Poojashree NS
Student Id: 015979795

Assignment 1

1. Select your Top 5 concepts, formulas, definitions, etc that you have studied in the greatest detail, explain why they are important.
2. Write the definitions and give an example for each.

Concept 1: Clustering

Clustering is the process of dividing the unsupervised data into similar groups or clusters. Clustering algorithms are key in the processing data and identification of groups. There is no particular criteria to highlight good clustering.

- It enables businesses to approach customer segments differently based on their attributes and similarities, which again helps in profit maximization.
- It can help in dimensionality reduction if the dataset consists of too many variables. Irrelevant features can be identified easily and removed from the dataset.
- It also enable users to identify unusual data object ie., outlier in the dataset.

Example: Consider an E-commerce company that wants to provide discounts for its customers who are less frequent and share the same interest in order to increase customer base. During such use cases, one can perform clustering based on Invoice and product purchased to get the list of customers who are eligible for discounts.

Concept 2: K-means Clustering

K-means is one of the simple and most popular algorithms used for clustering the unsupervised data. K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

K-means is relatively simpler to implement and provides efficient performance even with larger datasets by providing guaranteed convergence. On the other hand, K ie., number of clusters has to be manually updated which affects the performance. To overcome this elbow method is used.

Working of K-means:

- Specify number of clusters K.

- Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids
- Compute the sum of the squared distance between data points and all the centroids.
- Assign each data point to the closer cluster (centroid)
- Compute the centroids for the clusters by taking the average of all the data points that belong to each cluster.

Repeat until there are no changes to the centroids.

Formula:

Euclidean distance : Distance between data points (x1, y1) and (x2, y2) is

$$d = SQRT((x2 - x1)^2 + (y2 - y1)^2)$$

Concept 3: Elbow Method

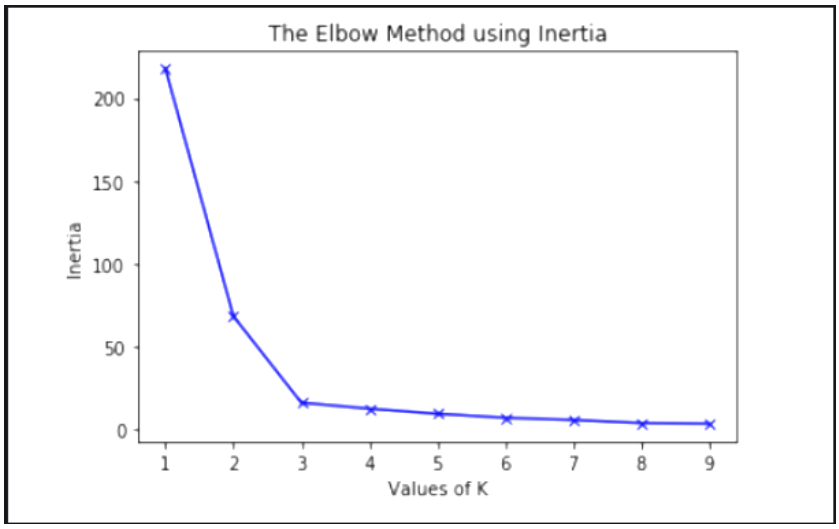
Fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which data may be grouped. The elbow method is used to determine this optimal value of K.

Elbow method internally calculates : Distortion and Inertia

Distortion: It is calculated as the average of the squared distances from the cluster centers of the respective clusters.

Inertia: It is the sum of squared distances of the samples to their closest cluster center.

In this method, for each value of K distortion and Interia will be calculated. The point after which the distortion/inertia starts decreasing in a linear pattern, that point will be considered as K.



Concept 4 : Silhouette Score

Silhouette Coefficient or score is a metric used to calculate the goodness of the clustering technique. Its values from -1 to 1.

1 means clusters are well apart from each other and clearly distinguished.

0 means clusters are indifferent or distance between clusters are not significant.

-1 means clusters are assigned in the wrong way

Formula:

$$Silhouette\ score = (b - a)/max(b,a)$$

Where,

a = average intra-cluster distance between each point within a cluster.

b = average inter-cluster distance i.e the average distance between all clusters**.**

When dealing with higher dimensions, the silhouette score is quite useful to validate the working of clustering algorithms as we can't use any type of visualization to validate clustering when dimensions are greater than 3.

Concept 5 : Rand Index

Rand Index is a way to compare similarities of results between 2 different clustering methods. It can take values between 0 and 1.

0 indicates the two clustering methods do not agree on the clustering of any pair elements.

1 indicates that two clustering methods perfectly agree on the clustering of every pair of elements.

Formula:

$$Rand\ Index = (a+b) / (nC2)$$

Where,

a = The number of times a pair of elements belongs to the same cluster across two clustering methods.

b = The number of times a pair of elements belong to different clusters across two clustering methods.

nC2: The number of unordered pairs in a set of *n* elements.