12/08/21

Machine learning Algorithm

K Means Clustering

- Simplest & popular unsupervised machine learning algorithm.

Objective of K-Means :

~ Grap similar data points together and discover underlying patterns.

~ Cluster refers to a collection of data pants aggregated together because of certain similarities.

~ Target number = K

~ Centroid = Imaginary / real location representing the center of the cluster.

~ K means algorithm identifies k number of centroids & and allocates every data point to the nearest cluster.

~ We need to keep the centtroids as small as possible.

~ K-Means

Averaging of data that is finding the centroid.

Clustering.

- Technique to get an intuition about structure of the data.
- defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are similar while the data points in different cluster are different.

- Clustering Analysis :-

→ Done on the basis of ~~samples~~ features where we try to find subgroups of samples based on features.

eg: Used in market segmentation :

→ To find customers that are similar to each other whether in terms of behaviour/ attributes.

→ Considered an unsupervised learning method since we don't have the ground truth to compare the output.

→ We only want to investigate the structure of the data by grouping the data points into distinct subgroups.

K-Means Algorithm

- Iterative algorithm
  - Partition the dataset into Kpre-defined distinct non-overlapping subgroups (cluster)
  - Each dataset belongs to only one group.
- Tries to make intracluster data points as similar as possible. While keeping the clusters as different (far) as possible.

K-Means Works as follows:

(1) - Specify number of clusters K.

(2) - Initialize centroids by first shuffling the dataset & then randomly selecting K data points for the centroids without replacement.

(3) Keep iterating until there is no change to the centroids.

↳ Compute the sum of the squared distance between data points & all centroids.

↳ Assign each data point to the closest cluster (centroid)

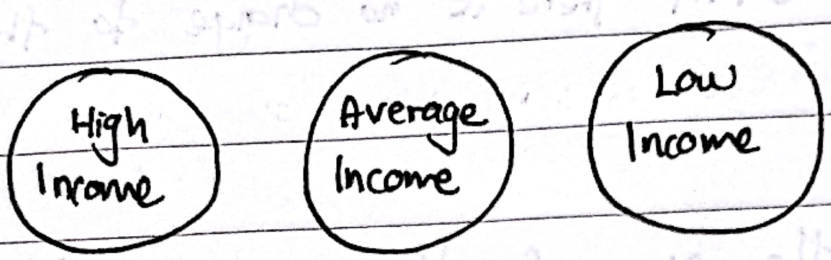
↳ Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

14/08/21

More detailed Explanation

What is clustering?

eg - A bank want to give credit cards to customers.



- The groups are known as clusters & the process of creating these groups is known as clustering.

~ Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data.

Application of Clustering in Real-World

- Customer Segmentation
  - In banks, e-commerce, advertising.

- Document Clustering
  - have multiple document, clustering help us to group these documents to be in the same clusters.

- Image Segmentation
  - club similar pixels in the image together.

- Recommendation Engines.
  - Recommend playlists of songs in spotify.

Introduction to K-Means Clustering

-> K-Means Clustering Technique:

↳ An algorithm that tries to minimize the distance of the points in a cluster with their centroid.

8 points to apply K-Means.

1. Choose the number of cluster k.
2. Select k random points from the data as centroids.
   -> if we have #2 clusters, hence $k = 2$.
3. Assign all the points to the closest cluster centroid.
4. Recompute the centroids of newly formed clusters.
5. Repeat step 3 & 4.
6. Stopping Criteria for K-Means Clustering.
   - Centroids of newly formed cluster do not change.
   - Points remain in the same cluster
   - Maximum number of iterations reached.