k vectors. This requires an incredibly large amount of computational effort (I believe it would be roughly n^n computations). When the textbook describes this method, it mentioned how this computational complexity makes using this method infeasible for large datasets. They did mention there are some methods of speeding it up, however in general k-nearest neighbors is not used for large datasets.

b. Since the silhouette value cares about the dissimilarity of a cluster to its neighbors along with the similarity of values inside of the cluster, it is reasonably easy to think about a potential set of feature vectors that cannot be efficiently clustered. After double checking the definition of the silhouette value, I noted that clusters with a size of 0 have a silhouette value of 0 which means overfitting a given model will not help. As far as I can tell there is a way to produce a good silhouette value with a single cluster if all the feature vectors are tightly packed. With that information I think it reasonable to assume that when there is a set of feature vectors that is evenly dispersed over a wide range and clustering attempts will yield poor results. Large clusters will be having high internal dissimilarities and small clusters will be highly like their neighbors which will negatively affect their silhouette value. To what extent the silhouette value will be impacted I cannot say, bit I cannot think of a way to produce a good one in that situation. I would have to spend some more time researching and looking at the math to be fully confident in this assumption.