Benchmarking the initialisation methods

Mehmet Fatih Gülakar

May 2019

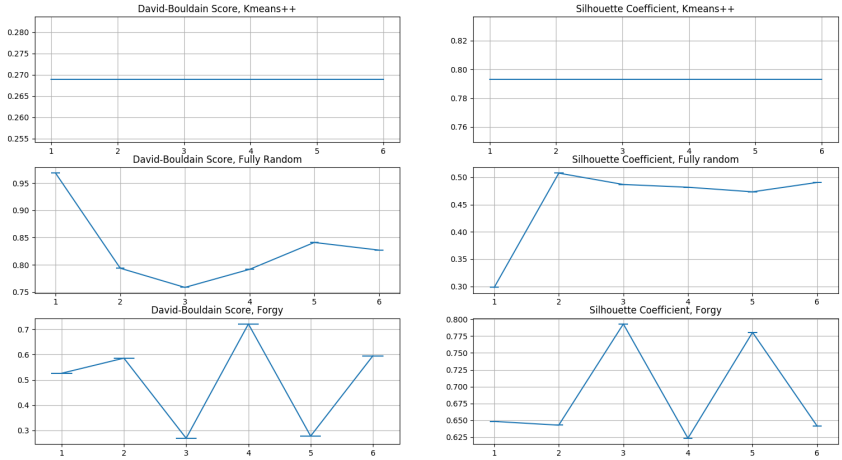
1 Setup

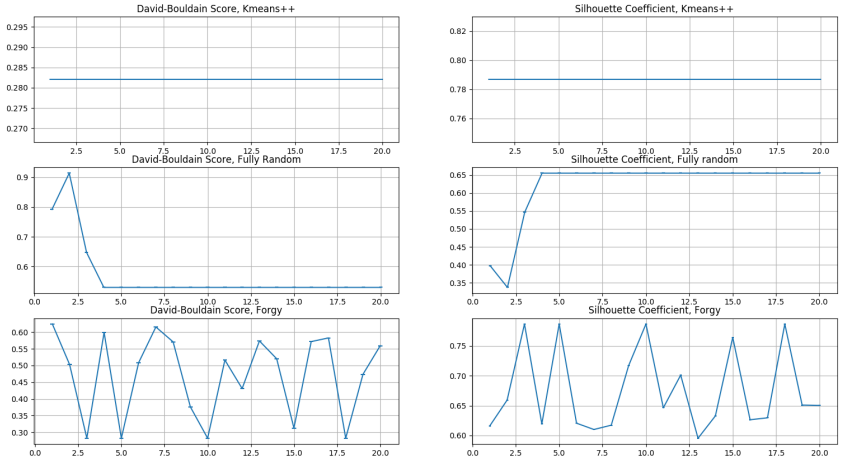
For the experiment, data set is created using make_blobs function with 150 samples and 0.85 standard deviation. k-means++, forgy and full random initialization methods are used. Random partition method could not be included unfortunately. Also, all generated data points and centroids(for the fully random method) are limited in (-8,8). Cluster number is chosen as 9 as assignment demands.

2 Results

Algorithm is run for 6 times for visualizing labels and centroids after each iteration. They can be seen below for all 3 method.

Also, David-Bouldin Score and Silhouette Coefficient are calculated after each iteration. It is done by both 6 and 20 iterations.





3 Discussion

For the first part, it is clear that initialization method determines whether algorithm will converge or not. K-means++ found the clusters correct at first iteration, while the other two method resulted in wrong clustering. Moreover, by taking into consideration the score table for 20 iteration, we can see the Forgy method could not converge. That does not mean fully random initialization is a better method than Forgy, since K-means clustering is a probabilistic algorithm, there can be cases where centroids do not converge. To sum up, what we can say about results is that K-means++ is the fastest and most accurate method, and other two method are very sensitive to probability of wrong centroid initialization.