CMPT 459: Data Mining, Spring 2024

Assignment 2

Tushrima Kelshikar 301357928

---

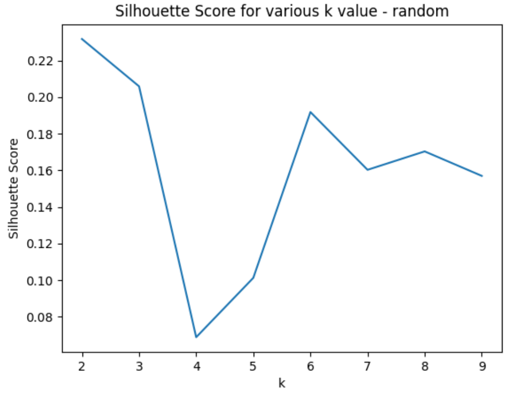
1) Implement KMeans

kmeans.py has the completed methods

- □ fit: this method takes input data **X** and performs the KMeans clustering algorithm to find **n_clusters** clusters. It returns an array of cluster labels indicating which cluster each data point belongs to.
- □ initialize_centroids: this method initializes cluster centroids using either random initialization or KMeans++ initialization. For random initialization, it randomly selects **n_clusters** data points as centroids. For KMeans++ initialization, it selects the first centroid randomly and then iteratively selects the remaining centroids by considering the distance from already selected centroids.
- □ update_centroids, euclidean_distance, silhouette: This method computes the silhouette coefficient for the clustering.It measures how similar an object is to its own cluster compared to other clusters. The silhouette coefficient is calculated for each data point and then averaged to get the overall score.

Command: python3 main.py --n-clusters 0 --data "./scRNAseq_human_pancreas.csv"

2) Random

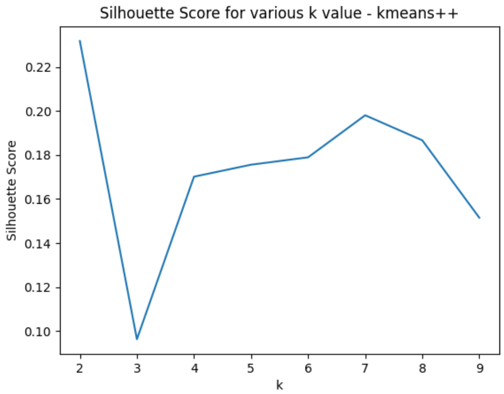


$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

By seeing the plot, k=2 is the best k.

k	**Silhouette score**
2	0.23176460528701007
3	0.20590956054070014
4	0.06880931139679504
5	0.10124764793488639
6	0.19185995776777615
7	0.16028377406584238
8	0.17035391093121305
9	0.1569706957465918

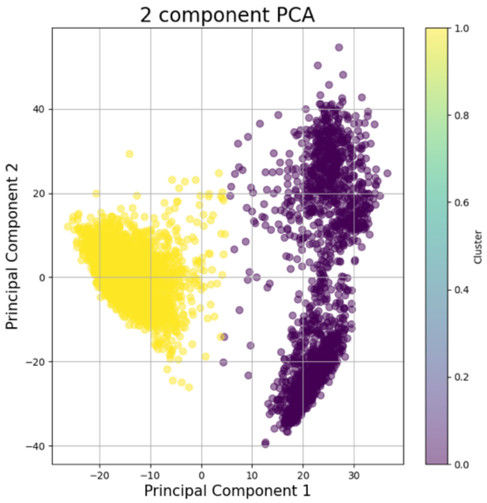
3) KMeans++



By looking at the plot, we can conclude k=2 is the best k value.

k	Silhouette score
2	0.23176460528701007
3	0.09635425096218096
4	0.17016004190953976
5	0.17560116685211102
6	0.178955168015265
7	0.19804643735056762
8	0.18668429987609392
9	0.1515185466094101

4) Best k



References

CMPT 459 Martin Ester Lecture notes

CMPT 419 Angelica Lim Assignment 2

https://towardsdatascience.com/create-your-own-k-means-clustering-algorithm-in-python-d7d4c9077670

https://stackoverflow.com/questions/1401712/how-can-the-euclidean-distance-be-calculated-with-numpy

https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb

https://medium.com/@avijit.bhattacharjee1996/implementing-k-means-clustering-from-scratch-in-python-a277c23563ac