MACHINE LEARNING

1. Movie Recommendation systems are an example of:

i) Classification ii) Clustering iii) Regression Options:

Answer – (a)

2. Sentiment Analysis is an example of:

i) Regression ii) Classification iii) Clustering iv) Reinforcement Options:

Answer – (d)

3. Can decision trees be used for performing clustering?

Answer - (a)

4. Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:

i) Capping and flooring of variables ii) Removal of outliers Options:

Answer – (a)

5. What is the minimum no. of variables/ features required to perform clustering?

Answer – (b)

6. For two runs of K-Mean clustering is it expected to get same clustering results?

Answer – (b)

7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?

Answer – (a)

8. Which of the following can act as possible termination conditions in K-Means?

i) For a fixed number of iterations. ii) Assignment of observations to clusters does not change between iterations. Except for cases witha bad local minimum. iii) Centroids do not change between successive iterations. iv) Terminate when RSS falls below a threshold.

Answer – (d)

9. Which of the following algorithms is most sensitive to outliers?

Answer – (a)

10. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):

i) Creating different models for different cluster groups. ii) Creating an input feature for cluster ids as an ordinal variable. iii) Creating an input feature for cluster centroids as a continuous variable. iv) Creating an input feature for cluster size as a continuous variable.

Answer – (d)

11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?

Answer – (d)

12. Is K sensitive to outliers?

Answer – Yes, **The K-means clustering algorithm is sensitive to outliers**, because a mean is easily influenced by extreme values. The group of points in the right form a cluster, while the rightmost point is an outlier

K-Means clustering is an unsupervised learning algorithm which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centroid. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs.

But sometime K-Means algorithm does not give best results. It is sensitive to outliers. An outlier is a point which is different from the rest of data points. Let us look at one method for finding outliers of univariate data (one dimensional).

The lower quartile 'Q1' is median of first half of data. The upper quartile 'Q3' is median of second half of data. The interquartile range 'IQR' is difference of Q3 and Q1. An outlier is a point that is greater than (Q3 + 1.5*IQR) or lesser than (Q1–1.5*IQR). The given below code can be used to find the outliers.

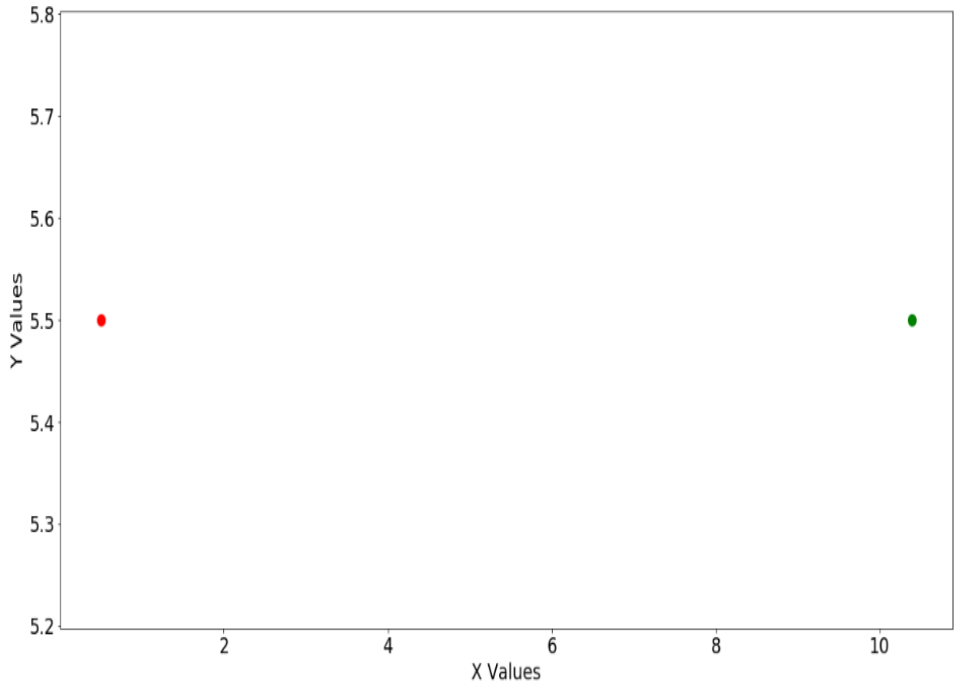
```
Q1 = np.percentile(data, 25, interpolation = 'midpoint') # The lower quartile Q1 is calculated.Q3 = np.percentile(data, 75, interpolation = 'midpoint') # The upper quartile Q3 is calculated.IQR = Q3 — Q1 # The Interquartile range is calculated.Q3 + 1.5*IQR, Q1–1.5*IQR # The outlier range is calculated.
```

We shall discuss the methods for finding outliers of multivariate data in another article.

Let us take an example to understand how outliers affect the mean of data using python.

```
X = list(np.random.rand(100)) # 'X' is a list of 100 random numbers between 0 and 1.
Y = list(np.linspace(1,10,100)) # 'Y' is a list of 100 random numbers equally spaced between 1 and 10.plt.figure(figsize=(20,10)) # Size of figure is adjusted.
plt.xticks(fontsize=20) # Size of number labels on x-axis is adjusted.
plt.yticks(fontsize=20) # Size of number labels on y-axis is adjusted.
plt.xlabel('X Values',fontsize=20) # x-axis is labelled.
plt.ylabel('Y Values',fontsize=20) # y-axis is labelled.mean_X = sum(X)/len(X) # 'mean_X' is the mean value of 'X'.
```

```
mean_Y = sum(Y)/len(Y) # 'mean_Y' is the mean value of 'Y'.
plt.plot(mean_X,mean_Y,'ro',markersize = 10) # The mean value (mean_X,mean_Y)
point is plotted.outlier = 1000 # An outlier of value 1000.
X.append(outlier) # The outlier is added to 'X'.
Y.append(Y[99] + Y[1] — Y[0]) # An extra number is added to 'Y' such equal
spacing still holds.mean_X_new = sum(X)/len(X) # 'mean_X_new' is new mean
value of 'X'.
mean_Y_new = sum(Z)/len(Z) # 'mean_Y_new' is new mean value of 'Y'.
plt.plot(mean_X_new,mean_Y_new,'go',markersize = 10) # The mean value
(mean_X,mean_Y) point is plotted in green.
```

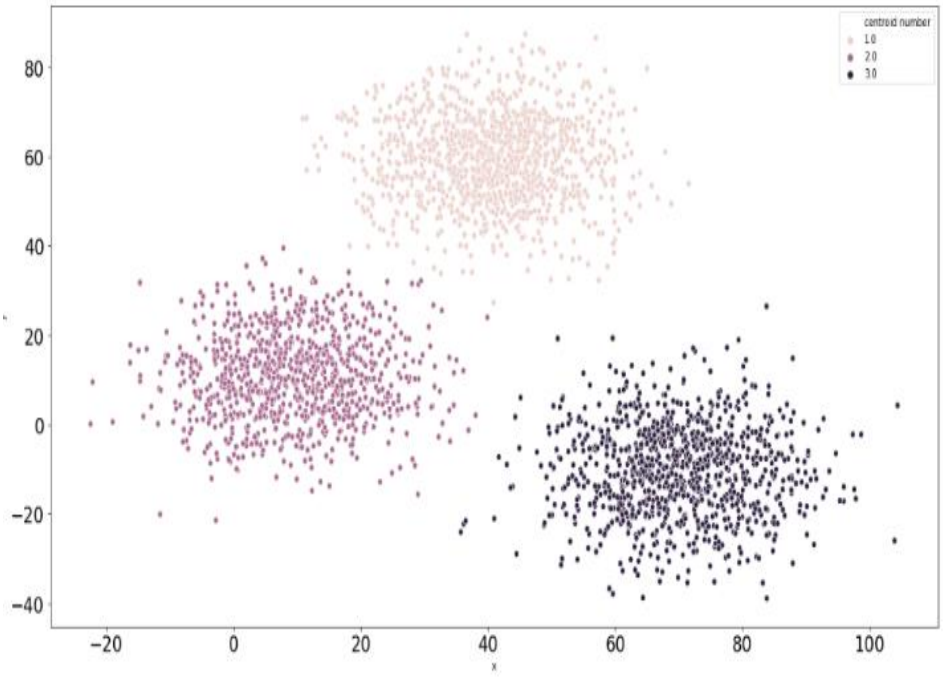


The red point is mean of data excluding outlier. The green point is mean of data including outlier.

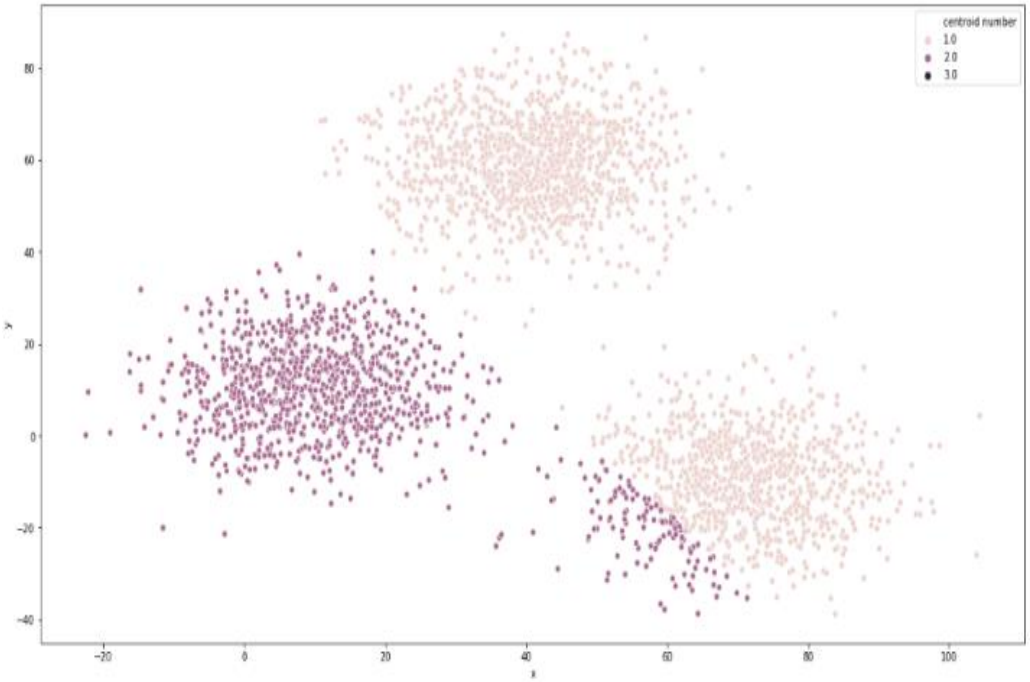
We observe that the outlier increases the mean of data by about 10 units. This is a significant increase considering the fact that all data points range from 0 to 1. This shows that the mean is influenced by outliers.

Since K-Means algorithm is about finding mean of clusters, the algorithm is influenced by outliers. Let us take an example to understand how outliers affect the K-Means algorithm using python.

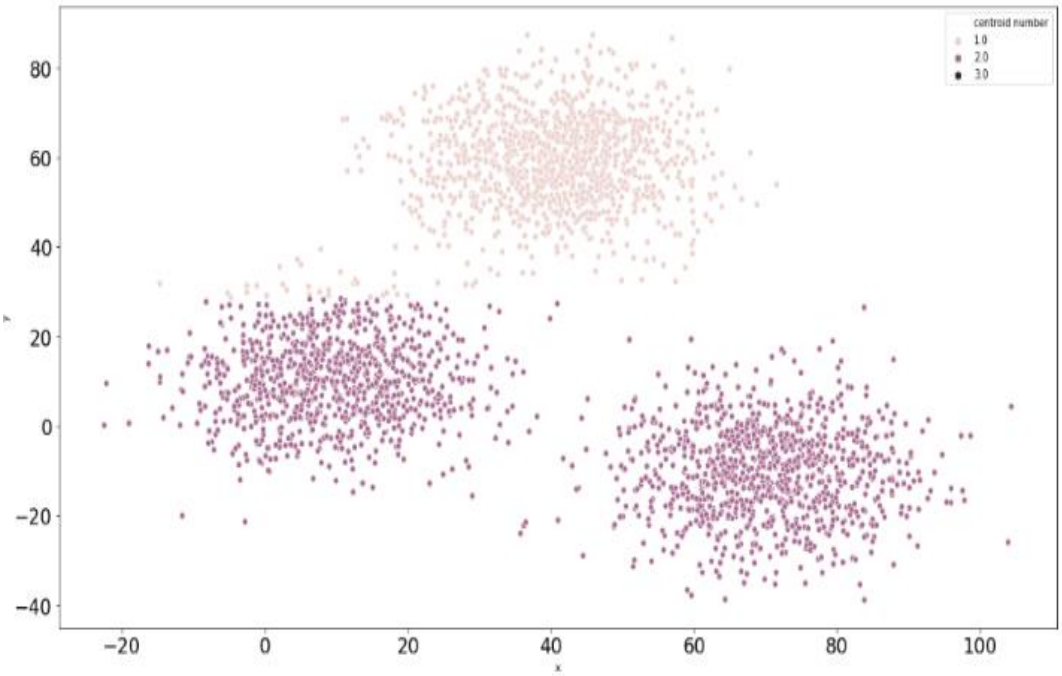
We have a 2 dimensional data set called 'cluster' consisting of 3000 points with no outliers. We get the following scatter plot after K-means algorithm is applied.



Now we add 60 outliers to ‘cluster’ data set. The outliers is about 2 percent of non-outliers. We get the following scatter plots for different values of outliers after K-means algorithm is applied.



The outliers are not shown in the scatter plot. Only the 3000 non outlier points is shown in the scatter plot for sake of better visualisation. The outliers form a seperate cluster represented by centroid number = 3.



The outliers are not shown in the scatter plot. Only the 3000 non outlier points is shown in the scatter plot for sake of better visualisation. The outliers form a seperate cluster represented by centroid number = 3.

We observe that the outliers show up as a separate cluster and also cause other clusters to merge which suggests clustering was not efficient when outliers were included in data set.

Even though the outliers were about 2 percent of non-outliers which is common in real world data sets, they had a significant impact on clustering. Hence it is better to identify and remove outliers before applying K-means clustering algorithm. We would be looking at ways of identifying and removing outliers from datasets in subsequent articles.

13. Why is K means better?

Answer - K-means clustering is a very famous and powerful unsupervised machine learning algorithm. It is used to solve many complex unsupervised machine learning problems. Before we start let's take a look at the points which we are going to understand.

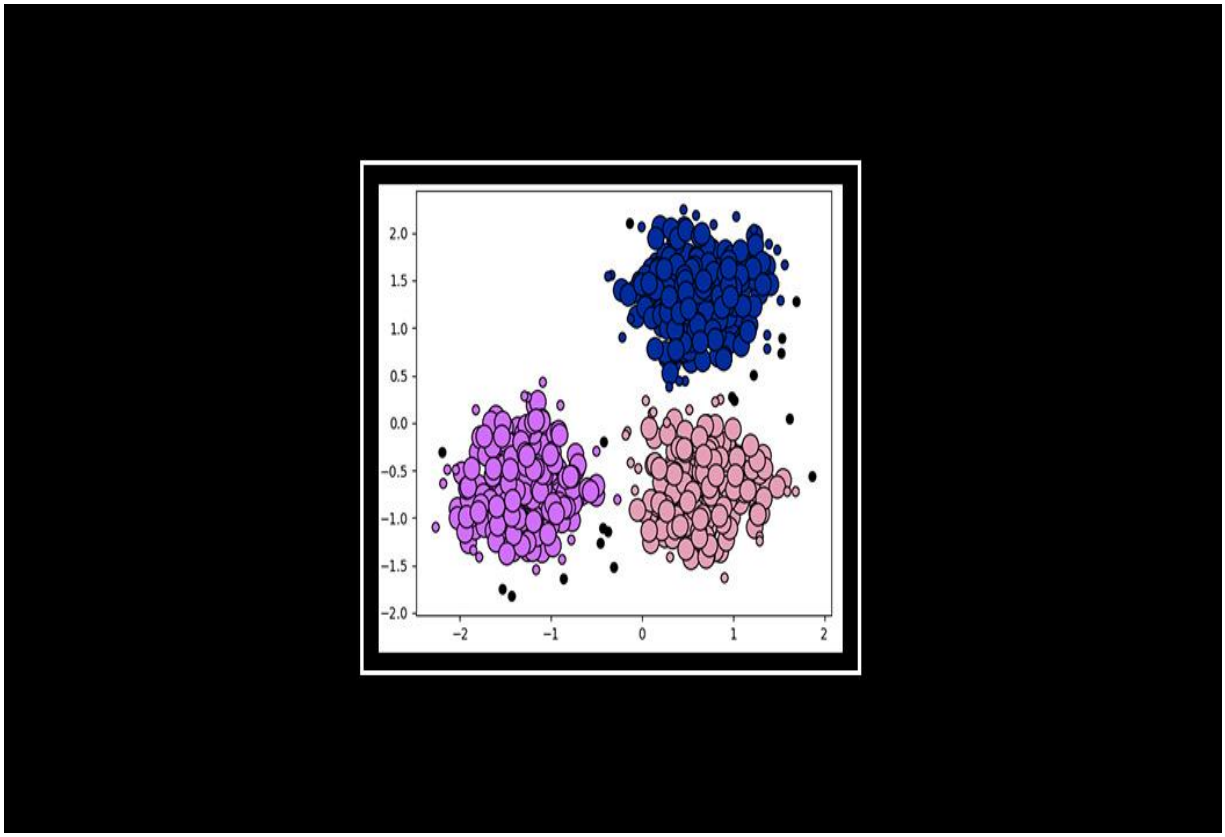


Table Of Contents

Introduction

Let us understand the K-means clustering algorithm with its simple definition.

A K-means clustering algorithm tries to group similar items in the form of clusters. The number of groups is represented by K.

Let's take an example. Suppose you went to a vegetable shop to buy some vegetables. There you will see different kinds of vegetables. The one thing you will notice there that the vegetables will be arranged in a group of their types. Like all the carrots will be kept in one place, potatoes will be kept with their kinds and so on. If you will notice here then you will find that they are forming a group or cluster, where each of the vegetables is kept within their kind of group forming the clusters.

Now we will understand this with the help of a beautiful figure.

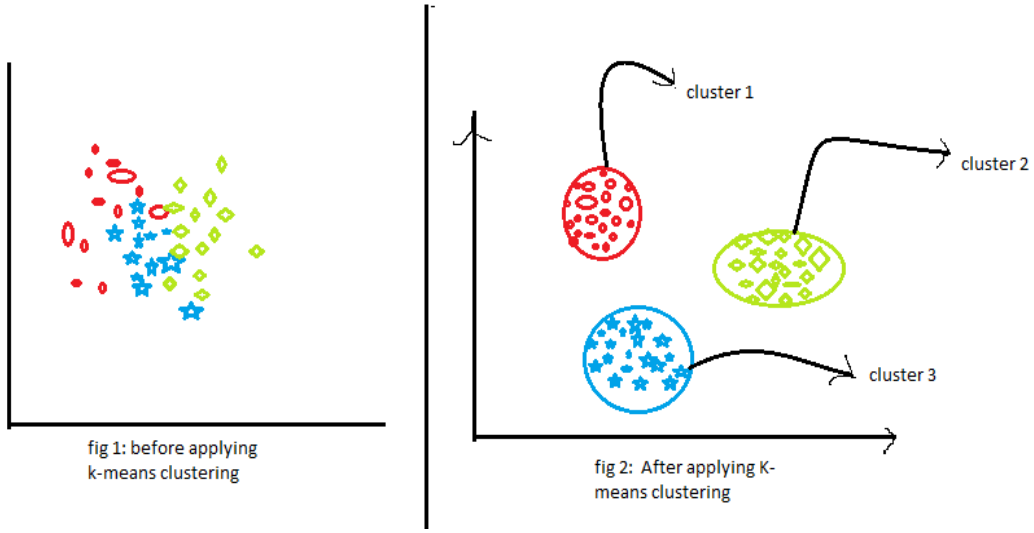


fig 1: before applying k-means clustering

fig 2: After applying K-means clustering

Now, look at the above two figures. what did you observe? Let us talk about the first figure. The first figure shows the data before applying the k-means clustering algorithm. Here all three different categories are messed up. When you will see such data in the real world, you will not able to figure out the different categories.

Now, look at the second figure(fig 2). This shows the data after applying the K-means clustering algorithm. you can see that all three different items are classified into three different categories which are called clusters.

How Does the K-means clustering algorithm work?

k-means clustering tries to group similar kinds of items in form of clusters. It finds the similarity between the items and groups them into the clusters. K-means clustering algorithm works in three steps. Let's see what are these three steps.

1. Select the k values.
2. Initialize the centroids.
3. Select the group and find the average.

Let us understand the above steps with the help of the figure because a good picture is better than the thousands of words.

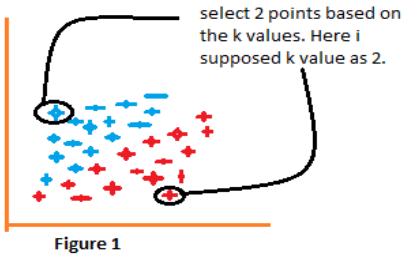


Figure 1

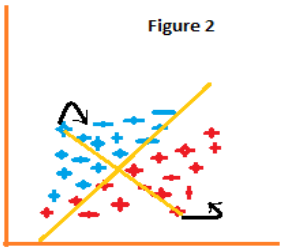
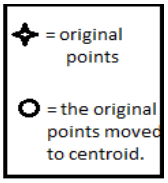
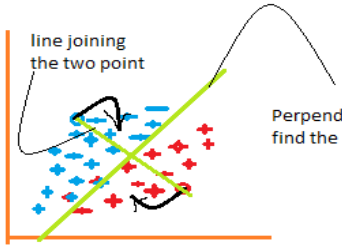


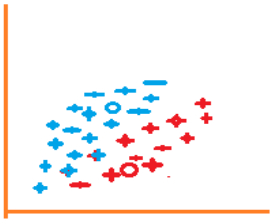
Figure 2

F2: Find the average of all the blue points and red points and move the selected points to **centroid.**





F3: Some of the red points changed to blue points, that means they belong to the group blue now. Again the repeat the same process.



F4: The same process has been applied here. This process will be continued until we get the **two complete different cluster.**

We will understand each figure one by one.

- Figure 1 shows the representation of data of two different items. the first item has shown in blue color and the second item has shown in red color. Here I am choosing the value of K randomly as 2. There are different methods by which we can choose the right k values.
- In figure 2, Join the two selected points. Now to find out centroid, we will draw a perpendicular line to that line. The points will move to their centroid. If you will notice there, then you will see that some of the red points are now moved to the blue points. Now, these points belong to the group of blue color items.
- The same process will continue in figure 3. we will join the two points and draw a perpendicular line to that and find out the centroid. Now the two points will move to its centroid and again some of the red points get converted to blue points.
- The same process is happening in figure 4. This process will be continued until and unless we get two completely different clusters of these groups.

NOTE: Please note that the K-means clustering uses the euclidean distance method to find out the distance between the points.

You will find a lot of explanations regarding the euclidean distance on the internet.

How to choose the value of K?

One of the most challenging tasks in this clustering algorithm is to choose the right values of k. What should be the right k-value? How to choose the k-value? Let us find the answer to these questions. If you are choosing the k values randomly, it might be correct or may be wrong. If you will choose the

wrong value then it will directly affect your model performance. So there are two methods by which you can select the right value of k.

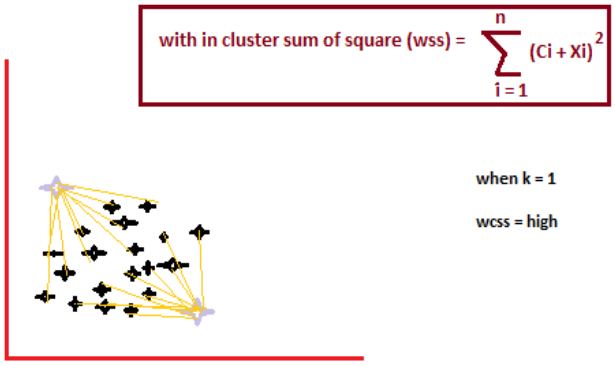
1. Elbow Method.
2. Silhouette Method.

Now, Let's understand both the concept one by one in detail.

Elbow Method

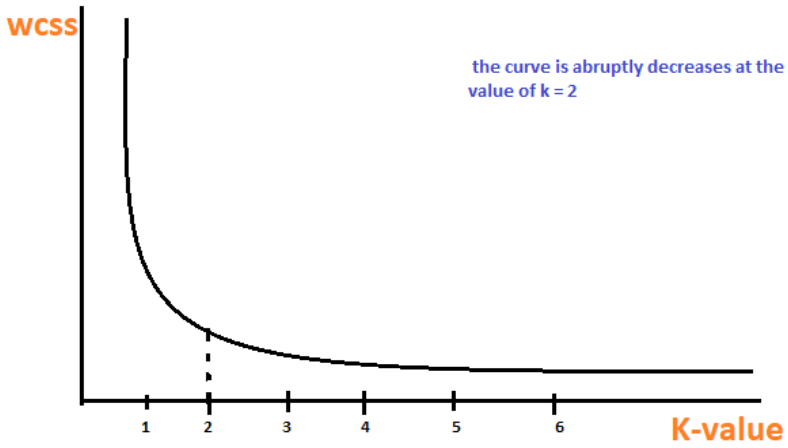
Elbow is one of the most famous methods by which you can select the right value of k and boost your model performance. We also perform the hyperparameter tuning to chose the best value of k. Let us see how this elbow method works.

It is an empirical method to find out the best value of k. it picks up the range of values and takes the best among them. It calculates the sum of the square of the points and calculates the average distance.



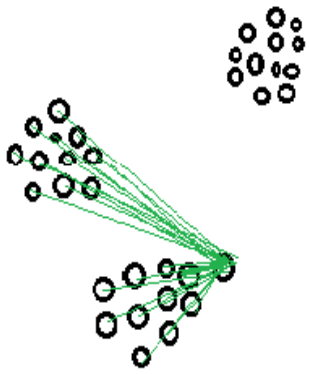
When the value of k is 1, the within-cluster sum of the square will be high. As the value of k increases, the within-cluster sum of square value will decrease.

Finally, we will plot a graph between k-values and the within-cluster sum of the square to get the k value. we will examine the graph carefully. At some point, our graph will decrease abruptly. That point will be considered as a value of k.



Silhouette Method

The silhouette method is somewhat different. The elbow method it also picks up the range of the k values and draws the silhouette graph. It calculates the silhouette coefficient of every point. It calculates the average distance of points within its cluster a (i) and the average distance of the points to its next closest cluster called b (i).



Note : The a (i) value must be less than the b (i) value, that is ai<<bi.

Now, we have the values of a (i) and b (i). we will calculate the silhouette coefficient by using the below formula.

in Worst case s(i) = -1

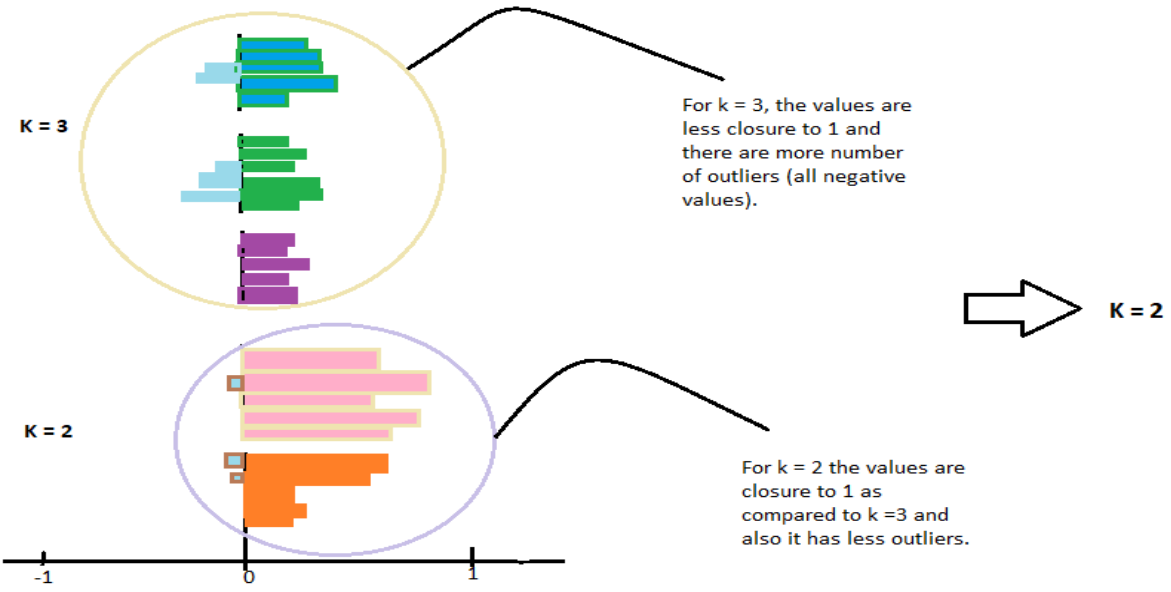
$$s(i) = \frac{b(i) - a(i)}{\text{larger of } b(i) \text{ and } a(i)}$$

a(i) = average distance inside cluster
b(i) = average distance nearest other cluster

Now, we can calculate the silhouette coefficient of all the points in the clusters and plot the silhouette graph. This plot will also helpful in detecting the outliers. The plot of the silhouette is between -1 to 1.

Note that for silhouette coeficient equal to -1 is the worst case scenario.

Observe the plot and check which of the k values is closer 1.



Also, check for the plot which has fewer outliers which means a less negative value. Then choose that value of k for your model to tune.

Advantages of K-means

1. It is very simple to implement.
2. It is scalable to a huge data set and also faster to large datasets.
3. it adapts the new examples very frequently.
4. Generalization of clusters for different shapes and sizes.

Disadvantages of K-means

1. It is sensitive to the outliers.
2. Choosing the k values manually is a tough job.
3. As the number of dimensions increases its scalability decreases.

14. Is K means deterministic algorithm ?

Answer - Is k-means a deterministic algorithm?

The basic 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.

Why is k-means non-deterministic?

The non-deterministic nature of K-Means is due to its random selection of data points as initial centroids. ... The key idea of the algorithm is to select data points which belong to dense regions and which are adequately separated in feature space as the initial centroids.

What is deterministic clustering?

Hierarchical Agglomerative Clustering is deterministic except for tied distances when not using single-linkage. DBSCAN is deterministic, except for permutation of the data set in rare cases. k-means is deterministic except for initialization. You can initialize with the first k objects, then it is deterministic, too.

Is k-means supervised or unsupervised?

K-means clustering is the unsupervised machine learning algorithm that is part of a much deep pool of data techniques and operations in the realm of Data Science. It is the fastest and most efficient algorithm to categorize data points into groups even when very little information is available about data.

Which is needed by K-means clustering?

Explanation: K-means requires a number of clusters. ... Explanation: Hierarchical clustering requires a defined distance as well. 10. K-means is not deterministic and it also consists of number of iterations.

Is K nearest neighbor the same as K-means?

K-means clustering represents an unsupervised algorithm, mainly used for clustering, while KNN is a supervised learning algorithm used for classification.

What is K-means used for?

The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.

Is K-means clustering machine learning?

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. ... Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

Is Random Forest non deterministic?

Like the name suggests, random forests do make use of randomness, or at least, pseudo-randomness. If we're only concerned about whether or not the algorithm is deterministic in the usual sense of the word (at least, within computer science), the answer is no.

Why is PCA deterministic?

PCA is a deterministic algorithm which doesn't have parameters to initialize and it doesn't have local minima problem like most of the machine learning algorithms has.

Is K nearest neighbors a deterministic algorithm?

K Nearest Neighbor (KNN) is a basic deterministic algorithm for locating which is widely used in fingerprinting approach. The performance of the KNN can be improved extensively by employing appropriate selection algorithm.

Why hierarchical clustering is deterministic?

The final cluster assignments are then represented by either the centroid or the medoid. Hierarchical clustering is deterministic, which means it is reproducible. However, it is also greedy, which means that it yields local solutions.

Can k-means be supervised?

You can have a supervised k-means. You can build centroids (as in k-means) based on your labeled data. Nothing stops you. If you want to improve this, Euclidean space and Euclidean distance might not provide you the best results.

Why k-means is unsupervised learning?

Example: Kmeans Clustering. Clustering is the most commonly used unsupervised learning method. This is because typically it is one of the best ways to explore and find out more about data visually.

Can k-means clustering used for supervised learning?

The k-means clustering algorithm is one of the most widely used, effective, and best understood clustering methods. ... In this paper we propose a supervised learning approach to finding a similarity measure so that k-means provides the desired clusterings for the task at hand.

What is K in K-means?

Introduction to K-Means Algorithm The number of clusters identified from data by algorithm is represented by 'K' in K-means. In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum.

Will K-means always converge?

The algorithm always converges (by-definition) but not necessarily to global optimum. The algorithm may switch from centroid to centroid but this is a parameter of the algorithm (precision , or delta). ... Max Num Iterations , if algorithm reaches that number of iterations stop the algorithm.

How do you interpret k-means clustering?

Interpreting the meaning of k-means clusters boils down to characterizing the clusters. A Parallel Coordinates Plot allows us to see how individual data points sit across all variables. By looking at how the values for each variable compare across clusters, we can get a sense of what each cluster represents.

Which is better KNN or SVM?

SVM take cares of outliers better than KNN. If training data is much larger than no. of features(m>>n), KNN is better than SVM. SVM outperforms KNN when there are large features and lesser training data.

Is K-nearest neighbor supervised or unsupervised?

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems.

How do we decide the value of k in KNN and K-means algorithm?

The optimal K value usually found is the square root of N, where N is the total number of samples. Use an error plot or accuracy plot to find the most favorable K value. KNN performs well with multi-label classes, but you must be aware of the outliers.