Machine Learning (CS60050)
Roll Numbers: 19CS10031
Names: Gandhi Abhishek Rajesh

Mini project -3 Code: CS1

Cricket Format Clustering using Single Linkage Hierarchical Clustering Technique

---

Code file consist of multiple classes and one main function.

bucket and k_means class

This class is used for implementing and storing the k_means algorithm and final buckets(distribution). Contains a comput_Silhouette for calculating the Silhouette coefficient.

Storage and Hierarchical_Clustering class

This class is used for implementing the Hierarchical Clustering single linkage algorithm and storing final sets. The solution finding process is implemented in two ways fit function uses dp Kruskal's algorithm like solution(faster), and fit_hard uses a naive approach calculating all possible distances at every step(very slow). Contains a comput_Silhouette for calculating the Silhouette coefficient.

Jagrad_similarity class

Calculates jagrad_similary when compute function is called, it returns a list containing jagrad similarity value for all cluster.

Save_to_file class

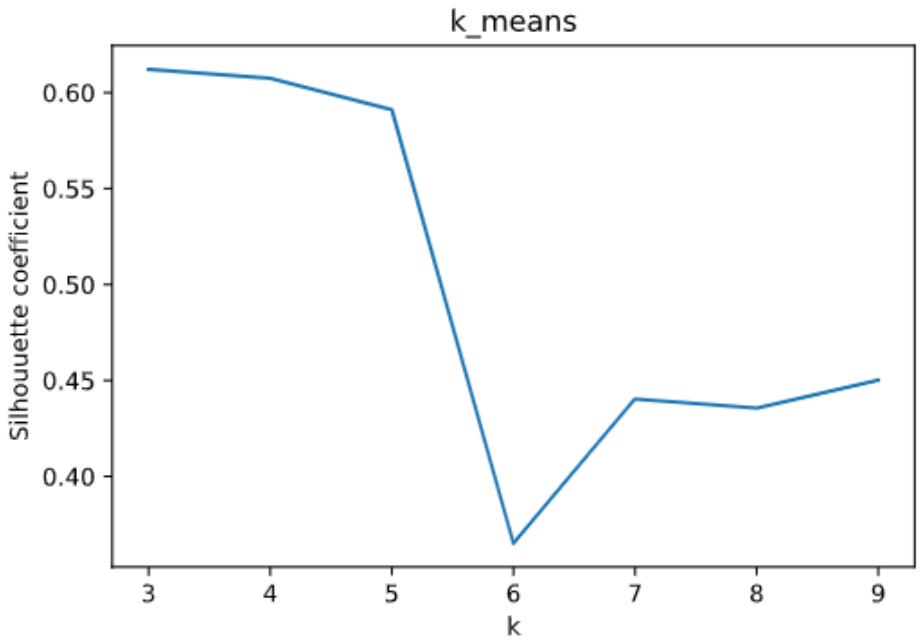
Create and save indexes in kmeans.txt and agglomerative.txt

Def main()

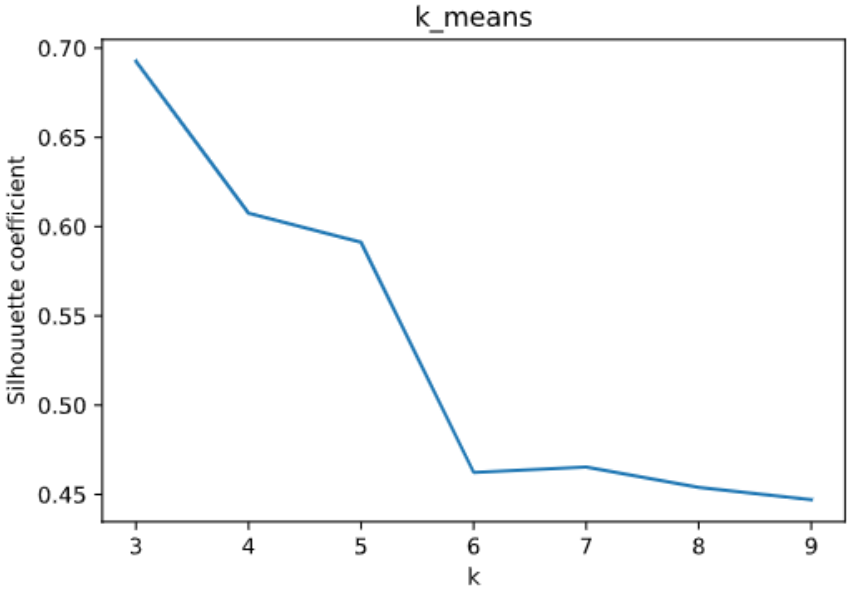
Defines and calls all the required class functions in provided order.

Silhouette coefficient vs. K Graphs

For k-means



Taking Random seed 8 (Graph 1)

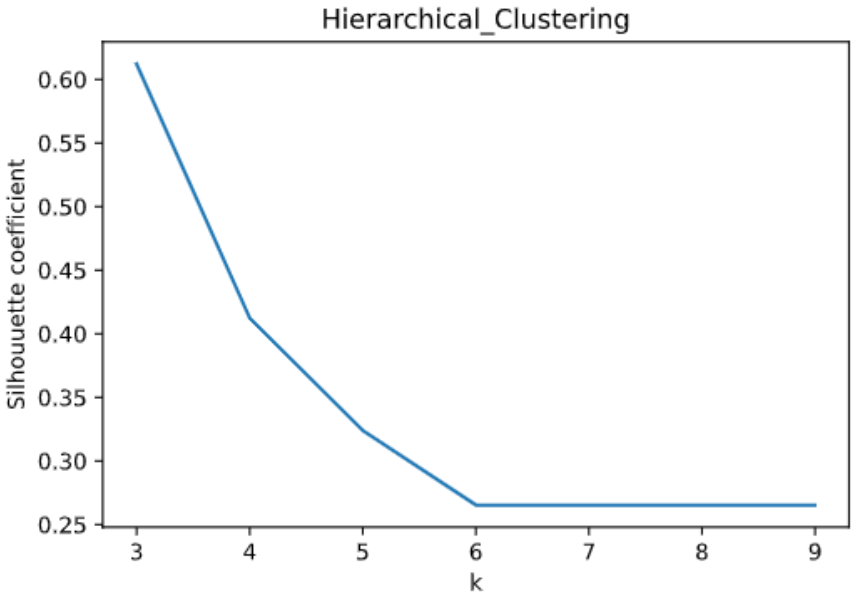


For random seed 7 (Graph 2)

From the above graphs, we can see that value best value of k is 3. After that value of the Silhouette coefficient decreases and finally stabilizes after 6.

So from these, we can conclude that our given data with the most accuracy can be visualized in 3 sets.

Hierarchical clustering



For Hierarchical clustering, we are ignoring all the sets formed which contain only one element.

We can observe a similar result maximum value at k = 3

So taking k = 3 for part 4

Critical observation:

In the CS1 dataset, there exist two completely different optimal value for k-means for some seeds; we get a higher value in the Silhouette coefficient but a lower value in jagard i.e

- Silhouette coefficient = 0.612 and jagard similarities as [1,1,1] max value in graph 1, by default i am putting seed for these value(seed 8).

After in other optimum value, we get,

- Silhouette coefficient= 0.6925 and jagard similarities as [0.992,0.367,0.362] max value in graph 2(seed 7).
- For the remaining seed, there is a little variation but very close to any one of these two.

Since jagard is a relative measure and the Silhouette coefficient is obsolete, I think the k-means result of 0.69 Silhouette coefficient is better than both the 0.61 result and the Hierarchical clustering result.

From these, we can conclude that k-mean is highly dependent on the initialisation value for smaller data sets.