MACHINE LEARNING – WORKSHEET (CLUSTERING) -2

13. Is K sensitive to outliers?

The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. Mean is greatly influenced by the outlier and thus cannot represent the correct cluster center, while medoid is robust to the outlier and correctly represents the cluster center. The k-means algorithm updates the cluster centers by taking the average of all the data points that are closer to each cluster center. When all the points are packed nicely together, the average makes sense. However, when you have outliers, this can affect the average calculation of the whole cluster. As a result, this will push your cluster center closer to the outlier.

14. Why is K means better?

K-mean is better because it has the following advantages:

* Relatively simple to implement.
* Scales to large data sets**.**
* Guarantees convergence.
* Can warm-start the positions of centroids.
* Easily adapts to new examples.
* Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

15. Is K means a deterministic algorithm?

K-means is a partitioning-based clustering algorithm. *k*-means method for clustering is an iterative process in which an initial partition of given *k* clusters is then improved by applying a search algorithm to the data. Simplifying, given a pre-defined number (*k*) of clusters, the algorithm:

|  |  |
| --- | --- |
| - | begins with an initial set of *k* cluster centers (*i.e.* the centroids) |
| - | (re)assigns objects to the closest centroids |

|  |  |
| --- | --- |
| - | Recalculates centroids according to new memberships of the data points. |
|  | repeats the last two steps until a consistent result is found or until the maximum number of iterations is reached. |

The k-means clustering is based on a non-deterministic algorithm. This means that running the algorithm several times on the same data, could give different results.