Decision Tree & Random Forest

Topics covered so far

1. Decision Trees
    a. Introduction
    b. Advantages & Disadvantages
    c. Building a DT
    d. Impurity measures
    e. Overfitting
2. Random Forest
    a. Bias-Variance Tradeoff
    b. Pruning
    c. Bagging
    d. Random forest

Discussion questions

1. What is a decision tree and how does it work?

2. How do we measure the impurity in a decision tree?

3. What is pruning? How does it avoid overfitting?

Decision Tree

- A decision tree is one of the most popular and effective supervised learning techniques for classification problems, that works equally 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 decide on going out to play tennis or not by looking at the weather conditions.
    - If it’s cloudy, then the person will go out to play.
    - If it’s sunny, the person will check the humidity level - if normal, the person will go out to play.
    - If it’s rainy, the person further checks the wind speed - if that’s weak, the person will go out to play.

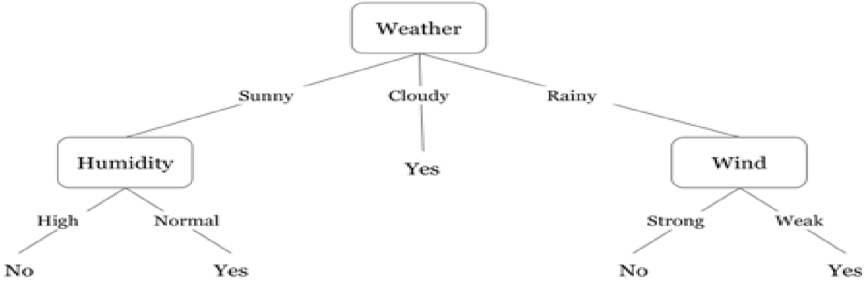


Image Source

Impurity Measures in Decision Trees

Impurity Measures: Decision trees recursively split features with respect to their target variable's purity. The algorithm is designed to optimize each split such that the purity will be maximized. Impurity can be measured in many ways such as Entropy, Information Gain, etc.

	GINI INDEX	ENTROPY	INFORMATION GAIN	VARIANCE
When to use	Classification Tree	Classification Tree	Classification Tree	Regression Tree
Formula	$G = 1- \sum^{c}_{i=1}(p_i^2)$	$E = -\sum P(X).logP(X)$	$IG\ (Y, X) = E(Y) - E(Y\|X)$	$V = \sum(x-\mu)^2/N$
Range	0 to 0.5 0 = most pure 0.5 = most impure	0 to 1 0 = most pure 1 = most impure	0 to 1 0 = less gain 1 = more gain	-
Characteristics	Easy to compute Non-additive	Computationally intensive Additive	Computationally intensive	The most common measure of dispersion

Pruning

- One of the problems with the decision tree is it gets easily overfit with the training sample and becomes too large and complex.
- A complex and large tree poorly generalizes to new sample data whereas a small tree fails to capture the information of the training sample data.
- Pruning may be defined as shortening the branches of the tree. It is 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 tree which helps in reducing the overfitting of the tree.

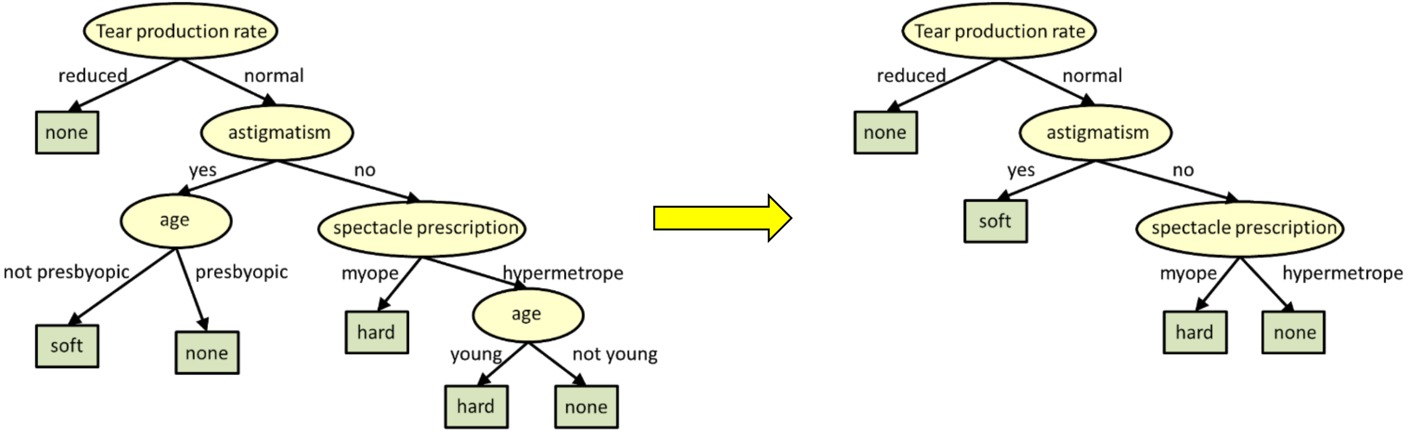


Image Source

Cost Complexity Pruning

- Cost Complexity Pruning is the most popular pruning technique for decision trees. It takes into account both the number of errors and the complexity of the tree.
- This technique is parametrized by the cost complexity parameter, ccp_alpha which reduces the complexity of the tree by controlling the number of leaf nodes, which eventually reduces overfitting. Greater values of ccp_alpha increase the number of nodes pruned.
- 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 and R(T) is the total misclassification rate of the terminal nodes.

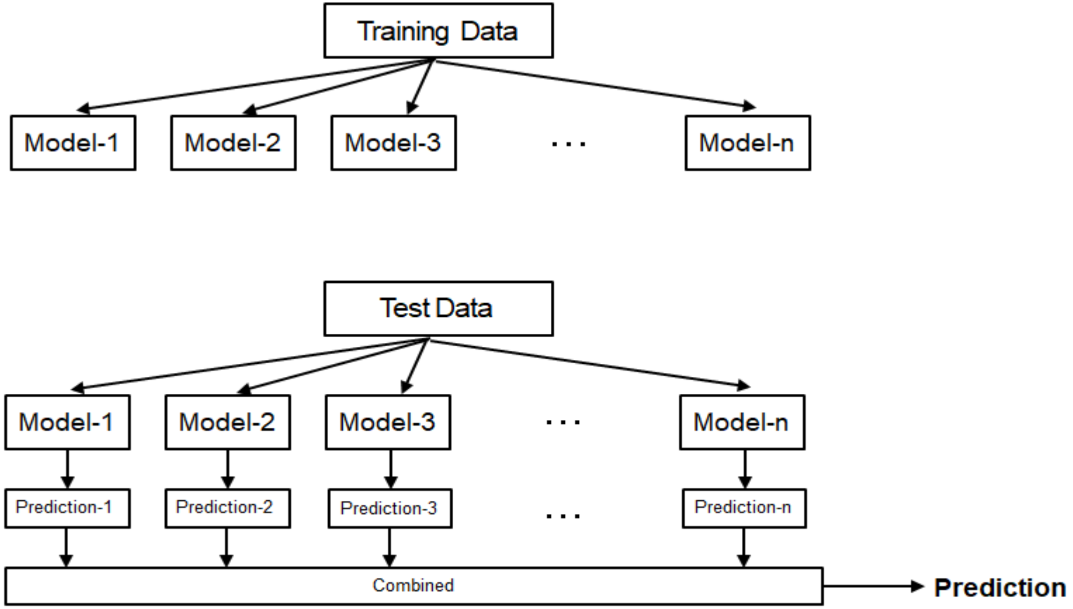
- Cost complexity pruning proceeds in the following stages:
  - A sequence of trees(T0, T1,...,Tk) for different values of alpha is built on the training data where T0 is the original tree before pruning and Tk is the root tree.
  - The tree Ti+1 is obtained by replacing one or more of the sub-trees in the predecessor tree Ti with suitable leaves.
  - The impurity of each pruned tree (T0, T1,...,Tk) is estimated and the best pruned tree is then selected based on the metric under consideration (using test data).

Discussion questions

1. What do you mean by ensemble learning?
2. What is bootstrap aggregation and how does it work?
3. What is a random forest and how is it useful?
4. What are the hyperparameters involved in a random forest model?
5. What are the advantages and disadvantages of the random forest algorithm?

Ensemble Learning

- Ensemble Learning is a paradigm of machine learning methods for combining predictions from multiple separate models.
- The central motivation is rooted under the belief that a committee of experts working together can perform better than a single expert.



Bootstrap Aggregation (Bagging)

- Bagging is a technique of merging the outputs of various models to get a final result
- It reduces the chances of overfitting by training each model only with a randomly chosen subset of the training data. Training can be done in parallel.
- It essentially trains a large number of “strong” learners in parallel (each model is an overfit for that subset of the data)
- Then it combines (averaging or voting) these learners together to "smooth out" predictions.

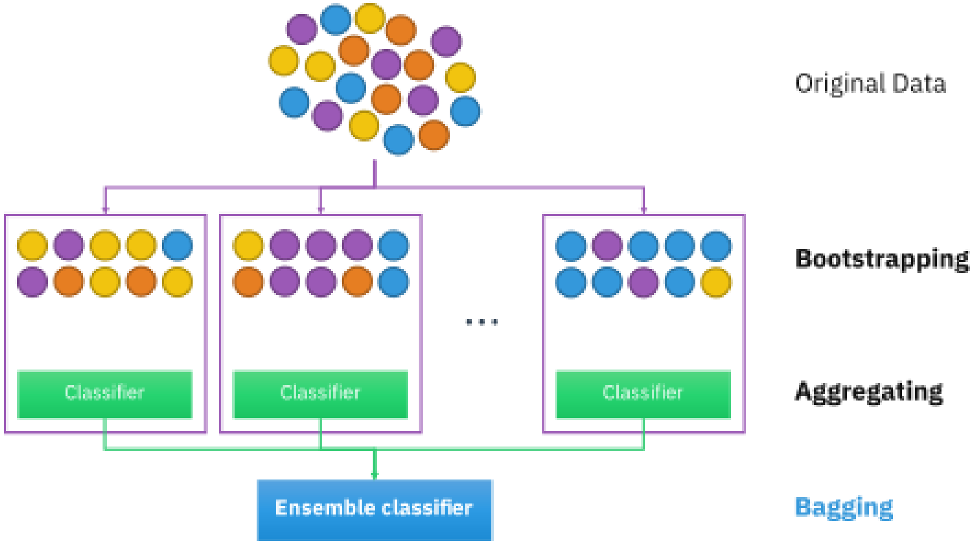


Image Source

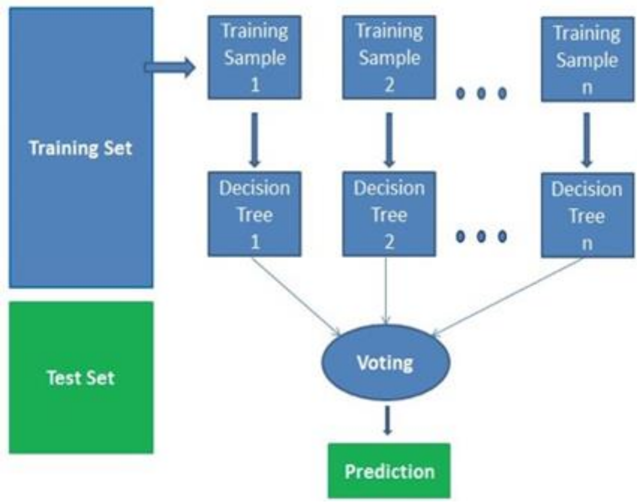
Random Forest

- Random Forest is a supervised machine learning algorithm which can be used for both classification and regression.
- It generates small decision trees using random subsamples of the dataset where the collection of the generated decision tree is defined as forest. Every individual tree is created using an attribute selection indicator such as entropy, information gain, etc.
- In classification, problem voting is done by each tree and the most voted class is considered the final result whereas in case of regression the average method is used to get the final outcome.
- Random Forest is used in various domains such as classification of images, feature selection and recommendation engines.

Steps involved in the Random Forest algorithm

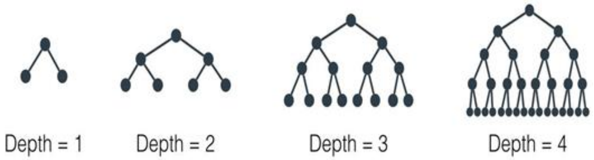
The following steps are involved in this algorithm:

1. Selection of a random subsample of a given dataset.
2. Using attribute selection indicators create a decision tree for each subsample and record the prediction outcome from each model.
3. Applying the voting/averaging method over predicted outcomes of individual models.
4. Considering the final results as the average value or most voted value.

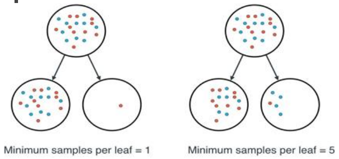


Hyperparameters in Random Forest

1. **Number of trees (n_estimators):**
    - It specifies the number of trees in the forest of the model.
    - The default value for this parameter is 10, which means that 10 different decision trees will be constructed in the random forest.



1. **Maximum Depth (max_depth):**
    - It specifies the maximum depth of the tree.
    - The default value is none which means the tree will expand until every leaf is pure



1. **The minimum number of samples per leaf (min_samples_leaf):**
    - It specifies the minimum number of samples required to be at a leaf node.
    - The default value is 1, which means that every leaf must have at least 1 sample that it classifies

1. **The minimum number of samples to split (min_samples_split):**
    - It specifies the minimum number of samples required to split an internal leaf node.
    - The default value for this parameter is 2, which means that an internal node must have at least two samples before it can be split to have a more specific classification.

Advantages and Disadvantages of Random Forest

Advantages:

- It can be used for both classification and regression problems.
- It is one of the most accurate algorithms because of the number of decision trees taking part in the process.
- It does not suffer from overfitting.
- It is used to select features of relatively more importance and helps in feature selection.

Disadvantages:

- The Random Forest algorithm is very slow compared to others because it calculates predictions for each decision tree for every sub sample and then votes on them to select the best one - which is time-consuming.
- It is difficult to explain the model as compared to a decision tree where you can easily make the decision following the path of the tree.

Case Study

Great Learning

Happy Learning !