Om Shete

THADOMAL SHAHANI TSEC ENGINEERING COLLEGE

Assignment 3

Q.1 Compare decision tree classification with K-NN classification.

⇒

K-NN Classification	Decison tree
1) Does not require specific training.	1) Require a proper training phase.
2) Zero learning, that is why called lazy algorithm	2) Builds a model based on training.
3) Data must always be available for making predictions.	3) Once training is completed training data is not needed.
4) Most interpretable algorithm.	4) It is also interpretable since we can see decision making in tree format.
5) Takes much time in decision making, since it have to traverse the whole dataset to calculate the distance with each point.	5) Once training is completed prediction time is less.
6) Since no training is required hence it zero(0)	6) Require initial training time to create decison node and branches.
7) Suitable for small size databases	7) Used for small as well as large size dataset.

Q.2 Compare decision tree classification with logistic regression classification.

Logistic Regression	Decision tree Regression
1) Logistic regression is a type of supervised learning algorithm used for classification tasks where the goal is to model predict probability that an instance belongs to given class or not.	1) A decision tree is a type of supervised learning that is commonly used in ML to model and predict outcomes based on input data.
2) Coefficients can be interpreted to understand the influence of each feature on the outcomes.	2) Highly interpreted, as the tree structure clearly shows the decision made at each node.
3) Less prone to over-fitting compared to decision tree.	3) Prone to overfitting.
4) It is majorly affected by noise.	4) It is majorly affected by noise.
5) Requires a large enough training dataset.	5) Can be trained on small training set.

Q.3 List down the attribute selection measures used by ID3 algorithm to construct a decision tree.

⇒

Attribute Selection Measures :

1) Gini Index
2) Information Gain
3) Entropy
4) Gain Ratio

1) Gini Index

⇒ All attributes are assumed to be continous valued. It is assumed that there exist several possible split values for each attribute.

- Gini index method can be modified for categorical attributes.
- Gini index is used in Classification and Regression Tree (CART).
- If a dataset T contains example from n classes, gini index, gini(T) is defined as -

$$Gini(T) = 1 - \sum_{j=1}^{n} (P_j)^2$$

- In the above equation P_j represents the relative frequency of class j in T.

- After splitting T into two subsets T_1 and T_2 with sizes N_1 and N_2, gini index of split data is,

$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2).$$

2) Information Gain (IG)

⇒ In this method all attributes are assumed to be categorical.

- The amount of information required to decide if a random record in S belongs to P or N is defined as,

$$I(p,n) = -\left(\frac{p}{p+n}\right) \log_2\left(\frac{p}{p+n}\right) - \left(\frac{n}{p+n}\right) \log_2\left(\frac{n}{p+n}\right)$$

- The higher the IG, the better the attribute is at classifying the example.

3) Entropy

⇒ Entropy measures the impurity or disorder in the dataset. It is a measure of the uncertainity in the dataset's classification.

$$Entropy(S) = -\sum_{i=1}^{c} P_i \log_2 P_i$$

- To quantify how mixed or impure a set of data is.
- The higher the entropy, the more uncertain or impure the data is.

c) Gain Ration

⇒ Gain ratio is an extension of IG that takes into account the intrinsic information of a split, penalizing attribute that result into a large number of branches with small sets of data.

$$\text{Gain Ratio}(A) = \frac{\text{Information Gain}(S, A)}{\text{Split Information}(A)}$$

Q.4 Explain the properties of Gini Index

⇒

- The Gini index is a measure used in decision trees to evaluate quality of the split.
- It quantifies the degree of impurity or disorder in a dataset.
- The Gini is used in Classification and Regression Tree (CART).
- If a dataset T contains example from n classes, gini index, gini(T) is defined as-

$$\text{Gini}(T) = 1 - \sum_{j=1}^{n} (p_j)^2$$

- In the above equation, P_j represents the relative frequency of class j in T.
- After splitting T in two subsets T_1 and T_2 with sizes N_1 and N_2, gini index of split data is :

$$gini_{(split)} = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)$$

- The attribute with smallest $gini_{(split)}(T)$ is selected to split the node.
- The Gini index ranges from 0 to 0.5.
- 0 indicates perfect purity, meaning all the instances belong to a single class.
- 0.5 represents maximum impurity, indicating that the instances are uniformly distributed across all classes.

Q.3 Discuss in brief pruning in decision tree.

⇒ - Pruning in a decision tree is a technique used to reduce the size of the decision tree by removing sections of tree that provide little to no additional predictive power.

- The main goal of pruning is to prevent overfitting and improve models ability to generalize to unseen data.
- Decision trees can grow very large and complex, capturing noise and outliers in the training data, which leads to overfitting.
- Pruning helps this mitigate by simplifying the tree.
- By removing unnecessary branches, pruning reduces the complexity of the model, leading to better performance on new, unseen data.
- There are two main types of pruning in decision tree :-

1) Pre-pruning

⇒ This involves halting the growth of the tree, before it becomes too complex.

- It sets a limit on tree depth, maximum samples per leaf, or minimum information gain required for a split.

2) Post-pruning

⇒ This involves first allowing the tree to grow fully and then pruning back unnecessary branches.

The pruning techniques are:

1) Cost-complexity pruning

⇒ This method assigns a score to each subtree primarily based on its accuracy and complexity, then selects the subtree with the lowest score.

2) Reduced Error Pruning

⇒ Removes the branches that do not significantly affect the overall accuracy.

27/9/24