Name: Dharnidhar Deshmukh

Ref. Code: IIPRSS5306

Assignment (Decision Tree)

1. **What is the main objective of decision tree algorithm?**

   The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

2. **What is homogeneity and how do you measure it?**

   Homogeneity measures how much the sample in a cluster are similar. It is defined using the Shannon's entropy.

   Analysing the Homogeneity of a Dataset:

   1) Calculate the median.
   2) Subtract the median from each value in the dataset.
   3) Count how many times the data will make a run above or below the median (i.e., persistence of positive or negative values).
   4) Use significance tables to determine thresholds for homogeneity.

3. **What are the various split criterias?**
   There are various split criteria in Decision tree 1) Gini impurity: It is a measure of non-homogeneity. It is widely used in classification tree, 2) Information gain: , 3) Sum of squared error.

4. **What is the advantage of Decision tree over linear/logistic regression?**
   1) Decision trees supports non linearity, where LR supports only linear solutions. When there are large number of features with less data-sets(with low noise), linear regressions may outperform Decision trees/random forests. In general cases, Decision trees will be having better average accuracy.
   2) Decision Trees bisect the space into smaller and smaller regions, whereas Logistic Regression fits a single line to divide the space exactly into two. Of course for higher-dimensional data, these lines would generalize to planes and hyperplanes. A single linear boundary can sometimes be limiting for Logistic Regression.

5. **Explain an example where decision tree will fail to perform.**
   1) Decision trees may fail if the data is probabilistic or if the data is noisy. Features in the tree are not weighted and simplicity is hard to control, with overfitting a constant problem while growing the tree.
   2) They can be extremely sensitive to small perturbations in the data: a slight change can result in a drastically different tree.
   3) They can easily overfit. This can be negated by validation methods and pruning, but this is a grey area.
   4) They can have problems out-of-sample prediction (this is related to them being non-smooth).

6. **What is over fitting?**
   Overfitting is an undesirable machine learning behavior that occurs when the machine learning model gives accurate predictions for training data but not for testing data.

7. **Can a regression model overfit, if yes then how?**
   Yes, In regression model overfitting occur frequently. If training accuracy is more than testing accuracy, then a model is overfitted. If we build a complex model for simple problem then it can be overfit.

8. **What is the best use case for decision tree?**
   Decision trees are used to solve classification problems and categorize objects depending on their learning features. They can also be used for regression problems or as a method to predict continuous outcomes from unforeseen data.

9. **Why is decision tree called a greedy technique?**
   It stops splitting once it reaches the maximum depth or cannot find a split to reduce impurity. This approach makes the decision tree a greedy algorithm. it greedily searches for an optimum split at the root node and repeats for the process at each subsequent level.

10. **What is pruning?**
    Pruning is the procedure that decreases the size of decision trees. It can decrease the risk of overfitting by defining the size of the tree or eliminating areas of the tree that support little power.

11. **What's the functioning difference between Decision Tree Regressor and Decision Tree Classifier?**
    Decision Tree classifier is used to solve classification problems. For example, they are predicting if a person will have their loan approved.
    Decision Tree Regressor is used to solve regression problems. For example, prediction of how many people will die because of an opiate overdose.