WEEK-2 REPORT (PART-2)

DASARI SRINITH (CS21BTECH11015)

kNN:

1. Is Euclidean Distance always the case?

   Solution:
   Although Euclidean distance is mostly used , there are different other ways like Hamming distance , Manhattan distance , etc.

2. Is it optimised to use kNN for large datasets?

   Solution:
   No , it is not . Because in this , we do not calculate any parameters or weights . We need to store all the data , all the time so, it is not recommended.

3. Choosing which values for k can be noisy and will have a higher influence on the result?

   Solution:
   Small values for K

4. Why is KNN algorithm called Lazy Learner?

   Solution: When it gets the training data, it does not learn and make a model, it just stores the data. It does not derive any discriminative function from the training data. So, KNN does not immediately learn a model, but delays the learning, that is why it is called lazy learner.

5. Is kNN used for regression or classification?

   Solution: It is used for both regression and classification.

K-Means:

1. what is k-Means Clustering?

   Solution:
   k-means clustering is a method of vector quantization that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

2. What are few stopping criteria for k-Means Clustering?

   Solution:

   (a) Convergence
   (b) The maximum number of iterations.
   (c) Variance did not improve by at least any 'a'

3. How is Entropy used as a Clustering Validation Measure?

   Solution: Entropy is used as an external validation measure by using the class labels of data as external information. The entropy of a cluster j is calculated by ,

$$E_{ij} = -\sum_i p_{ij} \log p_{ij} \tag{1}$$

4. What is the difference between the Manhattan Distance and Euclidean Distance in Clustering?

   Solution:

   (a) Manhattan distance captures the distance between two points by aggregating the pairwise absolute difference between each variable.
   (b) Euclidean distance captures the distance between two points by aggregating the squared difference in each variable.

5. Name any 2 methods for finding the final value of k.

   Solution:

   (a) Elbow method
   (b) Silhouette Method

Random forest:

1. Does Random Forest need Pruning? Why or why not?

   Solution:
   Random Forest usually does not require pruning because it will not over-fit like a single decision tree. This happens due to the fact that the trees are bootstrapped and that multiple random trees use random features so the individual trees are strong without being correlated with each other.

2. What are Ensemble Methods?

   Solution:
   Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.Random Forest is a type of ensemble method.

3. Few Advantages of using Random Forest.

   Solution:

   (a) It performs better in high dimensionality since the work is on subsets of data.
   (b) It has low bias, but moderate variance
   (c) The training speed is faster than decision trees

4. Difference between Bagging and Random forests is ..

   Solution:
   The difference between Random Forest and Bagging is the fact that for Random Forest only a subset of features out of all are selected in random and the best split feature from the subset is used to split each node in a tree.

5. Do Random Forests need pruning?

   Solution:
   Unlike a tree, no pruning takes place in random forest; i.e, each tree is grown fully. In decision trees, pruning is a method to avoid overfitting.