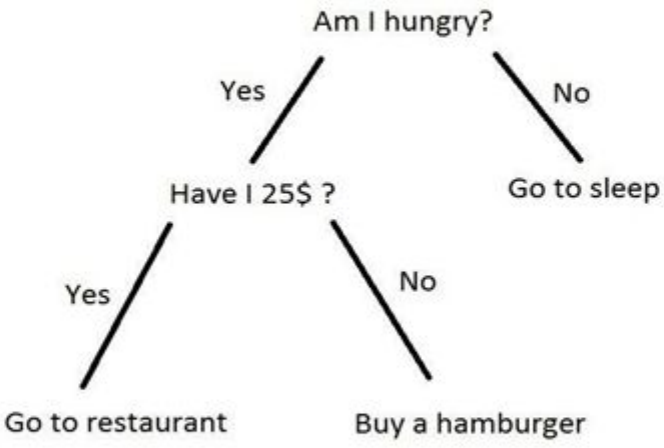
Decision Trees

Discussion questions

1. What is a Decision Tree and how does it work?
2. What are the common terminologies used in a Decision Tree?
3. What is Pruning? How does it avoid overfitting?
4. How to read python’s visual output of a Decision Tree?
5. What are the different ways to prune a tree? How to perform Pruning?

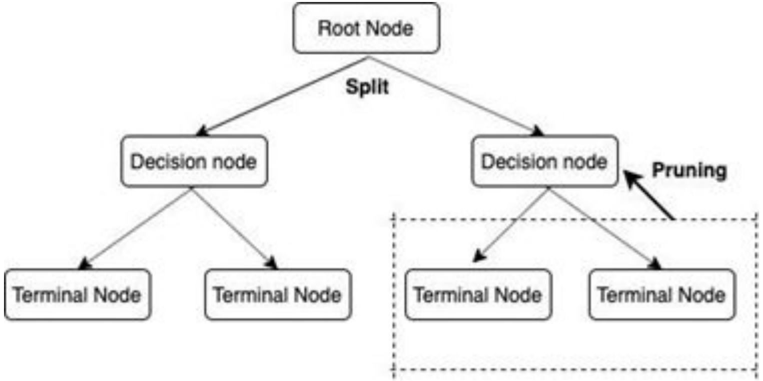
What is a decision tree and how does it work?

- A decision tree is one of the most popular and effective supervised learning techniques for classification problems that works well with both categorical and continuous variables.
- It is a graphical representation of all the possible solutions to a decision that is based on a certain condition.
- In this algorithm, the training sample points are split into two or more sets based on the split condition over input variables.
- A simple example of a decision tree can be - A person has to take a decision for going to sleep or restaurant based on parameters like he is hungry or has 25$ in his pocket.



What are the common terminologies used in decision trees?

1. **Root node -** Represent the entire set of the population which gets further divided into sets based on splitting decisions.
2. **Decision node -** These are the internal nodes of the tree, These nodes are expressed through conditional expression for input attributes.
3. **Leaf node/Terminal node -** Nodes that do not split further are known as leaf nodes or terminal nodes.
4. **Splitting -** The process of dividing a node into one or more sub-nodes.
5. **Pruning -** It is the reverse process of splitting where the sub-nodes are removed.

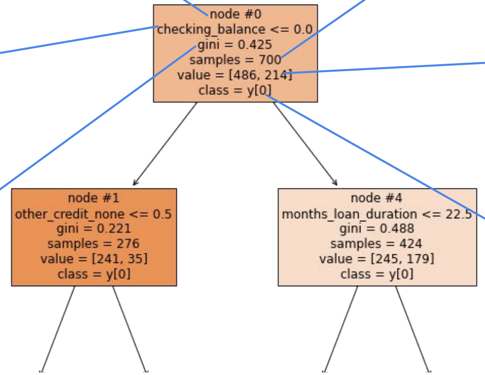


How to read python’s visual output of a decision tree?

Node number : Represents the number of node

Decision Rule : checking _balance <= 0 means that every observation with a checking balance of 0 and less will follow the True arrow i.e. go to the left node in the next level, and the rest will follow the False arrow that is going to the right node in the next level.

GINI: Refers to the quality of the split. Always a number between 0.0 and 0.5,
0.0 -> High purity in the node
0.5 -> High impurity in the node.



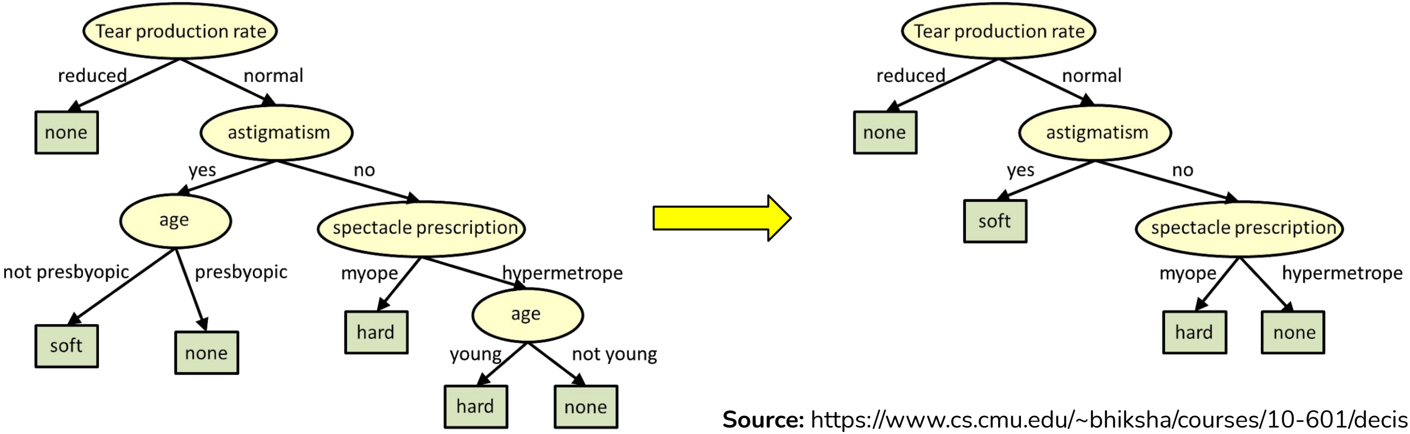
Samples: 424 means that there are 424 observations in this node.

Value: List that tells how many samples at the given node fall into each category. value = [486,214] means that of these 700 observations at this level 486 belong to class 0 and 214 to class 1 as the
Class = y[0] represents the majority class 0 in that node.

Color intensity represents the concentration of the majority class in a node. Darker the color higher the concentration of the majority class.

What is pruning? How does it avoid overfitting?

- One of the problems with the decision tree is it gets easily overfits the training data and becomes too large and complex.
- A complex and large tree poorly generalizes on new data whereas a small tree fails to capture the information of training sample data.
- Pruning shortens the branches of the tree. The process of reducing the size of the tree by turning some branch node into a leaf node and removing the leaf node under the original branch.
- By removing branches we can reduce the complexity of the tree which helps in reducing the overfitting of the tree.



Source: https://www.cs.cmu.edu/~bhiksha/courses/10-601/decisiontrees/

What are the different ways to prune a tree? How to perform pruning?

- There are two different ways to prune a tree:
- **Pre-Pruning**: Early stopping of a tree before it has fully grown.
- **Post-Pruning**: Building a full tree and removing the sub-tree nodes.

- To perform pre-pruning - Hyper-parameter tuning is one of the ways to tune different hyperparameters of a tree and restrict the growth of the tree by limiting the hyperparameters such as max_depth,min_samples_split, etc.

- To perform post-pruning - It can be done using the cost complexity pruning technique by choosing the appropriate value of ccp_alpha and removing the less significant sub-tree nodes.

greatlearning

Power Ahead

Happy Learning !