A Simple Explanation of K-Means Clustering

analyticsvidhya.com/blog/2020/10/a-simple-explanation-of-k-means-clustering

aditya610 October 4, 2020

Home » A Simple Explanation of K-Means Clustering
aditya610, October 4, 2020 Login to Bookmark this article
This article was published as a part of the Data Science Blogathon.

Overview

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.

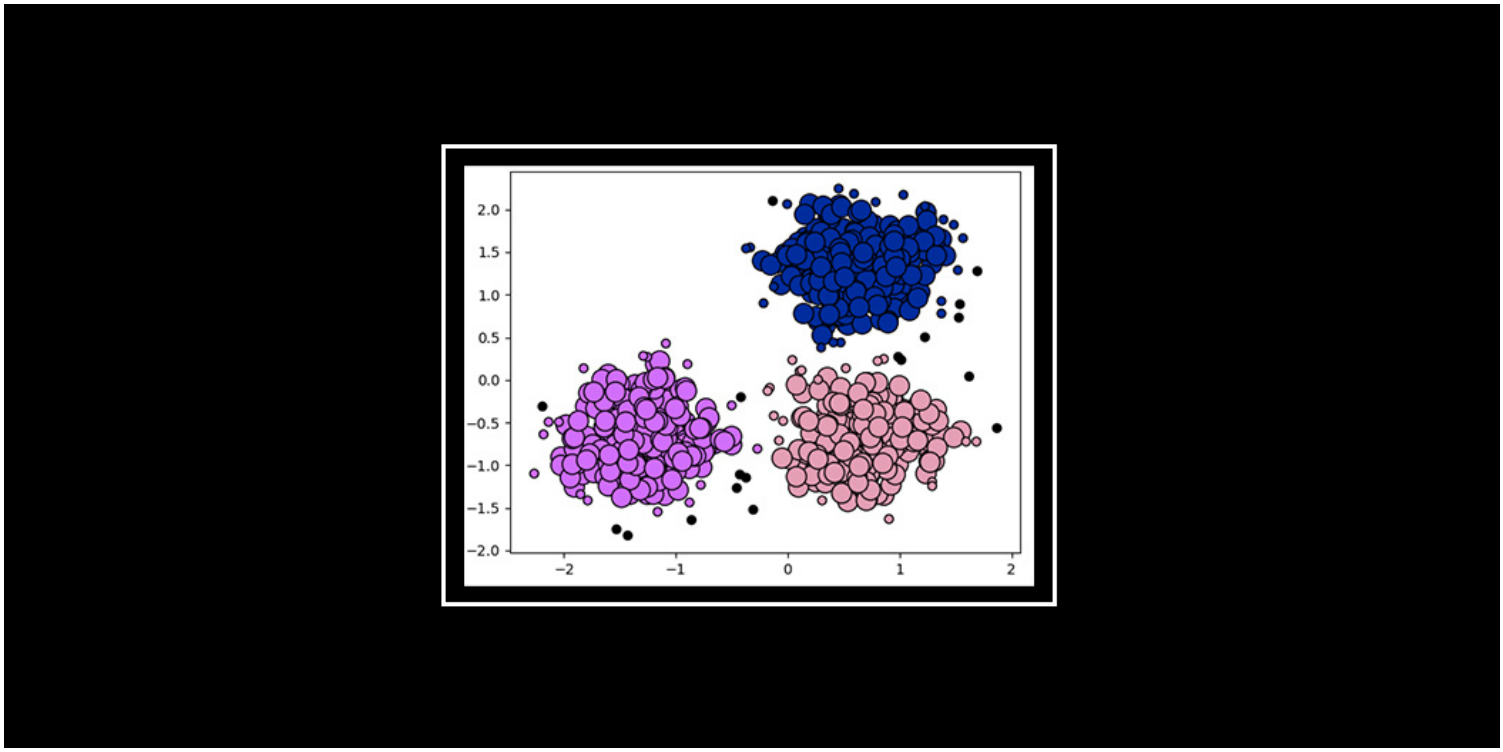


Table Of Contents

- Introduction
- How does the K-means algorithm work?
- How to choose the value of K?
    - Elbow Method.
    - Silhouette Method.
- Advantages of k-means.
- Disadvantages of k-means.

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.

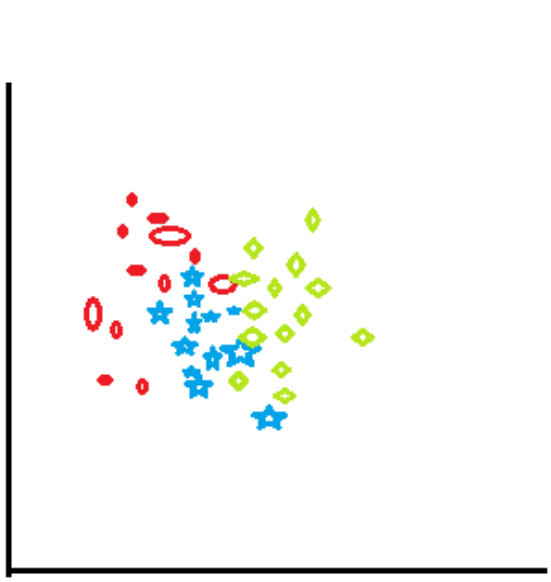


fig 1: before applying k-means clustering

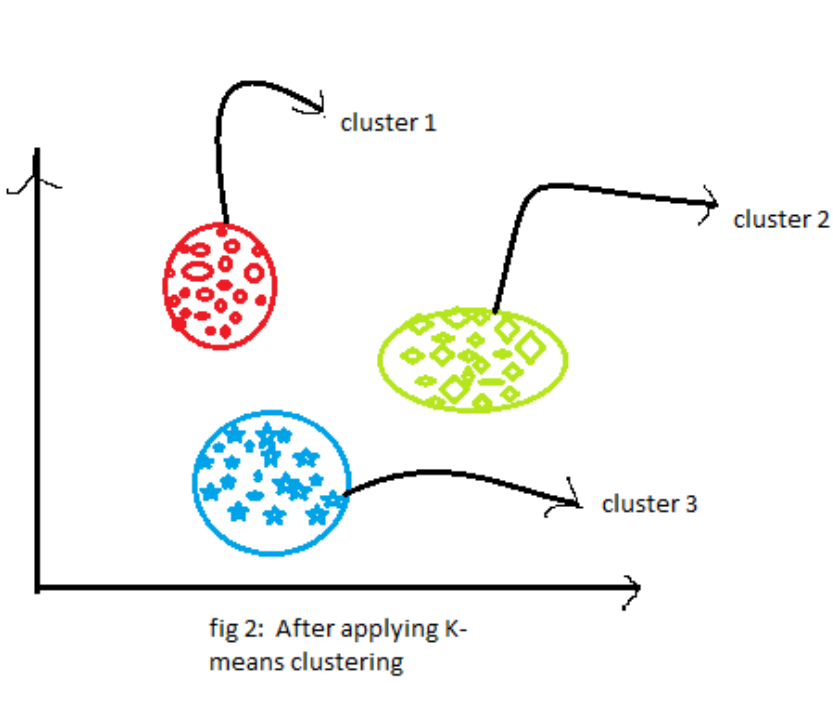


fig 2: After applying K-means clustering

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.

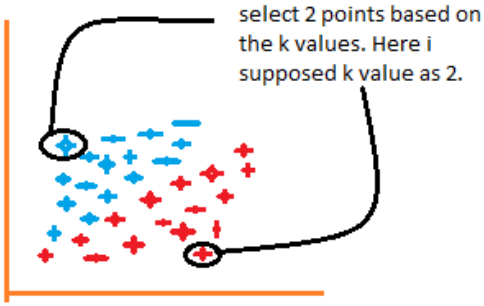


Figure 1

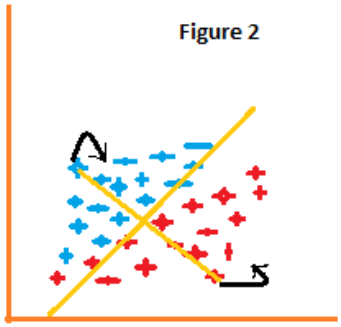
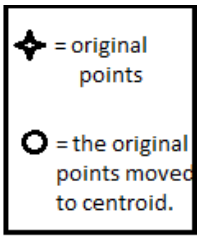
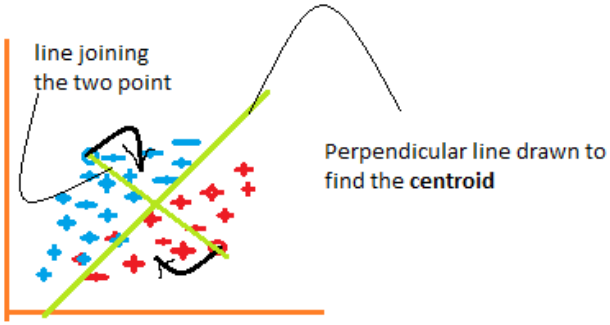


Figure 2



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.

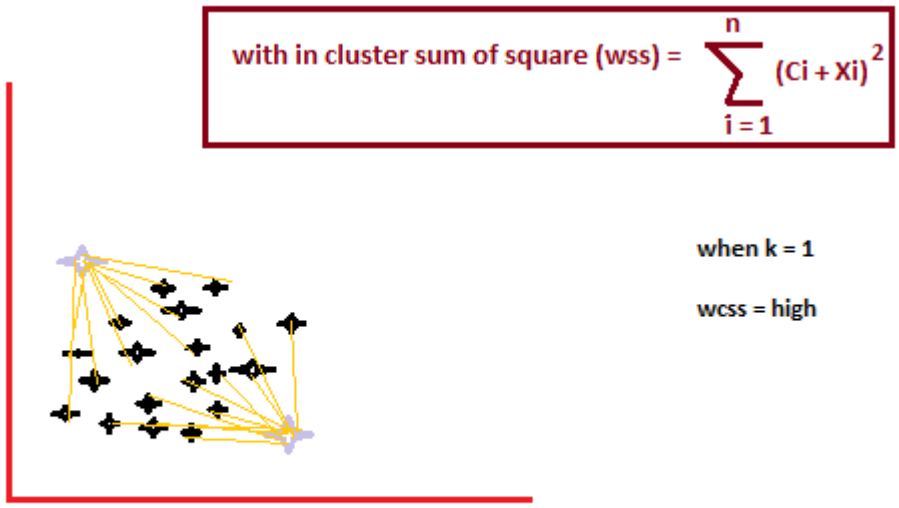
1. Elbow Method.
2. Silhouette Method.

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

Elbow Method

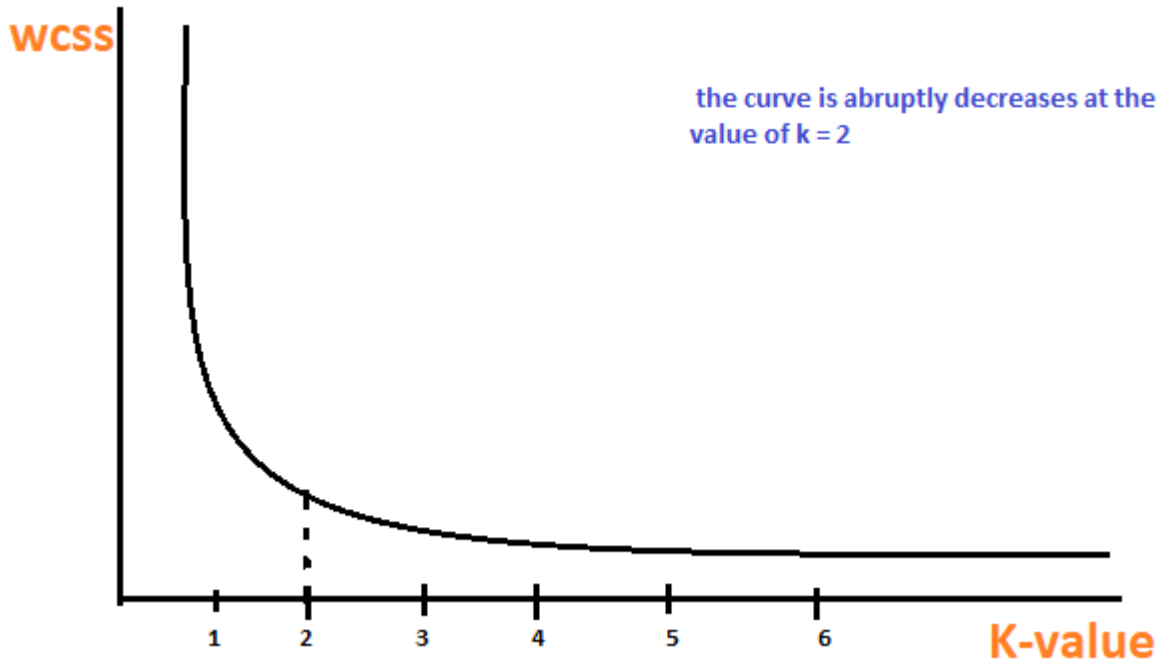
Elbow is one of the most famous methods by which you can select the right value of k and boost your model performance. We also perform the hyperparameter tuning to chose the best value of k. Let us see how this elbow method works.

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.



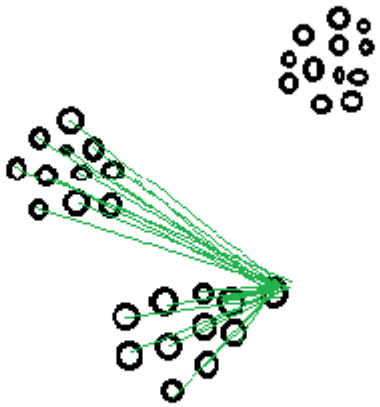
When the value of k is 1, the within-cluster sum of the square will be high. As the value of k increases, the within-cluster sum of square value will decrease.

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

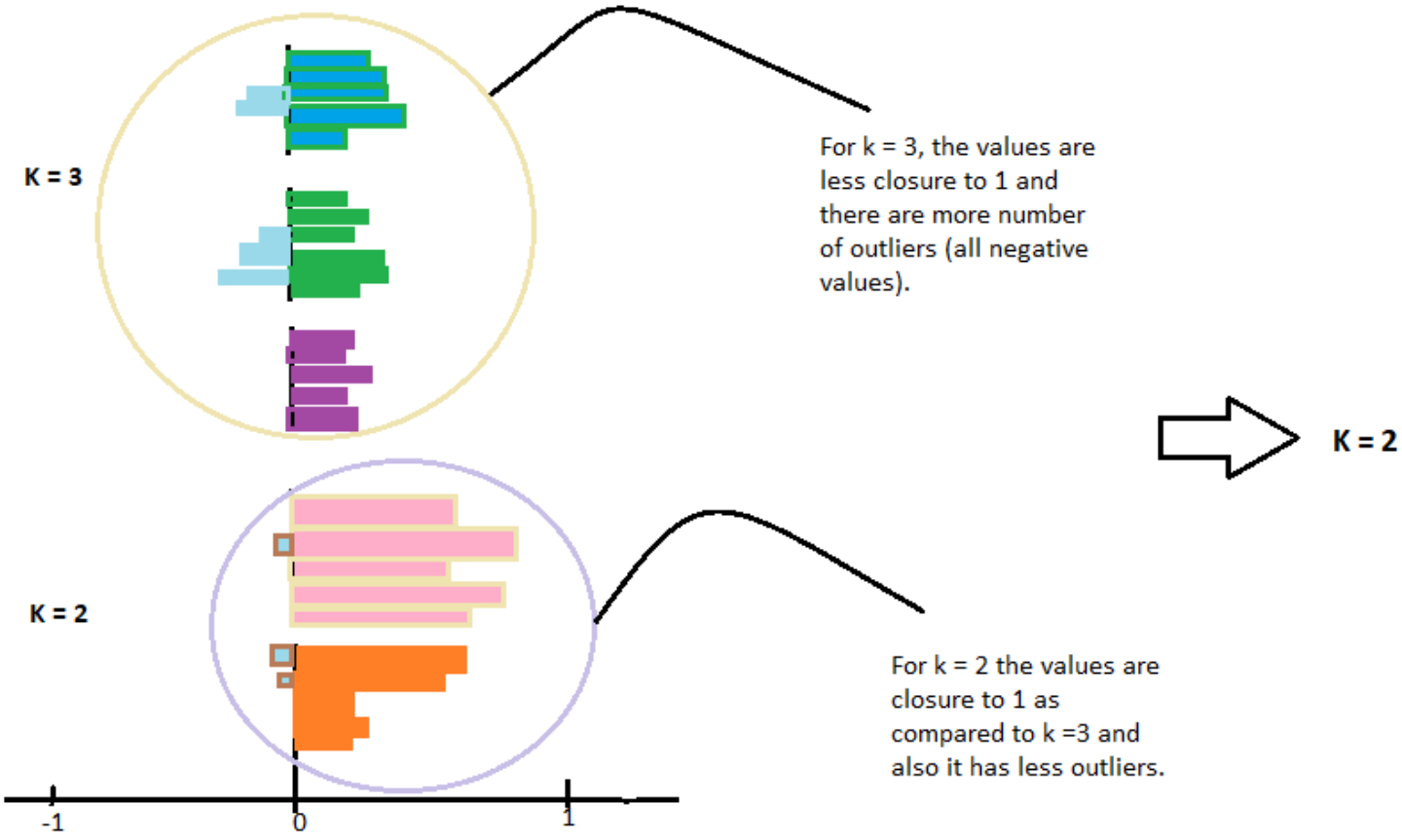
$$s(i) = \frac{b(i) - a(i)}{\text{larger of } b(i) \text{ and } a(i)}$$

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.

You can also read this article on our Mobile APP

This article is quite old and you might not get a prompt response from the author. We request you to post this comment on Analytics Vidhya's **Discussion portal** to get your queries resolved