DATA SCIENCE FOR ENGINEERS

WEEK-8

Distance Measures in KNN

1. What is Hamming distance in KNN?

   Answer:

   It is used for categorical variables. If the value (x) and the value (y) are the same, the distance D will be equal to 0
2. What is Euclidean Distance in KNN?

   Answer:

   Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (y)
3. Is the value of K in KNN required to be odd always?

   Answer:

   Need not be odd always. It can be even as well.

K- Means Clustering

1. What is elbow method?

   Answer:

   The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters.
2. What is WCSS?

   Answer:

   Within-Cluster-Sum-of-Squares (**WCSS**). **WCSS** is the sum of squares of the distances of each data point in all clusters to their respective centroids. The idea is to minimize the sum.

3. In which situation K Means will not perform well?

   Answer:

   K means performs poor

   - K-means is sensitive to outliers in the data set. Because, k-means tries to optimize the sum of squares. And thus, a large deviation (outlier) will get high weightage.
   - The K-means algorithm defines a cost function which computes Euclidean distance (or any other distance function) between two values. Hence, it performs poorly when it tries to calculate mean for categorical variables.

R Questions:

1. What is nstart in K means function?

   Answer:

   The kmeans() function has an nstart option that attempts multiple initial configurations and reports on the best one. For example, adding nstart=25 will generate 25 initial configurations.