K - Means Clustering

Linux Campus Club SJCE

Unsupervised Learning

- Unsupervised Learning is a type of Machine Learning Algorithm used to draw inferences from datasets consisting of input data *without labelled responses.*
- How is it different from Supervised Learning ??
- Some Unsupervised Learning Algorithms include
  - Clustering
  - Anamoly Detection

Clustering

- Grouping the set of objects(data samples) that have similar data features.
- These groups are referred to as Clusters.

Types of Clustering :

★ Hierarchical Clustering
  ○ Agglomerative algorithm
  ○ Divisive algorithm
★ Partitional Clustering
  ○ K - Means Clustering

Some applications of Clustering

- Market Segmentation
- Astronomical data analysis
- Social Network Analysis
- Organizing computer clusters

K - Means Clustering

- K-Means Algorithm is an algorithm to classify objects based on attributes/features into K number of group.

- The grouping is done by minimizing the sum of squares distances between data and the corresponding cluster centroid.

The algorithm

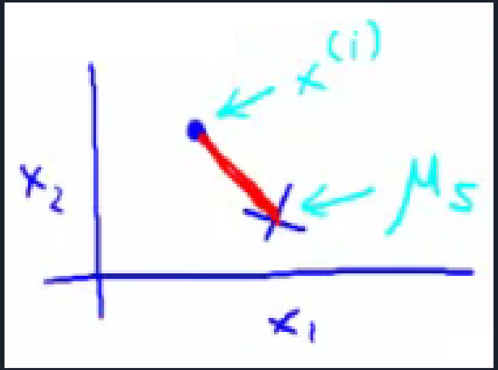
Input : Data Samples { x1,x2,x3......xm} and K (number of clusters).

- ➔ Randomly allocate K points as cluster centroids.
- ➔ Cluster assignment step:
    - ◆ Go through each data sample and depending on its's distacne from centroids, assign each sample to one of the centroids.
- ➔ Centroid updation:
    - ◆ Take the mean of the each clusters and shift the new centroid to the mean.

Repeat step 2 and step 3 until convergence or for specified number of iterations.

The Optimization Objective

- Like supervised, even unsupervised learning algorithms have an optimization objective.
- This is helpful in debugging.
- This cost function is referred to as distortion.



The red line in the figure indicates the distance between x^i (i^{th} data sample) and its corresponding cluster centroid.

Random Initialization

- The convergence of this algorithm also depends upon initialization step
- We might face the convergence problem if we randomly initialize the cluster centroids.
- One method is to randomly pick K training samples and set centroids to these sample values.
- Risk of local optimum.

How to choose the number of clusters? - The Elbow Method

- Normally K is chosen manually after data visualization.
- This technique will be ambiguous if data is distributed uniformly.

Elbow Method

- ➔ Plot cost function after convergence v/s number of clusters curve.
- ➔ The value at which elbow is obtained is our K.

Thank You