K-Means Report

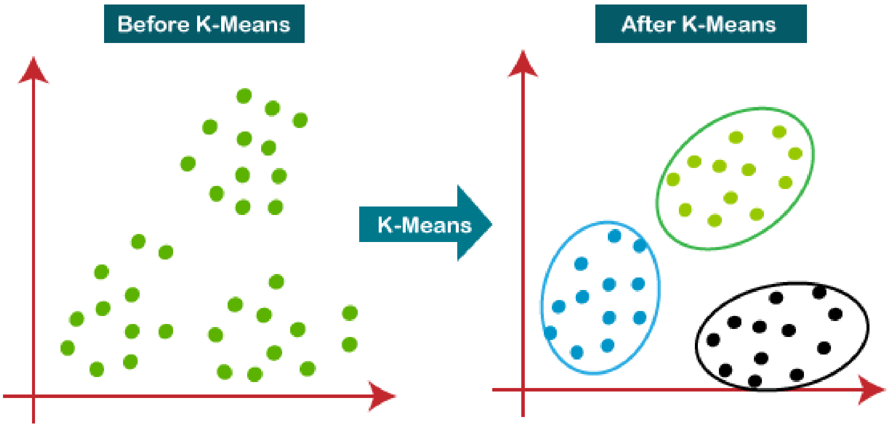
Donal Loitam

July 20, 2022

Contents

1 Introduction

- K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters.
- The number of clusters found from data by the method is denoted by the letter 'K' in K-means.
- The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.
- the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.



2 Key Points of the Algorithm

Let's take a simple example to understand this algorithm. So imagine you have a set of numerical data of cancer tumors in malignant or benign. However, you have no idea how to identify which tumor is what because nobody had the time to label the entire set of features (most data in the world are unlabeled). In this case, you need K-means algorithm because it works on unlabeled numerical data and it will automatically and quickly group them together into 2 clusters.
NOTE : For this example, we chose k=2 because, we already know it can either be malignant/benign But what if we don't actually know what the value of 'k' is ? There is a solution for it

Steps of the algorithm
Step 1: Initialization
The first thing k-means does, is randomly choose K examples (data points) from the dataset or any random K points (the blue and orange points) as initial centroids and that's simply because it does not know yet where the center of each cluster is. (a centroid is the center of a cluster).

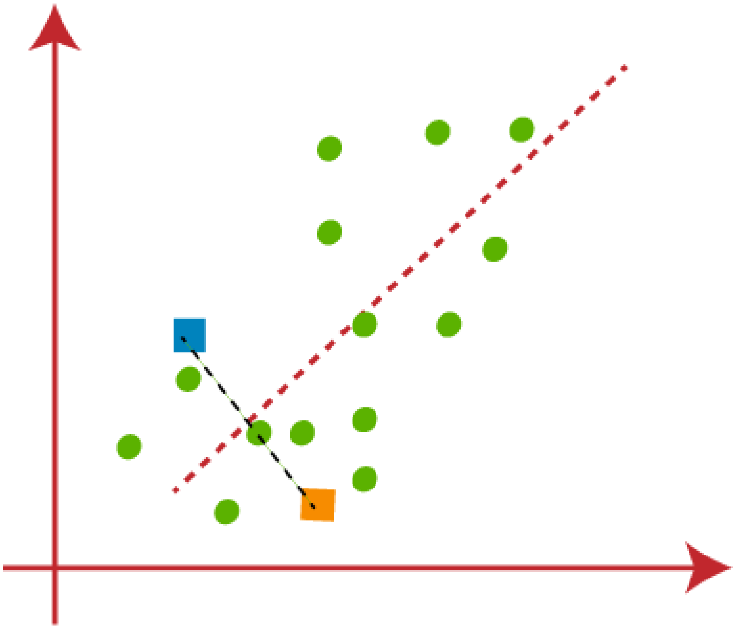


Figure 1: Initialization of centroids

Step 2: Cluster Assignment
Then, all the data points that are the closest (similar) to a centroid will create a cluster. If we're using the Euclidean distance between data points and every centroid, a straight line is drawn between two centroids, then a perpendicular bisector (boundary line) divides this line into two clusters.
From the above image, it is clear that points left side of the line is near to the K1 or blue centroid, and points to the right of the line are close to the yellow centroid. Let's color them as blue and yellow for clear visualization.

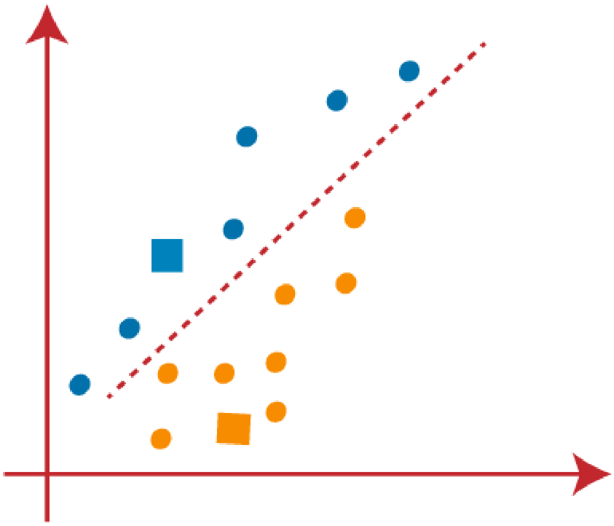


Figure 2: Cluster assignment

Step 3: Move the centroid

Now, we have new clusters, that need centers. A centroid's new value is going to be the mean of all the examples in a cluster. We'll keep repeating step 2 and 3 until the centroids stop moving, in other words, K-means algorithm is converged.

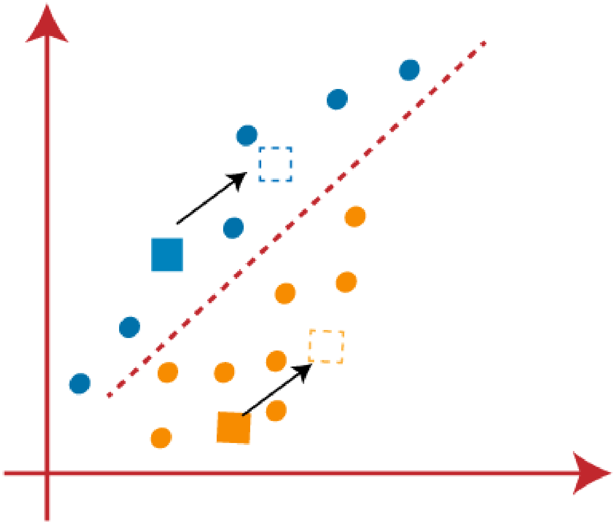
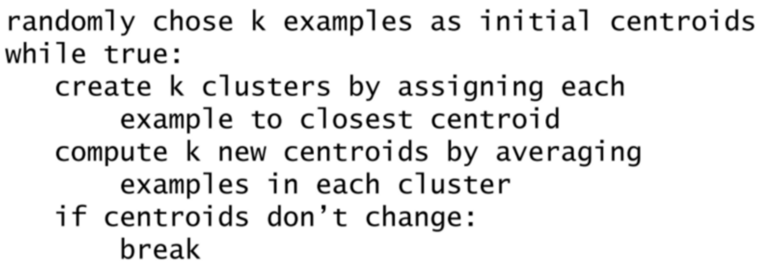


Figure 3: Changing Centroids

Here is the k-means algorithm as a psuedo code :



```
randomly chose k examples as initial centroids
while true:
   create k clusters by assigning each
       example to closest centroid
   compute k new centroids by averaging
       examples in each cluster
   if centroids don’t change:
       break
```

Figure 4: Pseudo code

Unlucky Centroids

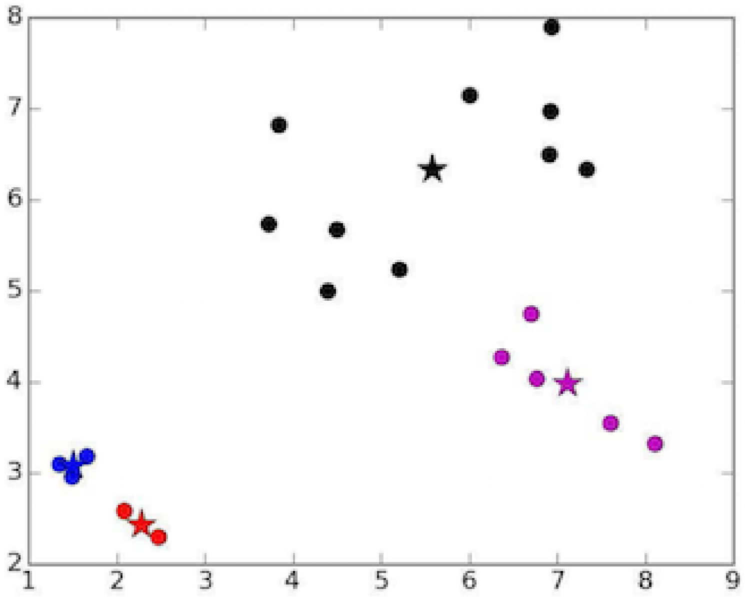


Figure 5: The blue and red stars are unlucky centroids :(

Choosing poorly the random initial centroids will take longer to converge or get stuck on local optima which may result in bad clustering. in the picture above, the blue and red stars are unlucky centroids. There are two solutions:

- Distribute them over the space.
- Try different sets of random centroids, and choose the best set.

3 Some Questions

1. Is Feature Scaling required for the K means Algorithm?
Ans Yes, K-Means typically needs to have some form of normalization done on the datasets to work properly since it is sensitive to both the mean and variance of the datasets.

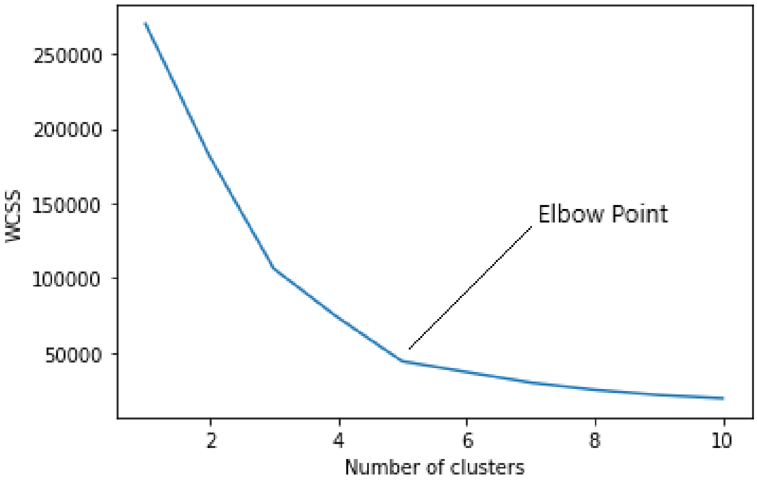
For Example, let's have 2 variables, named age and salary where age is in the range of 20 to 60 and salary is in the range of 100-150K, since scales of these variables are different so when these variables are substituted in the euclidean distance formula, then the variable which is on the large scale suppresses the variable which is on the smaller scale
2. Why is the plot of the within-cluster sum of squares error (inertia) vs K in K means clustering algorithm elbow-shaped? Discuss if there exists any other possibility for the same with proper explanation.
Ans. Let's understand this with an example, Say, we have 10 different data points present, now consider the different cases:

- k=10: For the max value of k, all points behave as one cluster. So, within the cluster sum of squares is zero since only one data point is present in each of the clusters. So, at the max value of k, this should tend to zero.
- K=1: For the minimum value of k i.e, k=1, all these data points are present in the one cluster, and due to more points in the same cluster gives more variance i.e, more within-cluster sum of squares.
- $1 < k < 10$: When you increase the value of k from 1 to 10, more points will go to other clusters, and hence the total within the cluster sum of squares (inertia) will come down. So, mostly this forms an elbow curve instead of other complex curves.

3. What are the challenges associated with K means Clustering?
Ans : Initialization sensitivity.
We were using randomization i.e, initial k-centroids were picked randomly from the data points.

This Randomization in picking the k-centroids creates the problem of initialization sensitivity which tends to affect the final formed clusters. As a result, the final formed clusters depend on how initial centroids were picked.
4. What are the ways to avoid the problem of initialization sensitivity in the K means Algorithm?
Ans. Repeat K means: It basically repeats the algorithm again and again along with initializing the centroids followed by picking up the cluster which results in the small intracluster distance and large intercluster distance.
5. How to decide the optimal number of K in the K means Algorithm?
Ans: Direct observation If we consider the data of a shopkeeper selling a product in which he will observe that some people buy things in summer, some in winter while some in between these two. So, the shopkeeper divides the customers into three categories. Therefore, K=3.

Elbow method: This method finds the point of inflection on a graph of the percentage of variance explained to the number of K and finds the elbow point.
As the number of clusters(K) increases, the WCSS value will start to decrease. WCSS value is largest when K = 1. When we analyze the graph we can see that the graph will rapidly change at a point and thus creating an elbow shape.


250000
200000
150000
100000
50000
WCSS
2
4
6
8
10
Number of clusters
Elbow Point