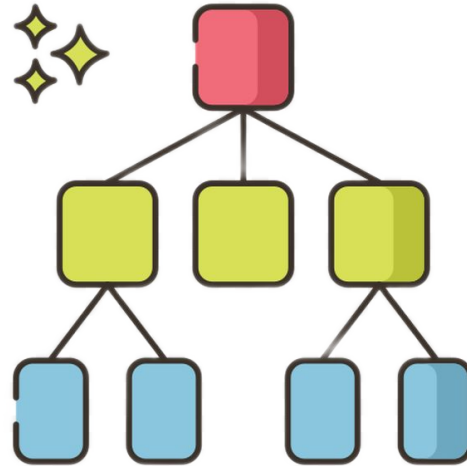
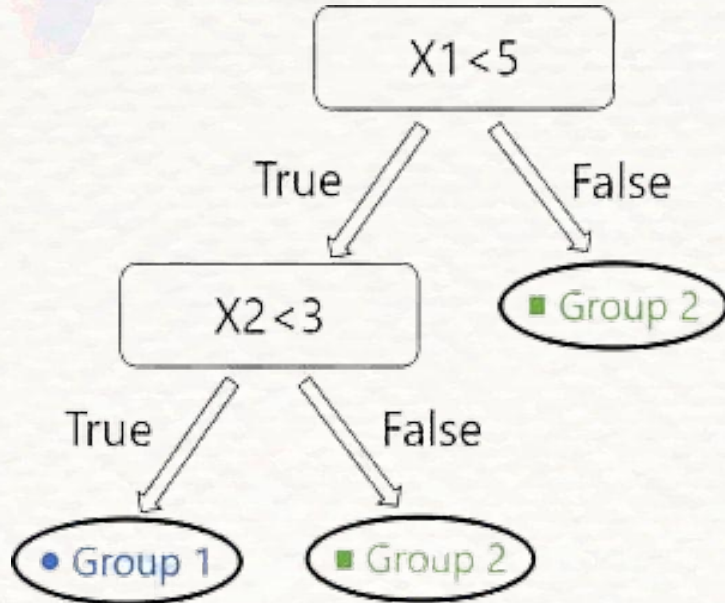


DECISION TREE PRUNING



Decision Tree:

A Decision Tree is a tree with nodes representing deterministic decisions based on variables and edges representing path to next node or a leaf node based on the decision.





Disadvantages of Decision Tree:

- In a Decision Tree, the choice is made to optimize the decision at each of the nodes. Choosing best result at each step does not result in global optimal result.
- They are prone to Overfitting, i.e., if the tree is deep, the number of samples considered at each decision becomes small.
- As the number of factors being considered increases, the data points available for the combination are not significant to assign sensible probability for outcome.

Stopping Criteria:

- All instances in the training set belong to a single value of y .
- The maximum tree depth has been reached.
- The number of cases in the terminal node is less than the minimum number of cases for parent nodes.
- If the node were split, the number of cases in one or more child nodes would be less than the minimum number of cases for child nodes
- The best splitting criteria is not greater than a certain threshold



Pruning Trees:


Using loose stopping criteria tends to generate large decision trees that are overfitted to the training set.

The overfitted tree is cut back into a smaller tree by removing sub-branches that are not contributing to the generalization accuracy. Pruning methods can improve the generalization performance of a decision tree, especially in noisy domains.





Some Heuristic Pruning Techniques:

- Cost Complexity Pruning
 - Reduced Error Pruning
 - Minimum Error Pruning (MEP)
 - Pessimistic Pruning
 - Error-Based Pruning etc.
- 



Cost Complexity Pruning:

Also known as Weakest Link Pruning or Error Complexity Pruning.

It proceeds in two stages:-

1. In the first stage, a sequence of trees T_0, T_1, \dots, T_k is built on the training data where T_0 is the original tree before pruning and T_k is the root tree.
2. In the second stage, one of these trees is chosen as the pruned tree, based on its generalization error estimation.

Note:- The tree T_{i+1} is obtained by replacing one or more of the sub-trees in the predecessor tree T_i with suitable leaves.

Calculation behind CCP:

This algorithm is parameterized by $\alpha(\geq 0)$ known as the complexity parameter.


$$\alpha = (\epsilon(\text{pruned}(T, t), S) - \epsilon(T, S)) / (|\text{leaves}(T)| - |\text{leaves}(\text{pruned}(T, t))|)$$

Where,

$\epsilon(T, S)$ indicates the error rate of the tree T over the sample S ,

$|\text{leaves}(T)|$ denotes the number of leaves in T ,

$\text{pruned}(T, t)$ denotes the tree obtained by replacing the node t in T with a suitable leaf.



The complexity parameter is used to define the cost-complexity measure $R\alpha(T)$ of a given tree T :-

$$R\alpha(T) = R(T) + \alpha|T|$$

Where,

$|T|$ is the number of terminal nodes in T ,

$R(T)$ is the total misclassification rate of the terminal nodes.

In the second phase, the generalization error of each pruned tree is estimated. The best pruned tree is then selected.



Dataset Division and Code:

If the given dataset is large enough, it is suggested to break it into a training set and a pruning set. The trees are constructed using the training set and evaluated on the pruning set. On the other hand, if the given dataset is not large enough cross-validation methodology can be used, despite the computational complexity implications.

Sample implementation Code Link:-

<https://github.com/g-4-gagan/MSc-Sem-1/blob/master/Data%20Mining/Cost%20Complexity%20Pruning.ipynb>

References:

- Chapter 7, Lior Rokach and Oded Maimon. 2014. Data Mining With Decision Trees: Theory and Applications (2nd. ed.). World Scientific Publishing Co., Inc., USA <https://doi.org/10.1142/9097>.
- Rebala G, Ravi A, Churiwala S (2019) An introduction to machine learning. Chapter 7) in book. https://doi.org/10.1007/978-3-030-15729-6_7
- <https://www.analyticsvidhya.com/blog/2020/10/cost-complexity-pruning-decision-trees/>



Submitted By:
GAGAN KUMAR SONI
Roll : 21