Rohan Nyati 500075940 (148) Batch-5 (AI & ML)

Assignment - 2

Q.1] Why tree pruning is useful in decision tree induction? What is a drawback of using a seperate set of tuples to evaluate pruning?

Soln] The decision tree built may overfit the training data. Tree pruning addresses this issue of overfitting the data by removing the least reliable branches. This results in more compact & reliable decision tree that is faster and more accurate in its ~~the~~ classification of data

The drawback of using a seperate set of tuples to evaluate pruning is that it may not be representative of the training tuple used to create the original decision tree. Furthermore, using a seperate set of tuples means there are less tuples to use for creation & testing of tree.

Q.2] Why naive Bayesian Classification is known as naive? Briefly outline the major ideas of naive bayes classification.

Soln] Naive Bayesian classification is called naive because it assumes class conditional independence. That is, the effect of an attribute value on a given class is independent of the values of the other attribute. This assumption is made to reduce computation costs and hence is considered as "naive".

The major idea is to try and classify data by maximizing $P(X|C_i)P(C_i)$ using the Bayes theorem of posterior probability.

Q.3] Compare the advantages and disadvantages of eager classification versus lazy classification.

Soln] In Eager classification, eager learners when given a set of training tuples, will construct generalization model before receiving new tuples to classify. We can think of a learned model as being ready to ~~eager~~ eagerly classify previously unseen tuples.

In lazy classification, the learners instead wait until the last minute before doing any model construction in order to classify a given test tuple. That is, when given a training tuple, a lazy learner simply stores it & wait until it is given a test tuple.

Q.4] Use an example to show why the K-means algorithm may not find the global optimum, that is optimizing within-cluster variation.

Soln] K means algorithm does not give optimal results when the clusters are of diff. density & size. If the initial centres are not rightly picked, the solution can get stuck on local maxima instead of going for global maxima or optimum.

The problem of clustering N patterns into M classes may be regarded as a case where K means algo. may not find the global optimum. The K-means algorithm is unable to avoid the clustering results and are trapped into a local optimum.

Q.5] Both K-means and K-medoids algorithms can provide effective clustering. Illustrate the strength & weakness of K-means in comparision with K-medoids.

Sol:] The K-means algorithm is a well known partitioning method for clustering. It takes K as input parameteral partition of a set of n objets from K clusters.
But this algorithm is not effective when used with global cluster. If the density cluster has different size, then this algorithm can't handle this problem.
The K-medoids algorithm is used to find medoid in a cluster which is centre located point of a cluster. It is more robust than K-means as in this we find K as a representative object to minimize the sum of dissimilarities of data objects.
This shows K-medoids is better in all aspects but with the drawback that the complexity is high as compared to K-means.