Clustering- K Means

Agenda

- Unsupervised Learning - What and Why
- What is Clustering
- K-Means Clustering theory
- K-Means Implementation
- Optimal K
- Advantages and Disadvantages
- Key points

Unsupervised Learning

- Dataset does not have labels
- Target column is not available
- Model takes variables as input and either transforms them into another type of features or a value that can be used to solve practical problems.
- Techniques - Clustering, PCA, Association, GANs, Autoencoder

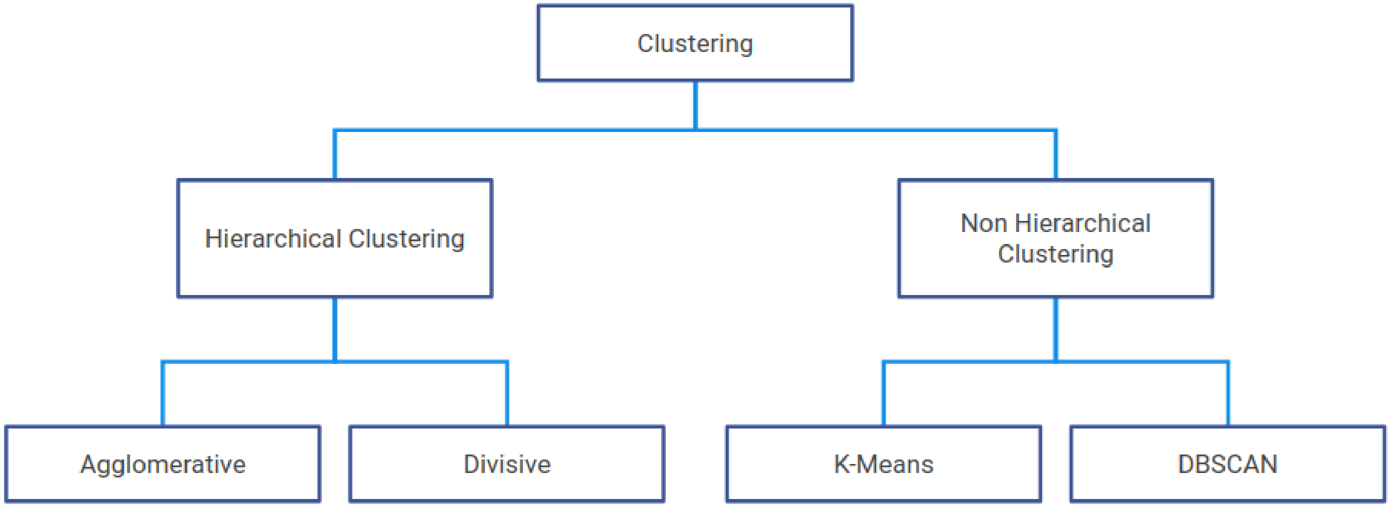
Uses of Unsupervised Learning

- Can be used as an exploratory technique to discover hidden structure and patterns of the data.
- Can be used to decide whether there is a need to separate models representing each cluster.
- Helps is simplifying the data representation
- Can be used for feature engineering through the centroid methods
- Can be used to find useful features for categorization
- Can be used to detect anomalous data points that do not fit into either group
- Used for density estimation in statistics

Clustering

- Used to discovers hidden structure and patterns in uncategorized data.
- Finds natural clusters (groups) existing in the data
- Number of clusters and their granularity can be adjusted

Types of clustering



Applications

- Image Processing
- Medical
- Customer segmentation

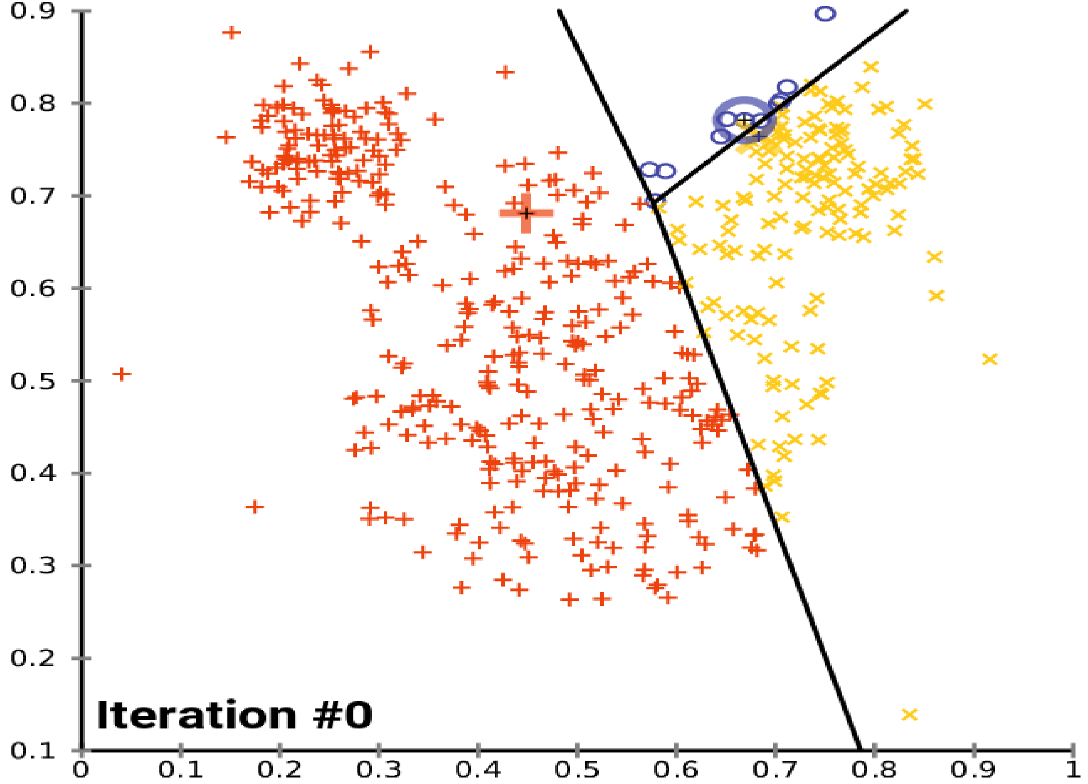
Distance Calculation

- Clustering methods attempts to group the objects based on the definition of similarity specified.
- The definition uses distance calculation for the same
- Lesser the distance, more similar the objects
- Degree of similarity (or dissimilarity) between the data points is a key to achieve the goal of clustering
- Some example of distance calculation are Euclidean distance, Manhattan distance, Jaccard distance, Cosine distance
- Euclidean distance is highly influenced by scale of each variable

K-Means Clustering

- Centroid based model
- Non-hierarchical clustering
- Considers that clusters are disjoint and there is no hierarchical relation between them
- K (number of cluster) should be specified
- K ranges from 1 to n (number of data points)
- Model clusters data into K clusters by segregating data into group of equal variance, minimizing within cluster sum of error (inertia)

K-Means Convergence



K Means Implementation Steps (1/2)

1. We specify a value for K
2. K centroids are randomly computed in the feature space

   Let’s say K =3 then C1, C2, C3 will be three clusters.

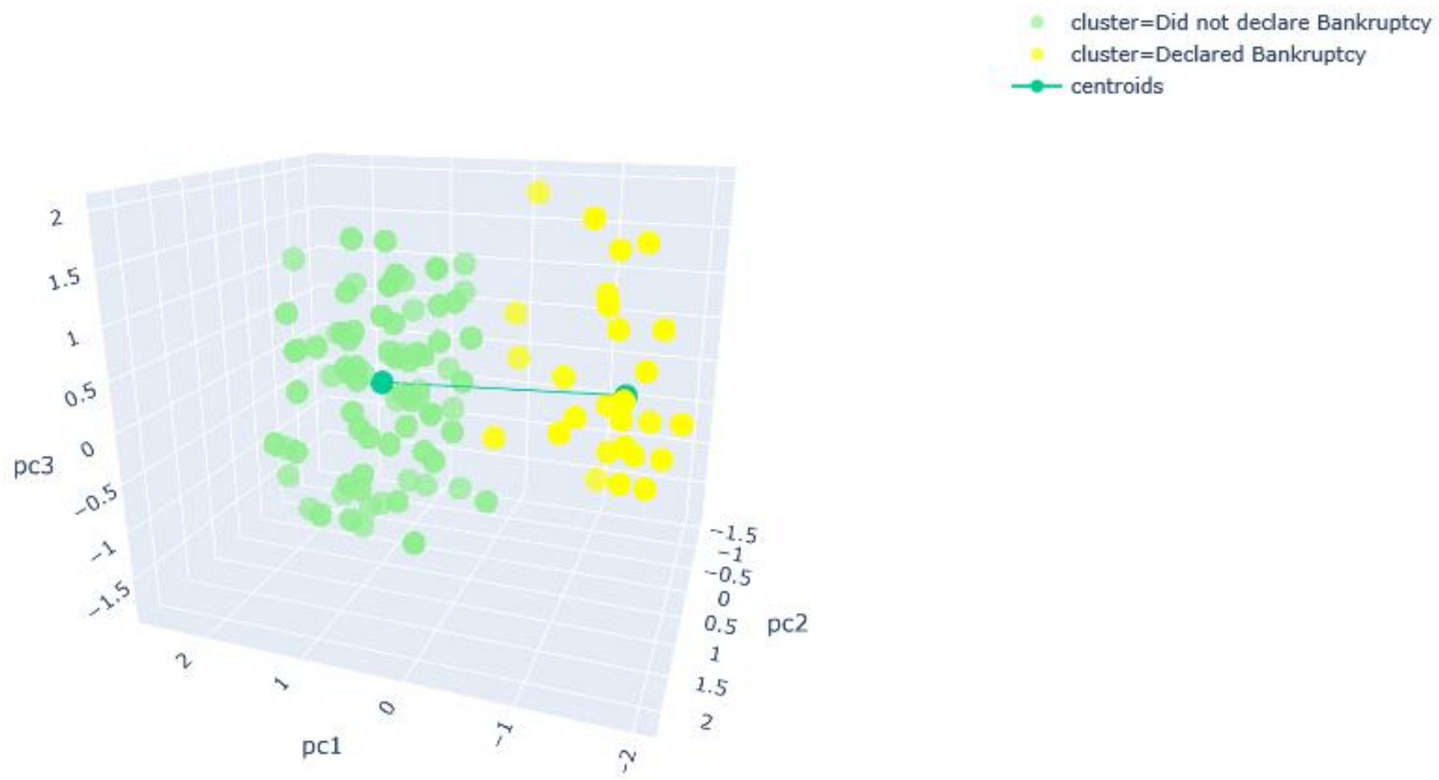
1. The distance from each example x to each centroid c using some distance metric is computed.

   So, if you have 3 clusters and 25 data points then it will return an array of [25,3] distances

K Means Implementation Steps (2/2)

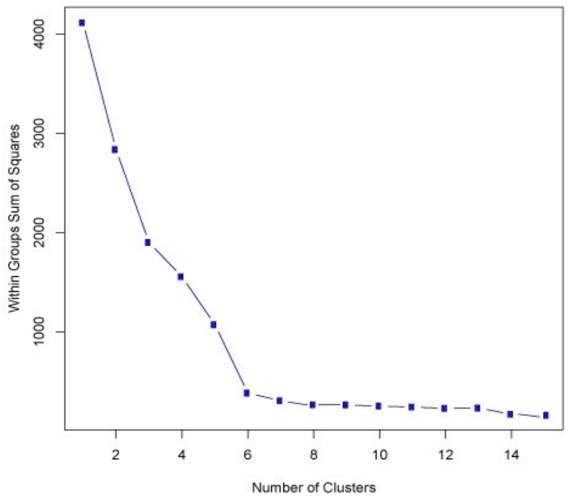
4. The closest centroid is assigned to each example.
    - This step will return an array of 25 numbers which are the distances of 25 points from their nearest cluster
    - The above distances is used to calculate the error/interia
4. New centroid are updated by computing centroids of the previous clusters
5. Above steps are repeated until the assignment of data points do not change after the centroids were recomputed

3D representation of clusters



Optimal K

- Elbow method is used to determine optimal K
- It measures homogeneity or heterogeneity within clusters as the number of clusters change
- One way is to minimize sum of squared errors in each cluster



Silhouette coefficient

- Used to study the separation distance between the resulting clusters
- Displays a measure of how close each point is to its own cluster compared to other clusters
- Range is [-1,1], where higher value indicates that the object is well matched to its cluster i.e. it’s a good fit

Ref - https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Silhouette coefficient

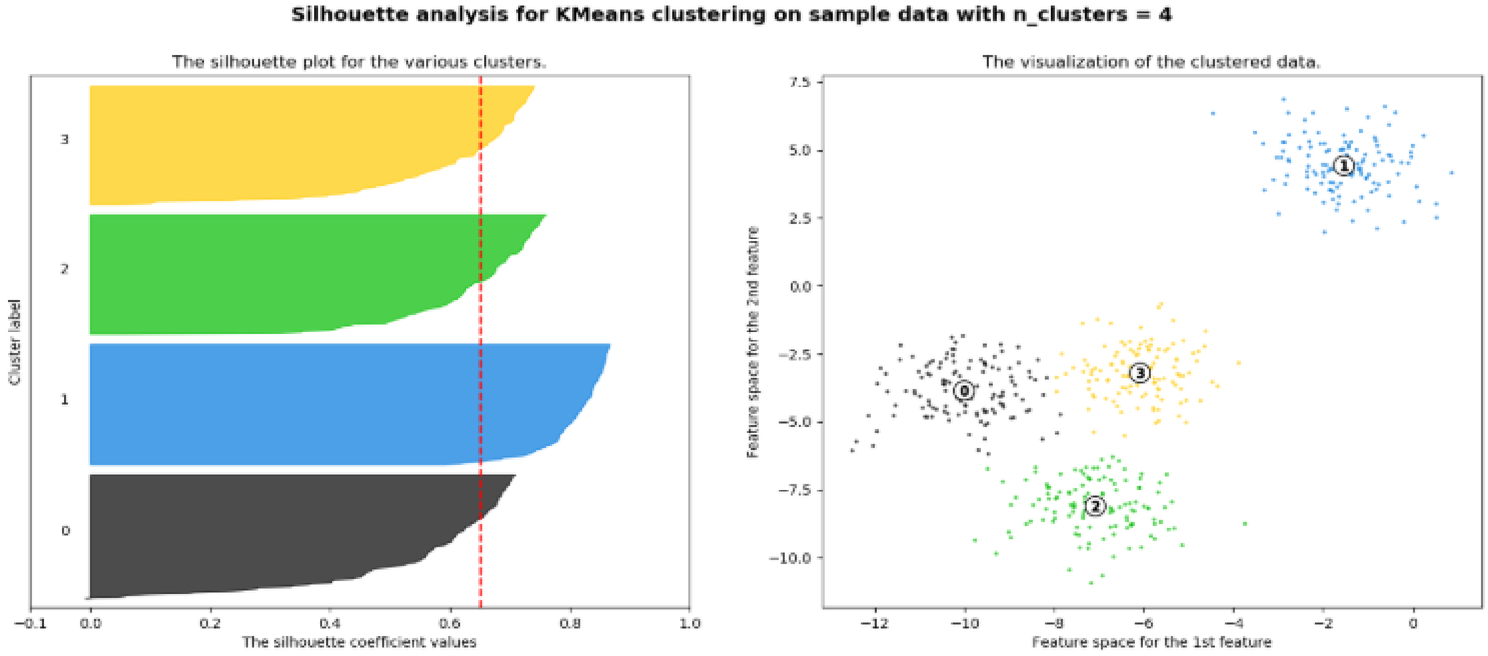


Image Credit- https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Advantages and Disadvantages

Advantages	Disadvantages
● Easy to understand and implement ● K clusters helps us in labelling the data ● Distance calculation is simple ● Helps to eliminate subjectivity from the analysis	● Computationally intensive ● Deciding K can be challenging ● Using correct distance metric can be a challenge ● Scaling is required ● Sensitive to the starting position of initial centroid ● Susceptible to curse of dimensionality

Key points

- Choice of distance measures play a key role in cluster analysis
- Knowledge of the distribution of data will help
- Knowledge about relationship between the attributes will help
- Knowledge about outliers in the data on the various dimension will help
- Euclidean distance is highly scale dependent. Hence standardizing the dimensions is a good practice
- Euclidean distance is sensitive to outliers. If the data has outliers that cannot be handled, use of Manhattan distance is preferred.
- Scikit-learn has implemented K-Means++ initialization scheme, which initializes centroids to be distant to one another