

Unsupervised Learning: K-Means Algorithm

Lecturer: Moustapha Cisse

Group 3



April 6, 2023

Group Members

- 1 Honorine Gnonfin
- 2 Binta Sow
- 3 Uriel Nguetack

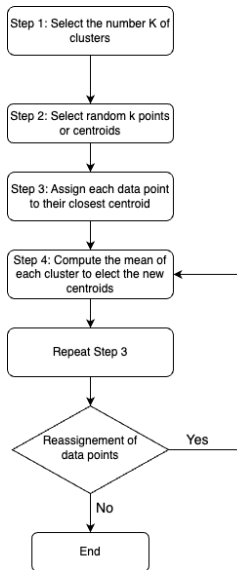
- 1 Introduction
- 2 How does K-Means works?
- 3 Methods to choose the optimal number of clusters
- 4 Pros and cons of K-Means
- 5 Implementation
- 6 Conclusion
- 7 References

Introduction

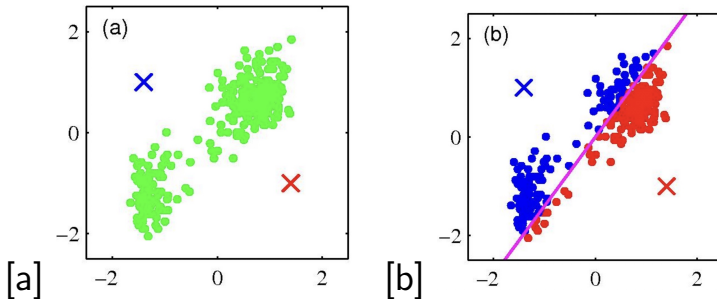
Clustering is a technique that group similar objects such that the objects in the same group are more similar to each other.

- Direct partitioning: we seek to partition the observations into a prespecified K number of clusters.
- Hierarchical clustering: we do not know in advance how many clusters we want.

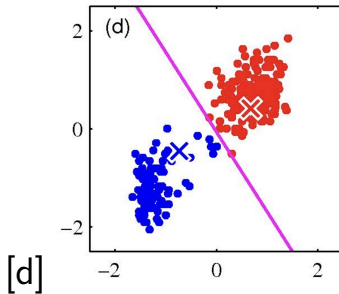
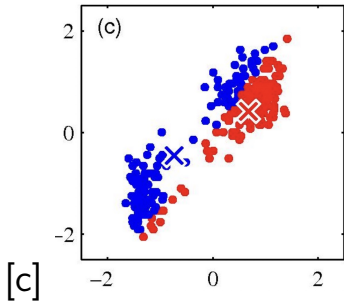
How does K-Means works?



How K-Means works?



How does K-Means works?



Methods to choose the optimal number of clusters

- Elbow Method
- Silhouette analysis

The Elbow Method: Minimization of WCSS

The objective function is given by:

$$\operatorname{argmin}_S \sum_{i=1}^k \sum_{x \in S_i} ||x - \mu_i||^2 \quad (1)$$

where $S = \{S_1, S_2, \dots, S_k\}$,
 μ_i is the centroid in the cluster i

Choose the number of clusters

The Silhouette Analysis

- We compute the silhouette coefficient for each data point.

$$S(i) = \frac{b(i) - a(i)}{\max \{b(i), a(i)\}} \quad (2)$$

where $a(i)$ is the average distance between i and all other points in the same cluster as i .

$b(i)$ is the average distance from i to all clusters to which i does not belong

- For each value of k , we compute:

$$\text{Average}(\text{silhouette})_k = \text{mean} \{S(i)\} \quad (3)$$

Pros and cons of K-Means Algorithm

Pros

- Linear time complexity and can be used with large datasets conveniently.
- Easy to implement
- Easy to interpret the clustering results

Cons





- Results will differ based on random centroid initialization.
- Sensitive to outliers
- Assume each cluster has roughly equal number of observations

IMPLEMENTATION

Conclusion

- Powerful and widely-used clustering algorithm that can be used to group data into similar clusters.
- It works by iteratively optimizing the placement of k centroids that represent the center of each cluster.
- Sensitivity to initial cluster centroids and its assumption that clusters have a spherical shape.
- Customer segmentation
- Understand what the visitors of a website are trying to accomplish
- Pattern recognition
- Data compression

References

-  Effect of Distance Metrics in Determining K-Value in K- Means Clustering Using Elbow and Silhouette Method. Autor: Danny Matthew SAPUTRA¹, Daniel SAPUTRA² and, Liniyanti D. OSWARI³ Year: 2019
-  CS229 Lecture Notes. Autor: Andrew Ng
-  <https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/>
-  <https://www.kdnuggets.com/2020/06/centroid-initialization-k-means-clustering.html>

References



<https://stackoverflow.com/questions/61462501/turning-a-matrix-into-a-vector>



<https://www.analyticsvidhya.com/blog/2021/05/guide-for-loss-function-in-tensorflow/>



https://en.wikipedia.org/wiki/K-means_clustering



[https://en.wikipedia.org/wiki/Silhouette_\(clustering\)](https://en.wikipedia.org/wiki/Silhouette_(clustering))

THANKS

FOR

LISTENING