Segmentation or Clustering

07 January 2022 20:31

Our learning so far :

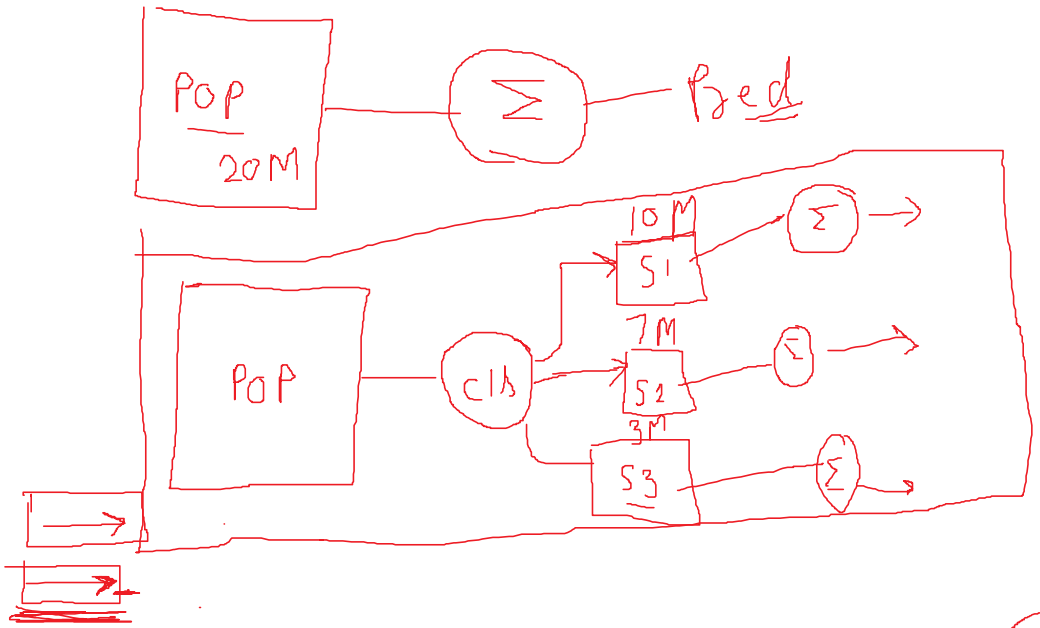
- Regression (Supervised machine learning)
  - Linear Regression
  - Decision Tree
- Classification (Supervised machine learning)
  - Logistic Regression
  - Decision Tree
  - Ensemble - Bagging models and Boosting models
- Segmentation or clustering
  - K-Means Algorithm
  - Agglomerative Approach (Hierarchical Clustering)

What is clustering ?

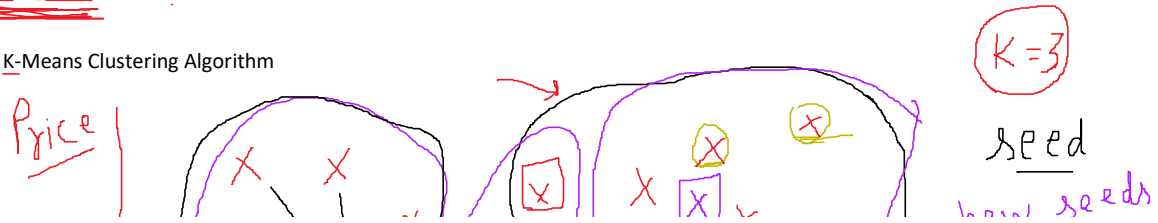
- Grouping on common factor
- It is an unsupervised machine learning
- It is not a predictive modelling technique but it is a helper for precise predictive models
- We create homogeneous groups from heterogeneous population based upon some similarity criteria

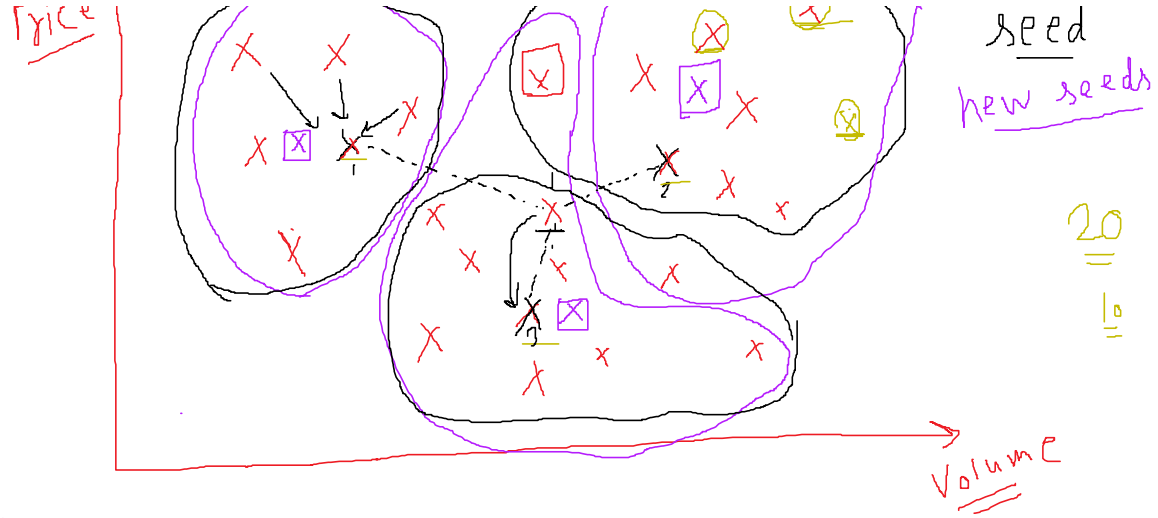
- On a ground you have 20k people, Ask is to create segments out of these people ?
  - Tell me the similarity criteria
    - Height
    - Age
    - Colour they are wearing
    - With kids or w/o kids
    - Married
    - Gender
    - Ethnicity
    - Type of employment ……
      - Cluster-1 :- All females, private employment, Age 30-35, Height between 5 to 5-5
      - Cluster-2 :- All male above 50 years height less than 5.5 and unemployed.

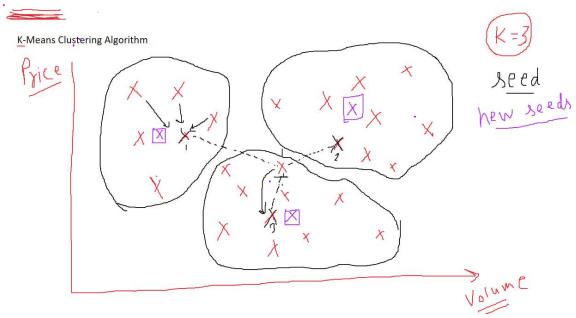
- What is the business values from these clusters ?
  - Large scale businesses can't treat each customer individually. Clustering is a proxy - to mimic the celebrity experience.

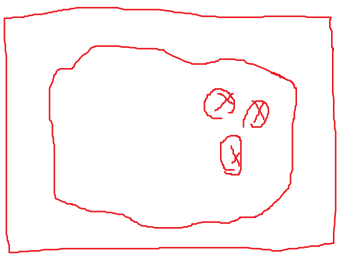


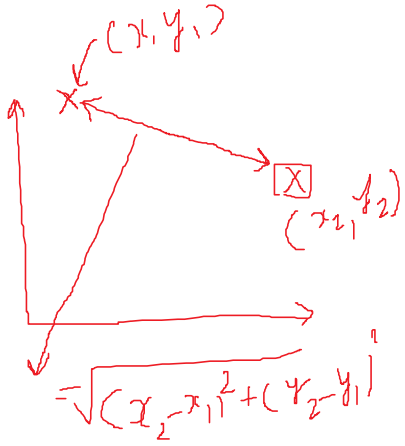
K-Means Clustering Algorithm











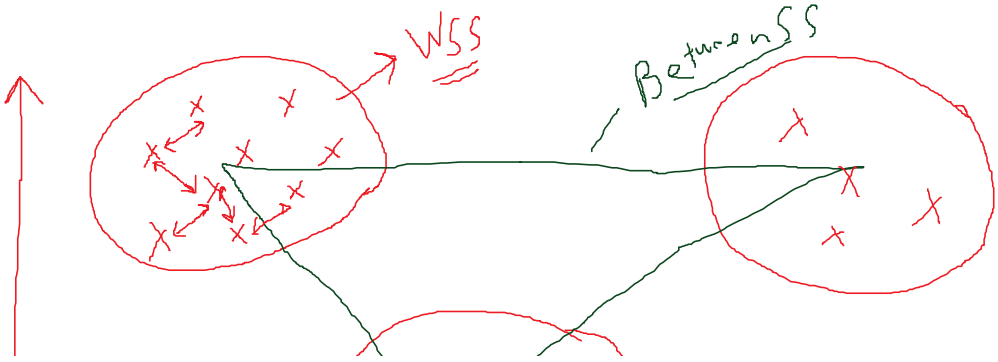
Steps

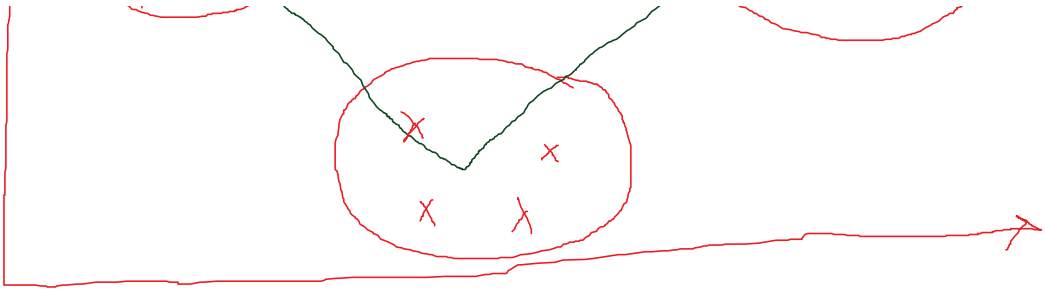
- Specify value of K=3 (how to find out optimal value of K, to be discussed ?)
- Pick the "random 3 data points" - Seeds
- Distance of each data points will be calculated from these three seeds
- Based upon distance proximity, data points will become part of one of the seed, you will have first set of clusters
- For each of these clusters, find out the data points which represent centre. This is my new seeds
- Re-compute distance
- Re-create clusters
- New seed
- ……..

If you will repeat this exercise for a good number of times, a point will come, after which data points will not change their cluster (Cluster boundaries will be freezed). This is when you will say that algorithm has converged and we are able to find the stable cluster. The seed of final clusters is called as "centroid".

$$\rightarrow (x_1, y_1, z_1, a_1, b_1) \quad (x_2, y_2, z_2, a_2, b_2)$$

$$= \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2 + (z_2-z_1)^2 + (a_2-a_1)^2 + \dots}$$





WSS - Within group sum of square - Average distance between the data points in a cluster

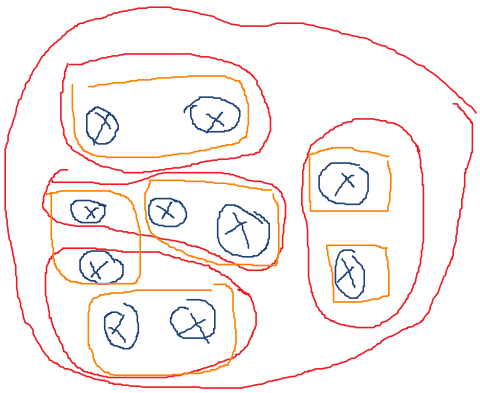
- If WSS value is less, datapoints within a cluster are very close to each other => Homogeneous (clusters are cohesive)
- Each cluster will have its own WSS value
- Total WSS =>Sum of all clusters WSS value.

Between SS - Between Group sum of square -Average distance between the clusters

- If Between SS is bigger => Clusters are well separated
- If you have 20 clusters, how may Between SS value ? Only 1.

How do you decide number of clusters ?

- Input from business
  - 10 marketing strategies, give me 10 clusters
- Business wants to know how many segments exists in the data
  - Data Driven approach
    - Elbow plot, Knee plot or Scree plot
    - Silhouette distance



Once you are done with clustering => Profiling of clusters (tell what is the unique property of each cluster)

It is not mandatory to use all the features as similarity criteria
Example - in your dataset you can have 500 features but only 50 used for clustering

Hierarchical Clustering

Also called as Agglomerative clustering (bottom up approach)

Consider each observation as a cluster in the beginning