Understanding different Clustering Techniques K-Means, kernel K-means, K-Medians, K-Medoids and their similarities, differences, and their usages

Author: Sujan Das(sujan2@illinois.edu)

This technical review will explain the popular clustering methods (Kmeans, Kmeans++, Kmedian, and Kmedoid), which is used in unsupervised learning for grouping similar type of elements together and building model based on that which provides the optimal result.

K-Means Clustering is one of the most effective algorithms in data science, which is used for unsupervised learning where data points are scattered and not defined in a group or categories (i.e., these are unlabeled data). This algorithm aims to group similar items in the form of clusters/groups where the variable K represents a number of groups/clusters.

Working principle of Kmeans algorithm Steps:

1. Choose the number of k clusters
2. Initializing random k data points for the centroids
3. Assign each data point to the closest centroid. That forms k clusters
4. Compute and place the new centroid of each cluster
5. Reassign each data point to the new closest centroid. If any re-assignment took place, then go to point 4 and execute; otherwise, your model is ready.

How to determine the optimal number of clusters K:

- Elbow Method
- Silhouette Method

Note: When the value of WCSS (Within Cluster Sum of Square) goes down, then the number of clusters increases, which does the grouping. We need to determine the optimal point where the WCSS value does not start dropping drastically. WCSS will be 0 when the number of clusters is equal to the number of points.

Advantages of K-means:

Kmeans clustering algorithm is very easy to implement. This algorithm
Can handle large datasets and provide faster performance. It also adopts new data points and assigns them to the correct group continuously.
It also generalizes clusters of different sizes and shapes.

Disadvantages of K-means:

Kmeans is very sensitive to outliers.
Selecting the value of K is not an easy job as we must find out the optimal value of K using the Elbow or Silhouette method.
If the number of dimensions increases, its scalability decreases.

Usage of Kmeans:

Here is the list of use cases for means

- Document Classification
- Delivery Store Optimization
- Identifying Crime Localities
- Customer Segmentation
- Fantasy League Stat Analysis
- Insurance Fraud Detection
- Rideshare Data Analysis
- Cyber-Profiling Criminals
- Call Record Detail Analysis
- Automatic Clustering of It Alerts

For an example, let me explain the usage for Customer Segmentation using Kmeans:
Clustering helps marketers improve their customer base, work on target areas, and segment customers based on purchase history, interests, or activity monitoring. The concept of segmentation relies on the high probability of persons grouped into segments based on common demands and behaviors to have a similar response to marketing strategies which can be done using the Kmeans clustering implementation. The result of this is to better understand its customer's needs and find out their relationship with the subscribers and improve customer satisfaction. These results enable the characterization of expenditure patterns for services that are continuously growing. This analysis model is very powerful for a large customer base.

K-Means++: K-Means++ is a smart centroid initialization technique that combats the problem of random initialization trap that occurred in Kmeans. Kmeans++ provides true clusters in results instead of non-desirable clusters.

Limitation of Kmeans: Kmeans cannot handle non-numerical(categorical) data.
Mapping categorical values to 1/0 cannot generate quality clusters for high-dimensional data. K-modes are used for those situations.

Kernel k-means:

Kernel k-means is an extension of the k-means algorithm that identifies nonlinearly separable clusters. To overcome the cluster initialization problem global kernel k-means algorithm provides a deterministic and incremental approach to kernel-based clustering.
It provides a way of adding clusters at each stage through the global search of several executions of kernel kmeans from initialization which does not depend on cluster initialization. It identifies nonlinearly separable clusters due to its search procedure and incremental nature of execution. To reduce the computational cost, it is going through the modifications, and we employ kernel k-means for MRI segmentation along with a novel kernel. This algorithm follows the same path

except for the calculation of distance; the kernel method is used instead of the Euclidean distance.

The main advantage of this algorithm is to able to identify non-linear structures and is suitable for real-life data set.
The disadvantage of this algorithm is we need to predefine the number of cluster centers. Also, this algorithm is very complex, and it has large time complexity.

Kmedians is a clustering algorithm. It is a variation of kmeans clustering where instead of calculating the mean for each cluster to determine its centroid, it calculates the median.

KMedoids is a clustering algorithm resembling the K-Means clustering technique used for unsupervised learning. The way it selects the cluster centroid differs from the K-Means algorithm. Kmeans calculates the mean of a cluster's data points and determines the center (that may not be the data point), while the KMedoid picks the actual data points from the clusters as their centers which are known as 'medoids' ('exemplars'). K-Medoids also differs from the K-Medians algorithm, which is the same as K-means, except that it chooses the medians (instead of means) of the clusters as centers.

Comparison between Kmeans, Kmedian, and Kmedoid:
k-means minimizes the distance between points within-cluster, which can be calculated using squared Euclidean distances. The arithmetic mean measure this. Squared deviations from the mean do not optimize distances'-medians minimize absolute deviations, which equals the Manhattan distance. The per-axis median should do this. K-medoids are more robust to outliers, thank-means, as it is considering more of a median-type approach to measuring the data. Here is the decision tree:

- Use Kmeans if the distance is squared Euclidean distance
- Use Kmeans if the distance is Taxicab metric
- For other distance Kmedoids can be used

To conclude, Kmeans clustering is very useful when we need to group or categorize similar data types of data from a huge dataset of unlabeled data for unsupervised learning. It also helps new data to be easily assigned to the correct group continuously that enable model to perform well and provide optimal result.

References:

1. https://en.wikipedia.org/wiki/K-means_clustering
2. https://dzone.com/articles/10-interesting-use-cases-for-the-k-means-algorithm
3. https://www.analyticsvidhya.com/blog/2020/10/a-simple-explanation-of-k-means-clustering/
4. https://stackoverflow.com/questions/21619794/what-makes-the-distance-measure-in-k-medoid-better-than-k-means
5. https://machinelearningjourney.com/index.php/2020/02/07/k-means-k-medians/
6. https://www.youtube.com/watch?v=WQB7PI6cRow&t=8s&ab_channel=PriyankaSharma

7. https://analyticsindiamag.com/comprehensive-guide-to-k-medoids-clustering-algorithm/