Weekly Report 1 - Decision Tree

Ganji Varshitha
AI20BTECH11009

Introduction

It is a supervised classification and regression ml algorithm which is non parametric. It is a hierarchical model and composed of internal decision nodes and terminal nodes.

Algorithm

The algorithm follows divide and conquer strategy as a test is applied to the input at each node and a branch is selected based on the outcome of the test. Since the rules are in terms of If Else statements it has high interpret-ability. There are two types of trees :

- Univariate trees : It uses single input dimension for split and for numeric attribute it results in binary split where as discrete attribute it results in multi-way split.
- Multivariate trees: It can use multiple attributes for splitting.

Goodness of the split is quantified by impurity measure of the node. If the node is pure, all the instances in that node will belong to the same class.
For node m, N_m instances reach m, N_m^i belong to C_i

$$\hat{P}(C_i|\mathbf{x}, m) \equiv p_m^i = \frac{N_m^i}{N_m} \tag{1}$$

$\therefore$ If the node is pure, p_m^i should be equal to 1 or 0. Impurity is measured by entropy. Entropy of a node is given by

$$I_m = -\sum_{i=1}^{K} p_m^i \log_2 p_m^i \tag{2}$$

where K is number of classes of output.
If the node is not pure, we recursively split to decrease the impurity of the node.
N_{mj} of N_m take branch j. N_{mj}^i belong to C_i

$$\hat{P}(C_i|\mathbf{x}, m, j) \equiv p_{mj}^i = \frac{N_{mj}^i}{N_{mj}} \tag{3}$$

$$I_m^{'} = -\sum_{j=1}^{n} \frac{N_{mj}}{N_m} \sum_{i=1}^{K} p_{mj}^i \log_2 p_{mj}^i \tag{4}$$

$I_m^{'}$ is the expected reduction in impurity after split which is known as **Information Gain**.
$\therefore$ We need to chose attribute which maximises the information gain at each split.
Other common measure of impurity is gini index.

$$\text{Gini index } = 1 - \sum_{j=1}^{c} p_j^2 \tag{5}$$

Key points

- The algorithm does not work well for linear data but if the relationship between input variable and target is non linear or complex, it will outperform the linear models.
- The model is highly prone to over-fitting and sensitive to outliers like noisy training instances.
- It also suffers excessive generalization error.

Unique points

- It is a discriminative model.
- Does not require normalization of inputs
- **Learning is greedy!** It splits to get maximum information gain recursively.
- Training and predicting the data is fast as we only store the parameters of node to split instead of entire training data.

Pruning

To reduce overfitting, we prune the decision tree.It can be done in 2 ways:

- Pre-pruning : We declare max depth and stop the node to split and make it a leaf node.
- Post-pruning: Pruning the sub trees after generating the whole tree. One common way is to prune the lower ends of tree that result in least information gain.

Figure 1: Algorithm to generate decision tree

```
GenerateTree(X)
    If NodeEntropy(X)< θ_I
      Create leaf labelled by majority class in X
      Return
    i ← SplitAttribute(X)
    For each branch of x_i
      Find X_i falling in branch
      GenerateTree(X_i)

SplitAttribute(X)
    MinEnt← MAX
    For all attributes i = 1,...,d
        If x_i is discrete with n values
          Split X into X_1,...,X_n by x_i
          e ← SplitEntropy(X_1,...,X_n)
          If e<MinEnt MinEnt ← e; bestf ← i
        Else /* x_i is numeric */
          For all possible splits
              Split X into X_1,X_2 on x_i
              e←SplitEntropy(X_1,X_2)
              If e<MinEnt MinEnt ← e; bestf ← i
    Return bestf
```

Questions

1. Pick the correct choice.

   A. Pre-pruning is faster, post-pruning is accurate

   B. Post-pruning is faster, pre-pruning is accurate

   Solution: A. Pre-pruning is faster, post-pruning is accurate

2. Information gain __________ at each node.

   A. Maximised

   B. Minimised

   Solution: A. Maximised

3. Let $\phi(p_1, p_2)$ be the function measuring impurity of a split. Which of the following is false?

   A. $\phi(0,1) = \phi(1,0) = 0$

   B. $\phi(1/2, 1/2) \geq \phi(p, 1-p)$, p $\in$ [0,1]

   C. $\phi(p, 1-p)$ is decreasing in p on [0,1/2]

   D. $\phi(p, 1-p)$ is decreasing in p on [1/2,1]

   Solution: C. $\phi(p, 1-p)$ is decreasing in p on [0,1/2]

4. Do we require normalization of features while training the model?
   A. Yes B. No
   Solution: B. No

5. If the measure of impurity is gini index,the feature with __________ gini index is selected.
   A. Highest   B. Least
   Solution: B. Least