K-means

Mannem Charan AI21BTECH11019

Abstract

This report consists of my basic understanding of one of the popular Ml methods ”K-means Clustering”.

1 K-means

K-means or K-means clustering algorithm is used to solve the clustering problem. It comes under unsupervised machine learning, where the data don't have any labels.

2 What is Clustering Problem?

Often it happens the we don't have labels for the data.So in those cases we try to divide the data into different groups intuitively.For example see the figure 2.1, we are trying to group the data points which are close to each other.

These groups are known as ”Clusters”. And this process of grouping is known as ”Clustering”.

3 Understanding K-means Clustering

In the k-means algorithm, we will try to divide the data set into ”k” clusters and a data point will become member of a cluster if it closest to that cluster in comparison to other clusters.It is kinda similar to k-NN algorithm but there the data we use have some pre-defined labels but here we are just trying to seperate the data into groups/clusters. The word ”means” represent center of the each cluster.And we want to make sure that

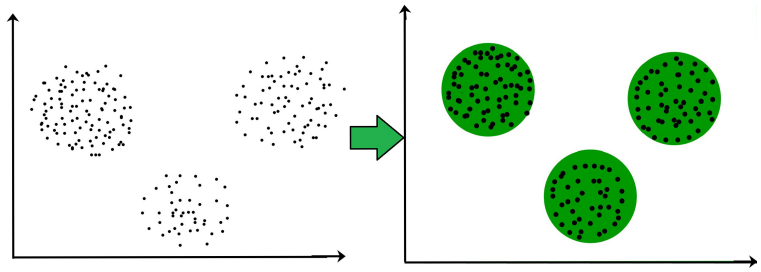

Fig. 2.1: An example of grouping/clustering

each point in a Cluster say A should be closest to the mean of A than the means of other clusters.In practice, the algorithm is done by following steps,

1) First we will intialise the value of "k", the number of clusters we want to make from the data.
2) Now we should generate k centers of clusters but initially we can't find the exact position of centers of the clusters.So we will generate "k" points randomly.
3) Since we initialised the mean points, we will assign the data points to each mean point depending how close they are to the mean point.To find "closeness" between data points and mean points we will use euclidean distance metric,

$$d(\mathbf{x}, \mathbf{c}) = \sqrt{\sum_{i=1}^{n} (x_i - c_i)^2} \tag{3.1}$$

where,

- $\mathbf{x}$ = data point
- $\mathbf{c}$ = mean/centroid point
- $\mathbf{n}$ = no. of features.

So for every data point we will find the closest mean point and assign the point to that cluster.

4) Then we will update the mean point of the cluster as the average of all the data points in that cluster.So the new mean point will be,

$$\mathbf{c'} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x_i} \tag{3.2}$$

where,

- $\mathbf{c'}$ = new mean point of the cluster
- N = The no. of data points in the cluster.
- $\mathbf{x_i}$ = i^{th} data point in the cluster.

5) Then repeat the step 3 and 4 until there is no change in the grouping of clusters.

4 Advantages of k-means

1) It is easy to implement.
2) One of the preffered methods if the dataset is large and don't pre-defined labels.
3) Guarantees the convergence.

5 Disadvantages of k-means

1) Uses lot of computational power.

2) Asks the user to give the ”k” value, can be resolved by plotting the squared error vs k curve.
3) Mean points can be dragged near to outlier points or outliers might get their own clusters instead of being ignored.
4) Causes problems for high-dimensional data because the distance measure will not be effective to seggregate the clusters.

6 Questions

1) Which part of machine learning k-means algorithm belongs to?
2) What is a cluster?
3) What is the difference betweeen the k-means and k-NN algorithms?
4) What are the challenges in k-means algorithm?
5) How the optimal k can be found for the given data?

7 Solution

1) k-means algorithm can under unsupervised learning where input data doesnot have pre-defined labels.
2) In machine learning, cluster is group of data points having some similarities.
3)
   - k-means algorithm used for unsupervised learning whereas k-NN is used for supervied learning.
   - The k in k-means represents the number of clusters to be grouped in the data whereas in k-NN it represents the no. of closest number you want to consider to classify the data.
   - In k-means we want to **group** the data based on how close the points are , whereas in k-NN we try to **classify** a data point to a class based on majority in neighbours.
   - K-means often has outliers problems whereas it can be resolved in k-NN by taking higher values of k.
4) The main challenges in k-means algorithm are,
   - Choosing the value of k.
   - Generating the k centroids/mean points.
5) The optimal k can be found for a given data using elbow method.In this method we try to plot within-cluster sum squared error(WSS) against different values of k. The sharp point of bend or the point where it looks like elbow, that point will be our optimal k value.The WSS is formulated as,

$$WSS = \sum_{j=1}^{k} \sum_{i=1}^{N_j} \left\| \mathbf{x_i^{(j)}} - \mathbf{C_j} \right\|^2 \tag{7.1}$$

where,

- k = no. of clusters
- $\mathbf{C_j}$ = centroid of j^{th} cluster.
- N_j = no. of samples in j^{th} cluster.

- $\mathbf{x_i^{(j)}}$ = i^{th} data point of j^{th} cluster.

We are taking the point where there is abrupt decrease because we want to make the centroid as close as possible to the all the points in the cluster.