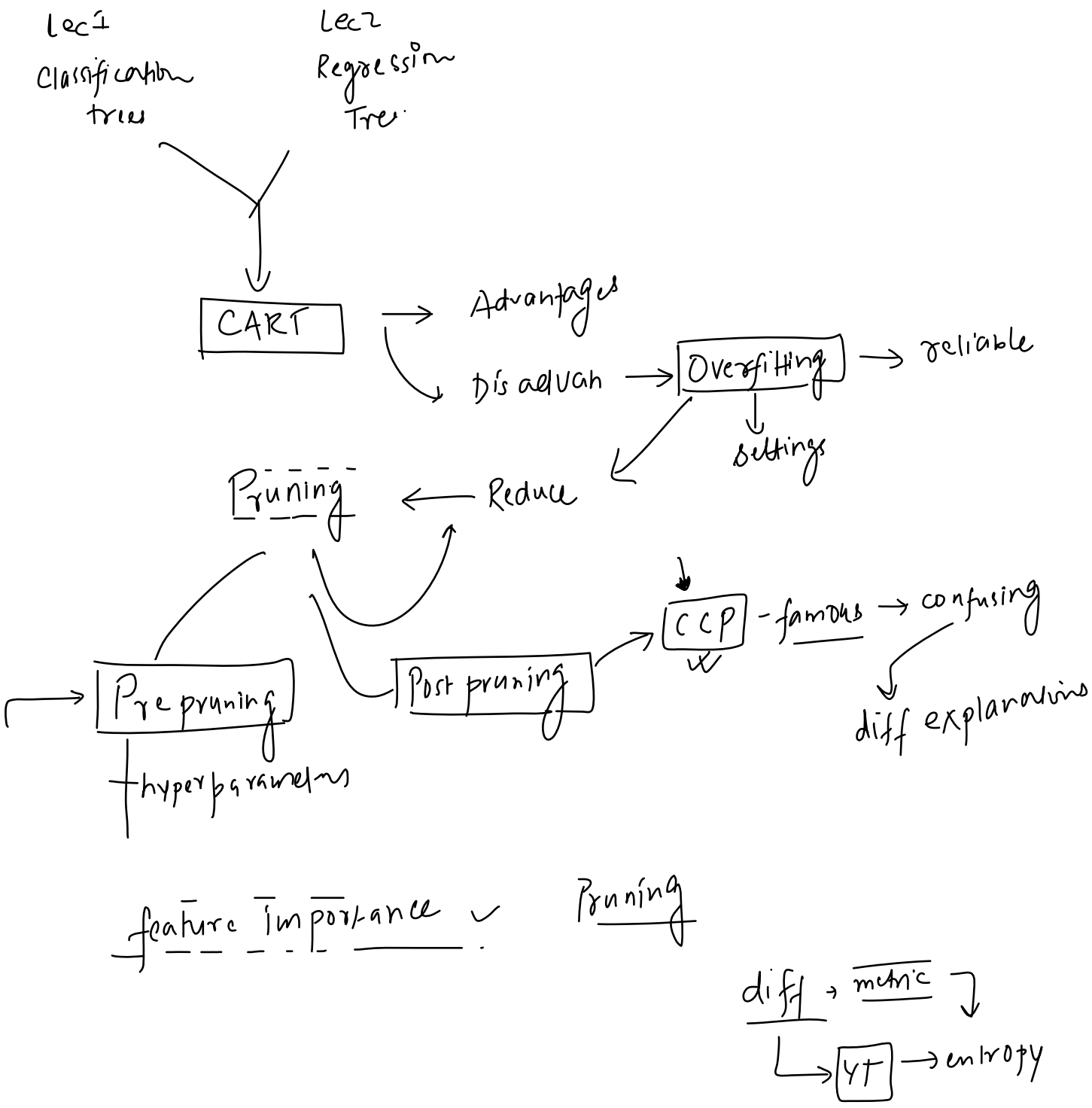
Recap

26 July 2023 15:14

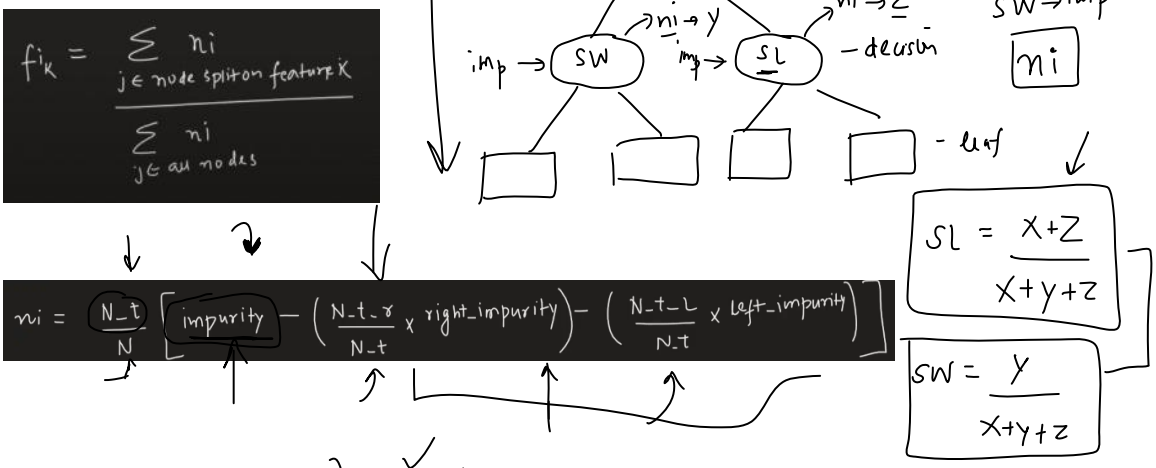


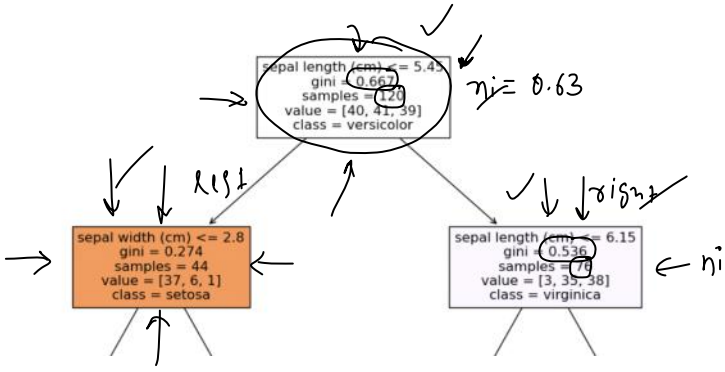
Feature Importance

27 July 2023 07:43

SL | SW | ni

The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance.

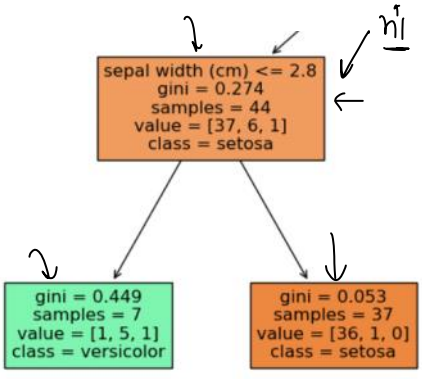




$\frac{120}{120}\left[0.667 - \left(\frac{76}{120} \times 0.536\right) - \left(\frac{44}{120} \times 0.27\right)\right]$

sepal length = $\frac{x+y}{}$

$SW = \frac{z}{x+y+z}$



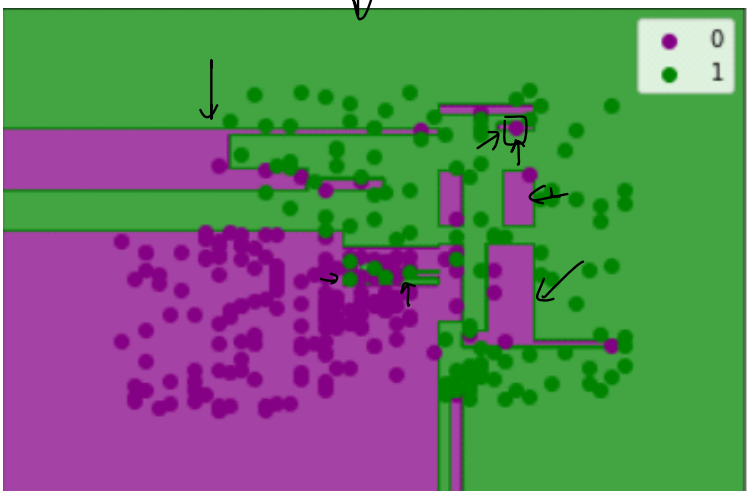
$\frac{44}{120}\left[0.274 - \left(\frac{7}{44} \times 0.449\right) - \left(\frac{37}{44} \times 0.05\right)\right]$

$ni = \frac{N\_t}{N}\left[\text{impurity} - \left(\frac{N\_t\_r}{N\_t} \times \text{right\_impurity}\right) - \left(\frac{N\_t\_l}{N\_t} \times \text{left\_impurity}\right)\right]$

The Problem of Overfitting

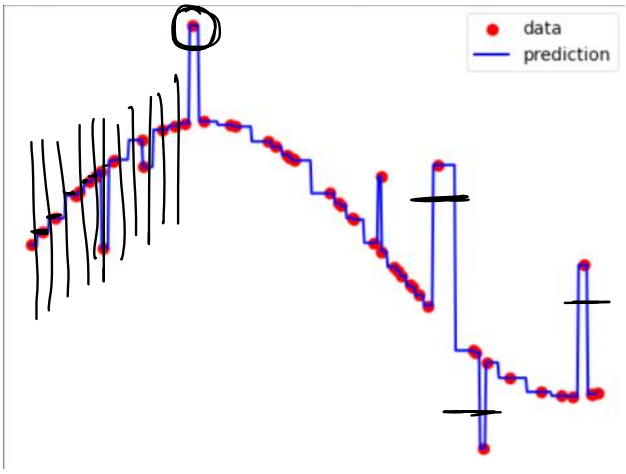
26 July 2023 15:15

test



Classification

Pruning

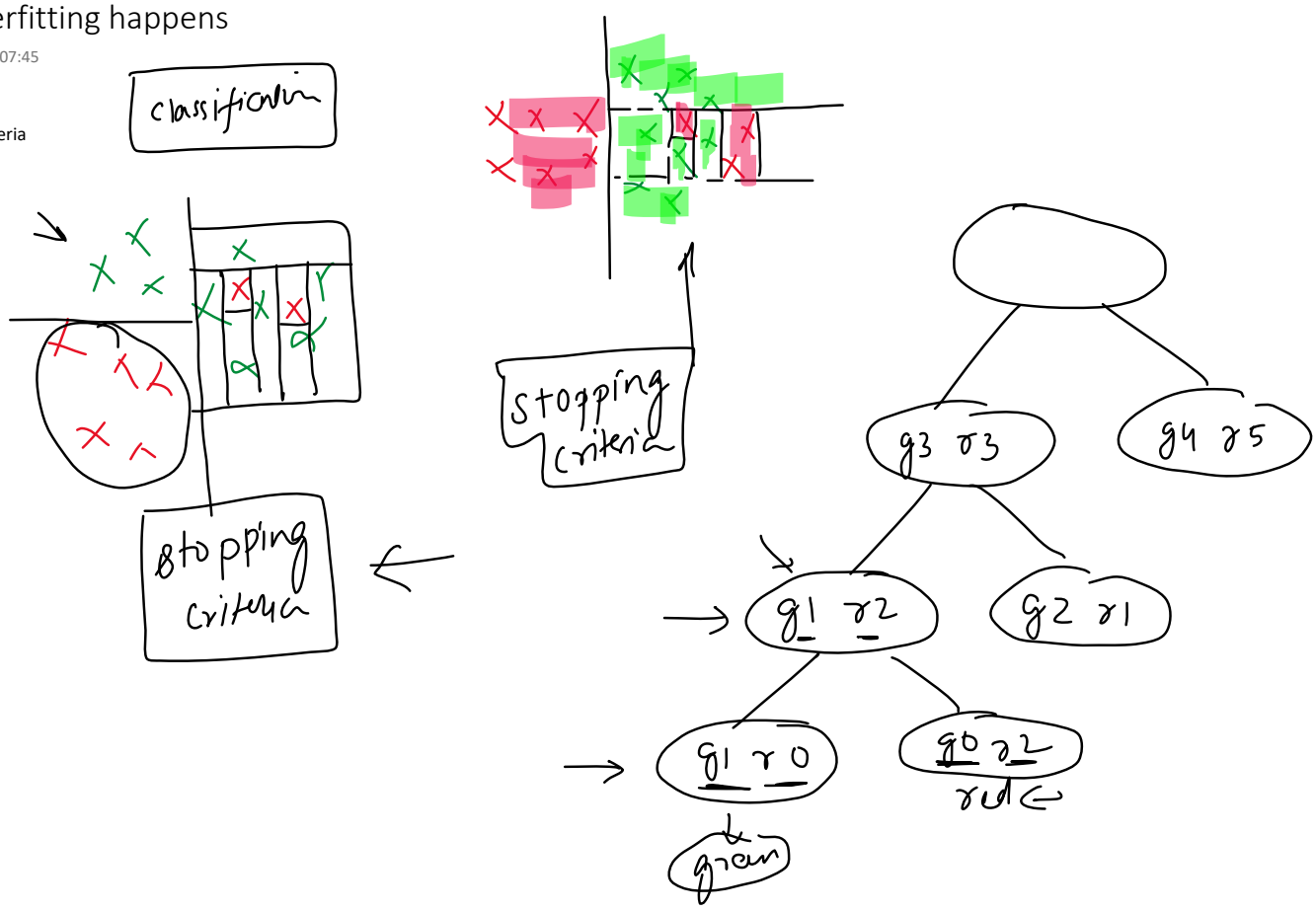


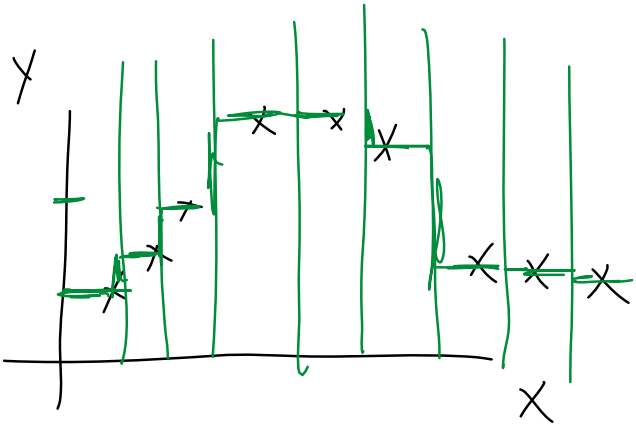
Regression

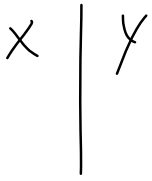
Why Overfitting happens

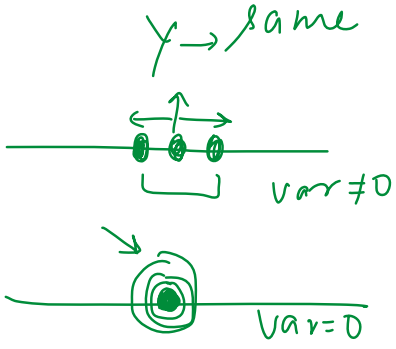
27 July 2023 07:45

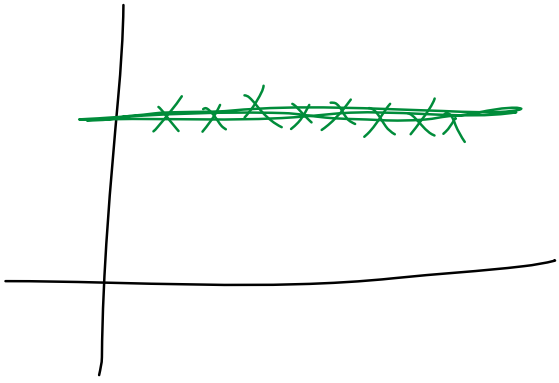
-> Stopping criteria





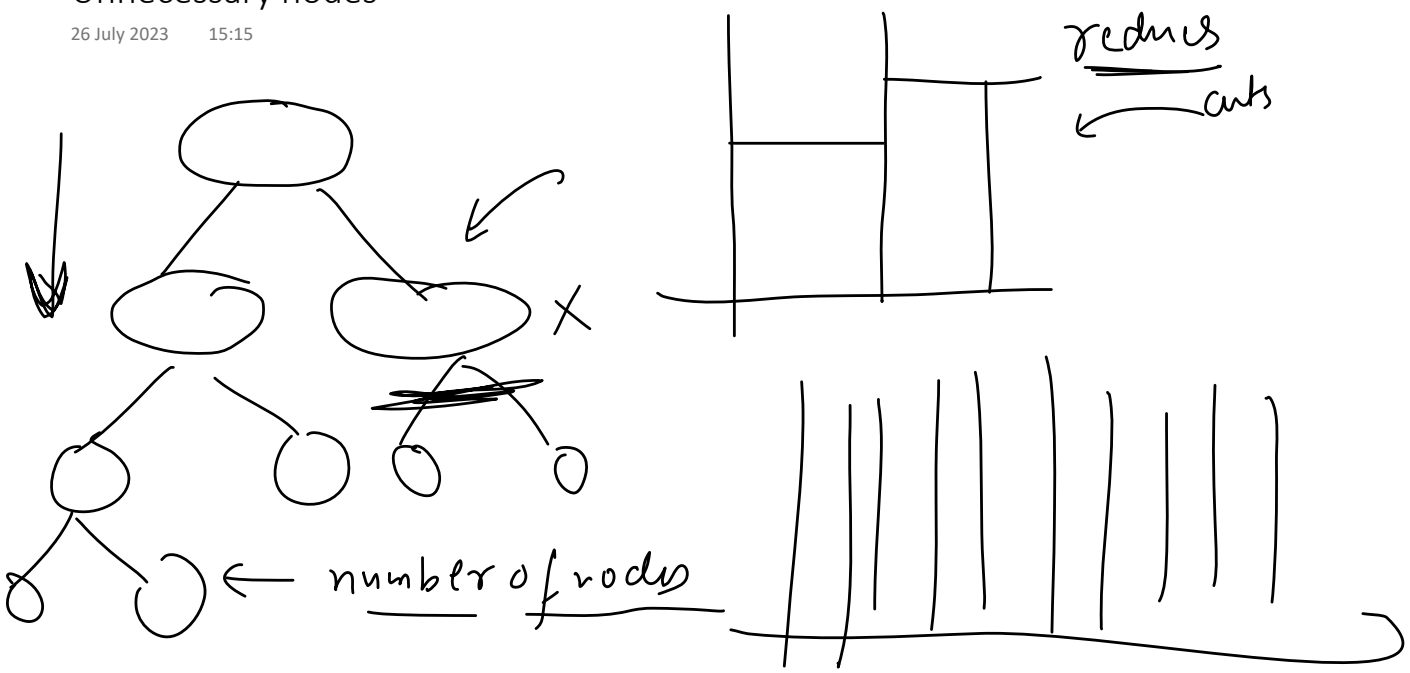


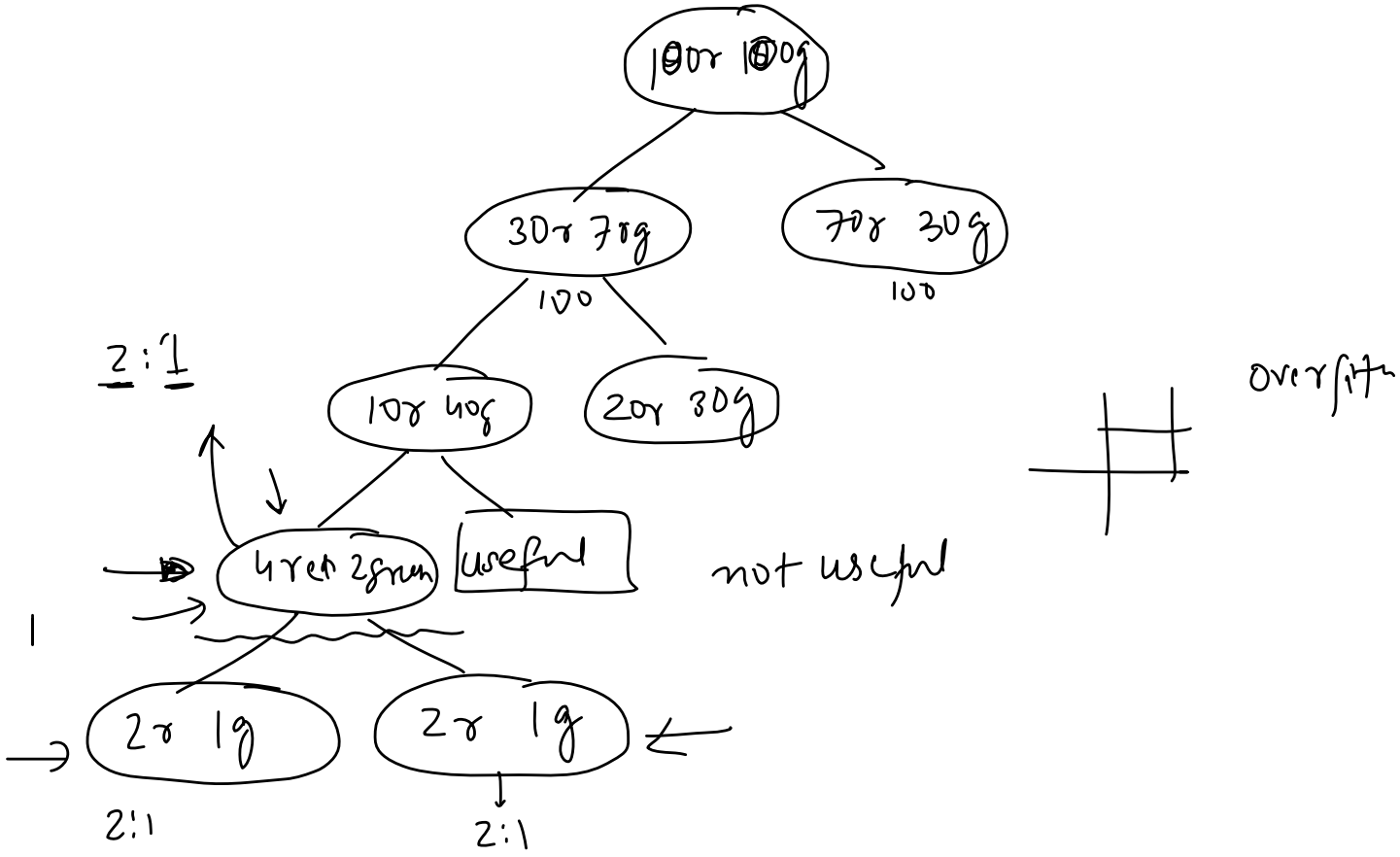




Unnecessary nodes

26 July 2023 15:15





dt → overfitting → stopping criteria

↓

deep tree (a lot of nodes)

↓

(not useful) → cut those nodes

Pruning

Pruning

(not useful) → cut those

↓

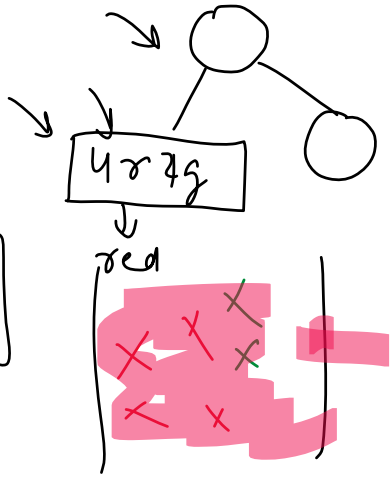
true size reduce ← reduce overfitting

Pruning & it's types

26 July 2023 15:15

Pruning is a technique used in machine learning to reduce the size of decision trees and to avoid overfitting. Overfitting happens when a model learns the training data too well, including its noise and outliers, which results in poor performance on unseen or test data.

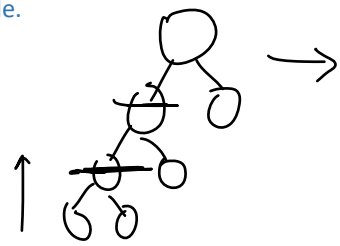
Decision trees are susceptible to overfitting because they can potentially create very complex trees that perfectly classify the training data but fail to generalize to new data. Pruning helps to solve this issue by reducing the complexity of the decision tree, thereby improving its predictive power on unseen data.



There are two main types of pruning: pre-pruning and post-pruning.

1. Pre-pruning (Early stopping): This method halts the tree construction early. It can be done in various ways: by setting a limit on the maximum depth of the tree, setting a limit on the minimum number of instances that must be in a node to allow a split, or stopping when a split results in the improvement of the model's accuracy below a certain threshold.

2. Post-pruning (Cost Complexity Pruning): This method allows the tree to grow to its full size, then prunes it. Nodes are removed from the tree based on the error complexity trade-off. The basic idea is to replace a whole subtree by a leaf node, and assign the most common class in that subtree to the leaf node.



Pre-pruning

26 July 2023 15:27

Pre-pruning, also known as early stopping, is a technique where the decision tree is pruned during the learning process as soon as it's clear that further splits will not add significant value. There are several strategies for pre-pruning:

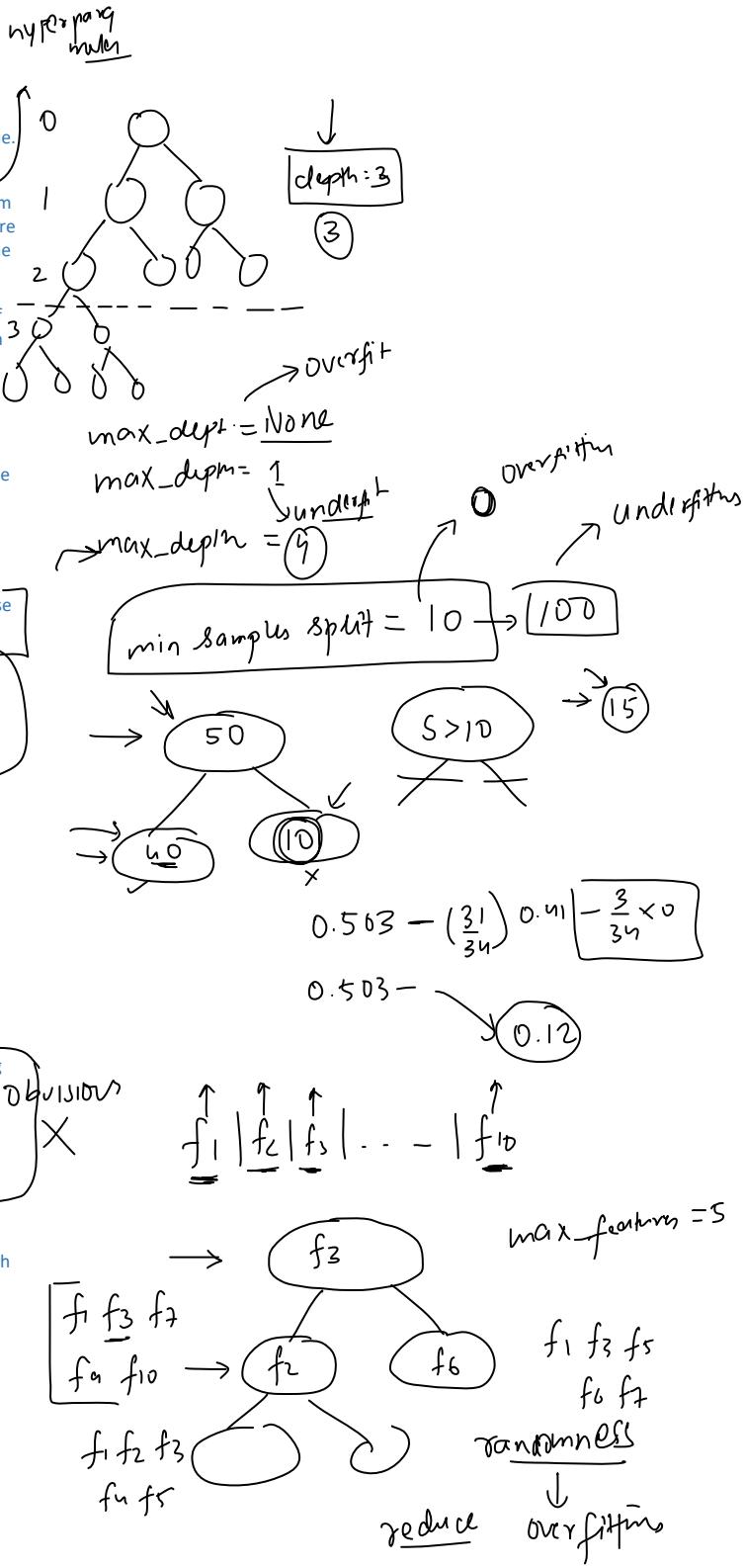
1. Maximum Depth: One of the simplest forms of pre-pruning is to set a limit on the maximum depth of the tree. Once the tree reaches the specified depth during training, no new nodes are created. This strategy is simple to implement and can effectively prevent overfitting, but if the maximum depth is set too low, the tree might be overly simplified and underfit the data.
2. Minimum Samples Split: This is a condition where a node will only be split if the number of samples in that node is above a certain threshold. If the number of samples is too small, then the node is not split and becomes a leaf node instead. This can prevent overfitting by not allowing the model to learn noise in the data.
3. Minimum Samples Leaf: This condition requires that a split at a node must leave at least a minimum number of training examples in each of the leaf nodes. Like the minimum samples split, this strategy can prevent overfitting by not allowing the model to learn from noise in the data.
4. Maximum Leaf Nodes: This strategy limits the total number of leaf nodes in the tree. The tree stops growing when the number of leaf nodes equals the maximum number.
5. Minimum Impurity Decrease: This strategy allows a node to be split if the impurity decrease of the split is above a certain threshold. Impurity measures how mixed the classes within a node are. If the decrease is too small, the node becomes a leaf node.
6. Maximum Features: This strategy considers only a subset of features for deciding a split at each node. The number of features to consider can be defined and this helps in reducing overfitting.

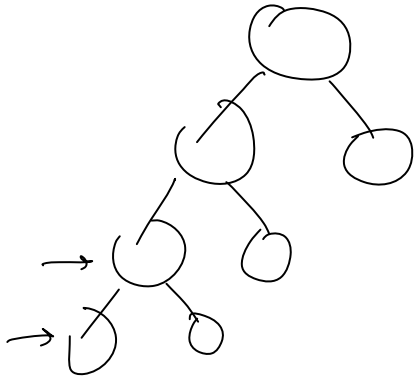
Advantages of Pre-Pruning:

1. Simplicity: Pre-pruning criteria such as maximum depth or minimum number of samples per leaf are easy to understand and implement.
2. Computational Efficiency: By limiting the size of the tree, pre-pruning can substantially reduce the computational cost of training and prediction.
3. Reduced Overfitting: By preventing the tree from becoming overly complex, pre-pruning can help avoid overfitting the training data and thereby improve the model's generalization performance.
4. Improved Interpretability: Simpler trees (with fewer nodes) are often easier for humans to interpret.

Disadvantages of Pre-Pruning:

1. Risk of Underfitting: If the stopping criteria are too strict, pre-pruning can halt the growth of the tree too early, leading to underfitting. The model may become overly simplified and fail to capture important patterns in the data.
2. Requires Fine-Tuning: The pre-pruning parameters (like maximum depth or minimum samples per leaf) often require careful tuning to find the right balance between underfitting and overfitting.
3. Short Sightedness - Can prune good nodes if they come after a bad node.





Post Pruning

26 July 2023 15:31

Post-pruning, also known as backward pruning, is a technique used to prune the decision tree after it has been built. There are several strategies for post-pruning:

1. Cost Complexity Pruning (CCP): Also known as Weakness Pruning, this technique introduces a tuning parameter (α) that trades off between tree complexity and its fit to the training data. For each value of α, there is an optimal subtree that minimizes the cost complexity criterion. The subtree that minimizes the cost complexity criterion over all values of α is chosen as the pruned tree.

2. Reduced Error Pruning: In this method, starting at the leaves, each node is replaced with its most popular class. If the accuracy is not affected in the validation set, the change is kept.

Advantages of Post-Pruning:

1. Reduced Overfitting: Post-pruning methods can help to avoid overfitting the training data, which can lead to better model generalization and thus better performance on unseen data.

2. Preserving Complexity: Unlike pre-pruning, post-pruning allows the tree to grow to its full complexity first, which means it can capture complex patterns in the data before any pruning is done.

3. Better Performance: Post-pruned trees often outperform pre-pruned trees, as they are able to better balance the bias-variance trade-off.

Disadvantages of Post-Pruning:

1. Increased Computational Cost: Post-pruning can be more computationally intensive than pre-pruning, as the full tree must be grown first before it can be pruned.

2. Requires Validation Set: Many post-pruning methods require a validation set to assess the impact of pruning. This reduces the amount of data available for training the model.

3. Complexity of Implementation: Post-pruning methods, especially those involving tuning parameters (like cost complexity pruning), can be more

complex to implement and understand than pre-pruning methods.

4. Risk of Underfitting: Similar to pre-pruning, if too much pruning is done, it can lead to underfitting where the model becomes overly simplified and fails to capture important patterns in the data.

Cost Complexity Pruning

26 July 2023 15:34