Random Forests

Joint Proseminar on Historical Milestones of Machine Learning

Kaan İmamoğlu, 451439
Bora Deren, 438377
July 12, 2024

Chair of Computer Science 13 (Computer Vision)
RWTH Aachen

Overview

- Decision Trees
- Main Concepts and Problems With Decision Trees
- Ensembles
- Generalization Error, Bias and Variance
- Conclusion

Decision Trees

1

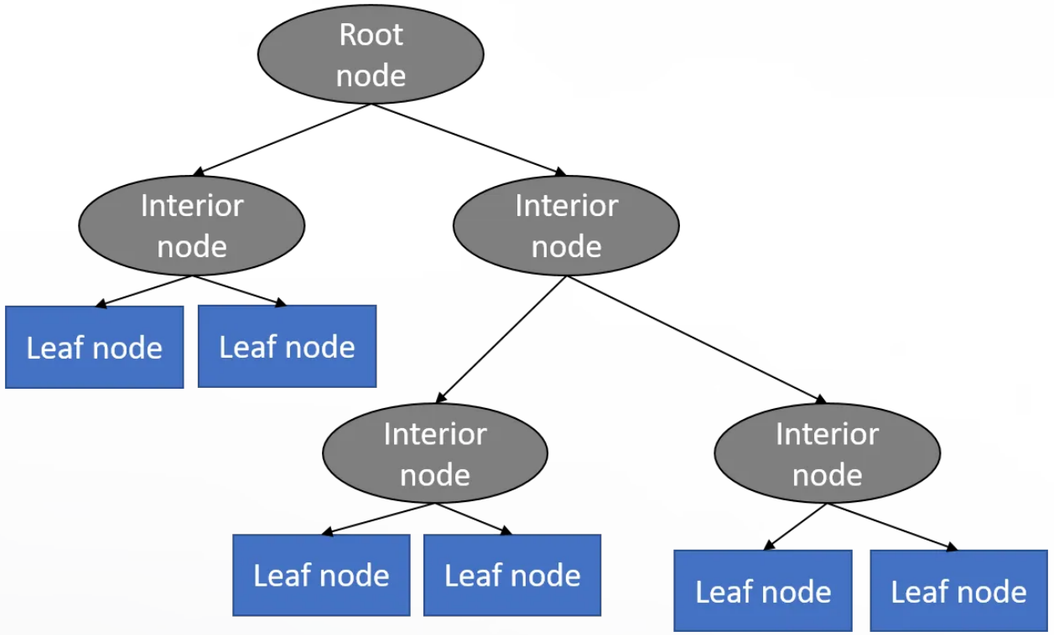
Definition

Decision tree is a machine learning technique used to make classification and regression predictions on datasets

2

Structure

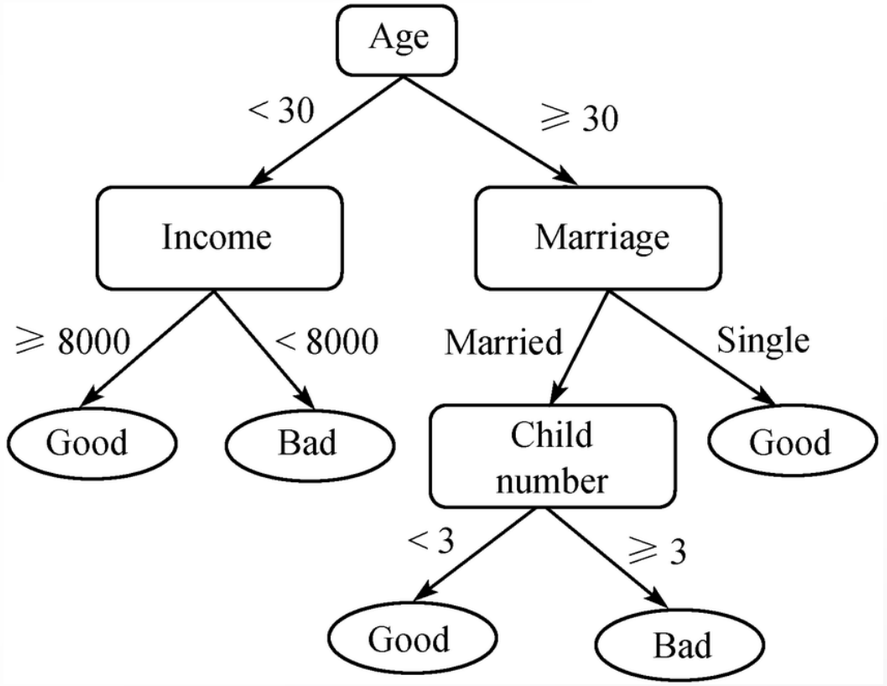
A decision tree consists of nodes forming a rooted tree.



https://medium.com/analytics-vidhya/a-deep-dive-to-decision-trees-6575e016d656

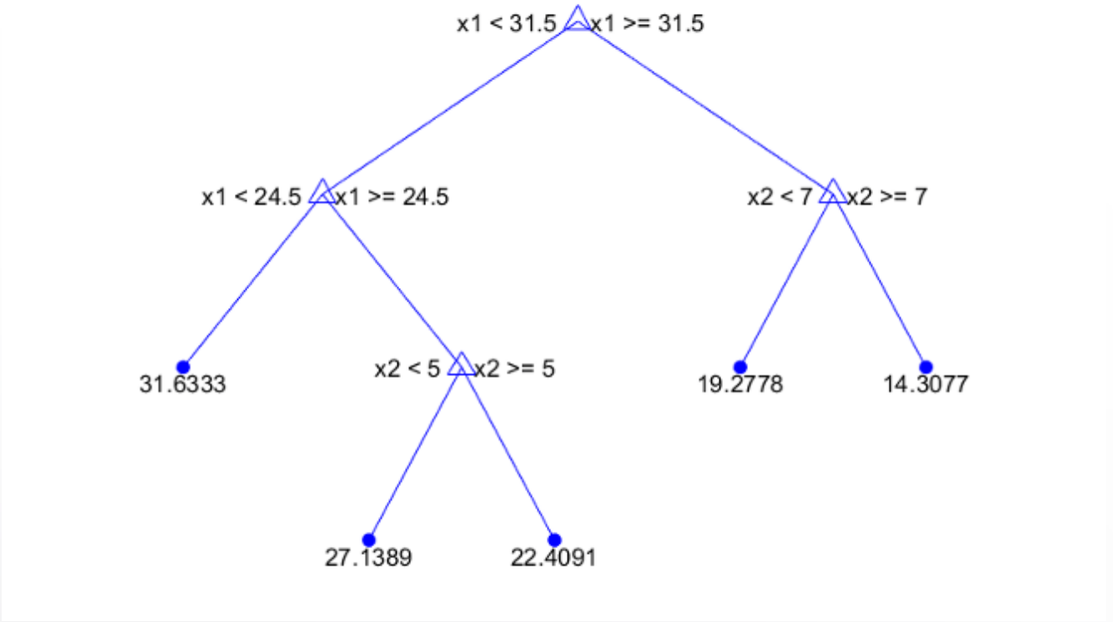
Decision Trees

Classification trees



https://www.researchgate.net/figure/An-example-of-classification-tree_fig3_233842752

Regression trees



https://de.mathworks.com/help/stats/train-regression-trees-using-regression-learner-app.html

Splitting criteria and Impurity

- One of the most important factors in construction of decision trees
- We need to find the best criteria to split the dataset for each node recursively
- The goodness of a split is quantified by an impurity measure
- A split is pure, if all instances choose the same branch and there is just one branch

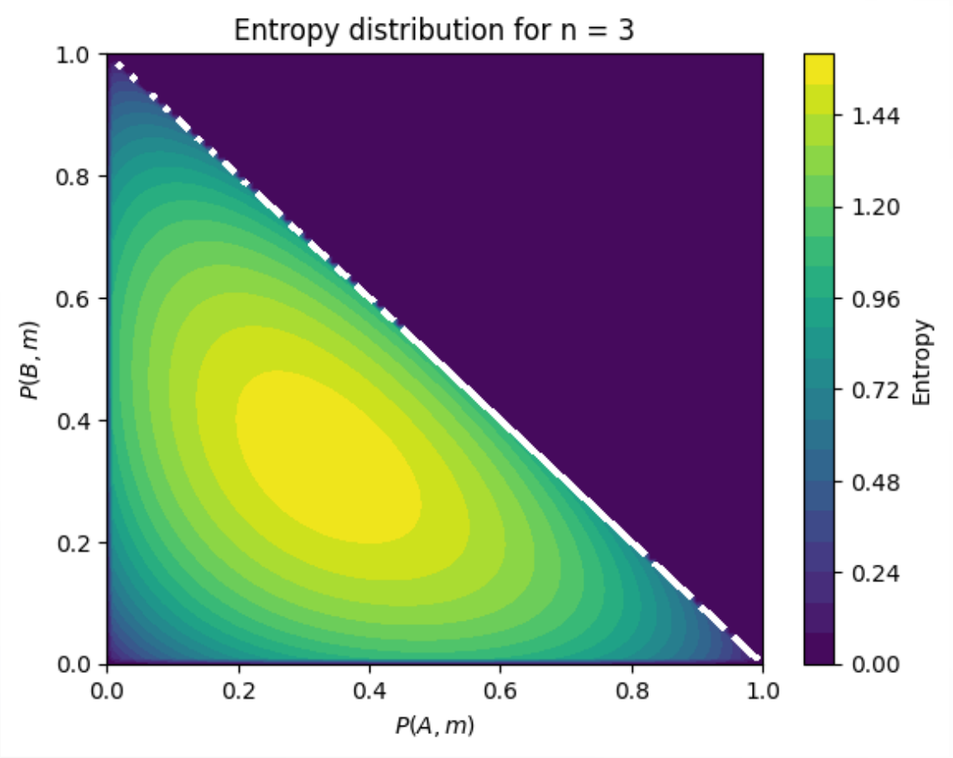
PassengerID	Age	Sex	Cabin	Survival Status
1	22	0	2	1
2	38	1	3	1
3	26	1	1	0
4	35	0	2	0
5	28	0	3	1
6	42	1	1	1
7	19	0	0	0
8	32	1	0	1
9	50	0	1	0
10	23	1	2	1

0: male, 1: female

Entropy

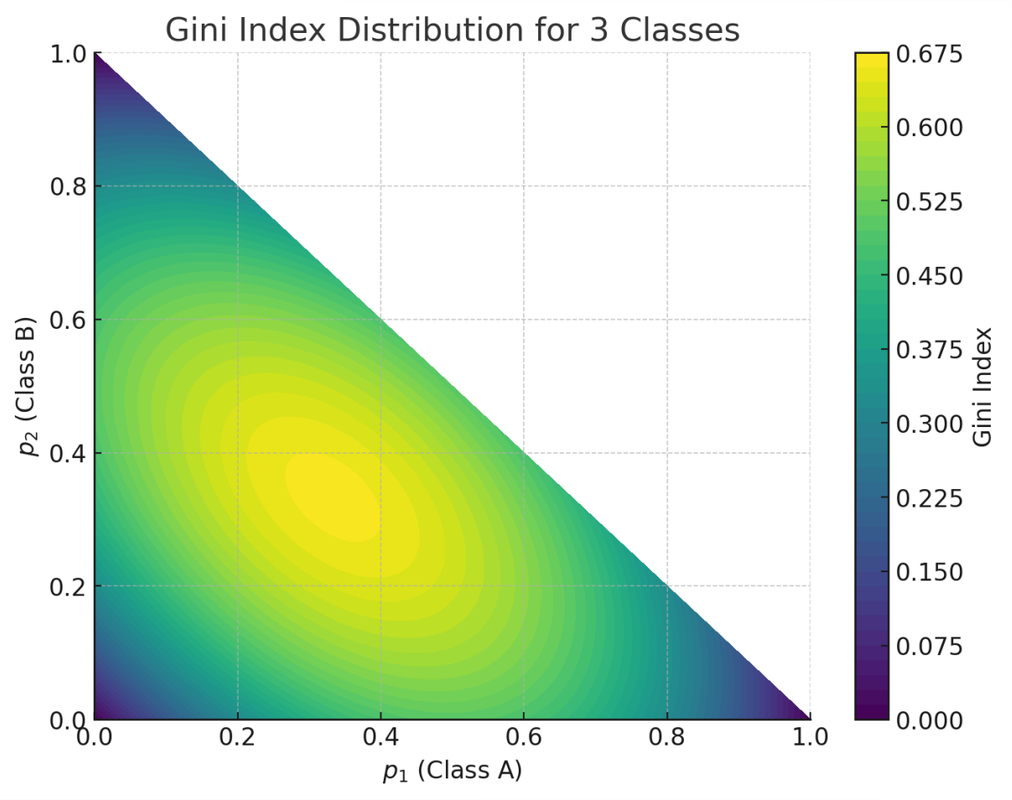
- Measure of impurity
- Refers to the homogeneity of a dataset
- It can values between 0 and log n
- It is calculated as:

  H(X) = $-\sum_{i=1}^{n} P(C_i, m) log_2(P(C_i, m))$, where n is the size of the dataset, $P(C_i, m)$ is the probability that an instance reaches class C_i in the condition it reached class m



Gini Index

- Measure of impurity
- Refers to the homogeneity of a dataset
- It can values between 0 and 1
- Towards 0, homogeneity increases and instances are equally distributed
- Calculated as: Gini(X) = $1 - \sum_{i}^{n} p_i^2$, where p_i is the probability of an instance belonging to class i



Information Gain

- Used to find best split
- Information gain of splitting S by feature x is defined as the subtraction of the average entropy of subsets from the entropy of S

PassengerID	Age	Sex	Cabin	Survival Status
1	22	0	2	1
2	38	1	3	1
3	26	1	1	0
4	35	0	2	0
5	28	0	3	1
6	42	1	1	1
7	19	0	0	0
8	32	1	0	1
9	50	0	1	0
10	23	1	2	1

0: male, 1: female

Summary

Concept	Definition
Entropy	Measures the impurity or randomness of a dataset
Information Gain	Measures the reduction in entropy after a split
Gini Index	Measures the diversity within the dataset

Decision Tree Algorithms

1

ID3 Algorithm

J. Ross Quinlan in 1986, ID3 uses information gain to choose the best local feature for splitting. It's a greedy approach that recursively builds the tree until a stopping condition is met.

2

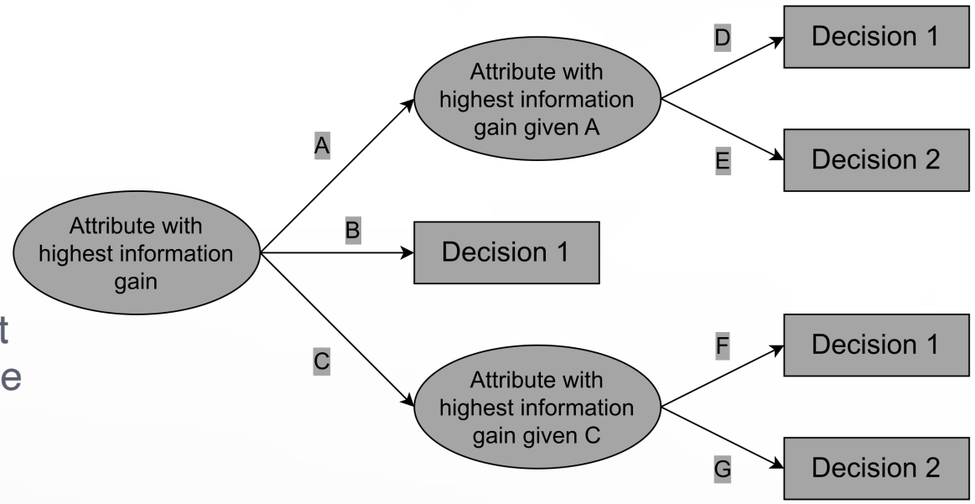
C4.5 Algorithm

C4.5 can handle missing values, and performs post-pruning. It uses the gain ratio instead of information gain for better feature selection.

3

CART Algorithm

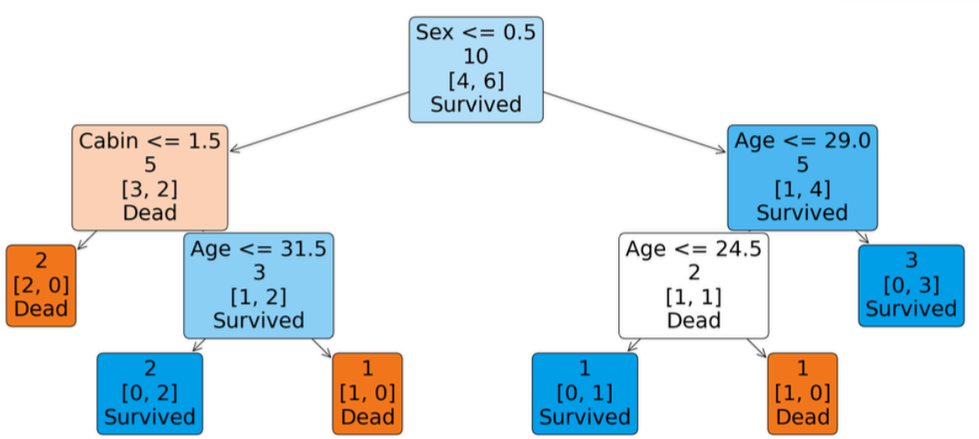
Leo Breiman in 1984, CART constructs binary trees. It uses the Gini index or entropy for classification.

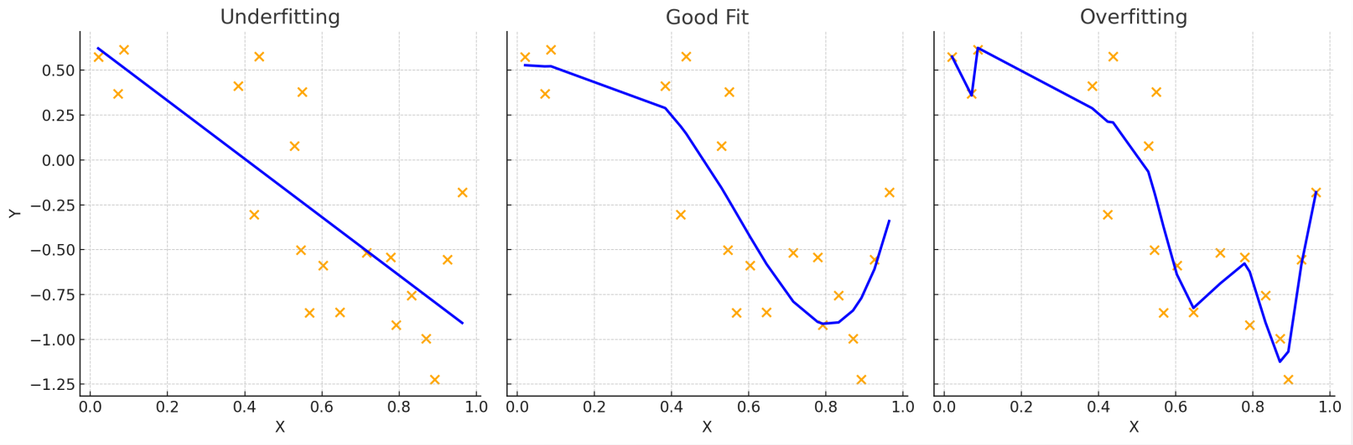


https://en.wikipedia.org/wiki/ID3_algorithm

Overfitting

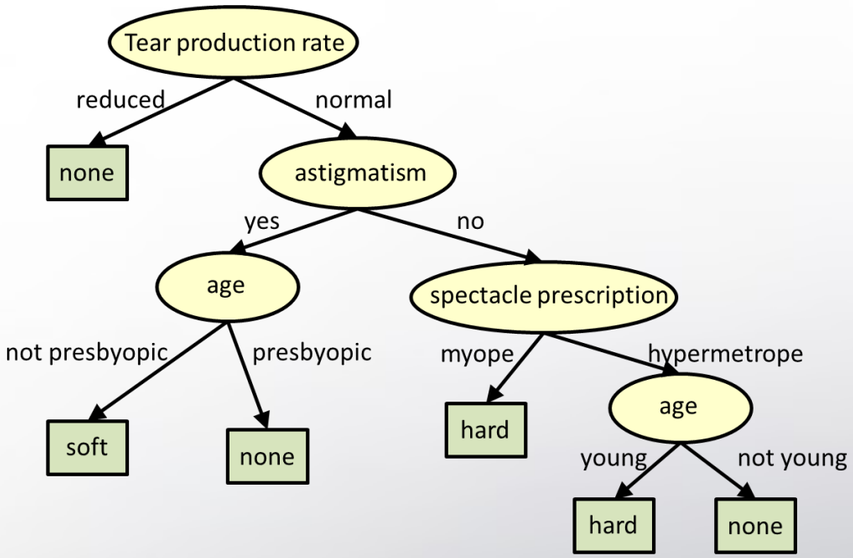
- One of disadvantages of decision trees
- Happens if the decision tree is too detailed
- Also occurs if the decision tree learns “too much” or memorizes the data set
- Small datasets or too deep constructed trees are causes of overfitting

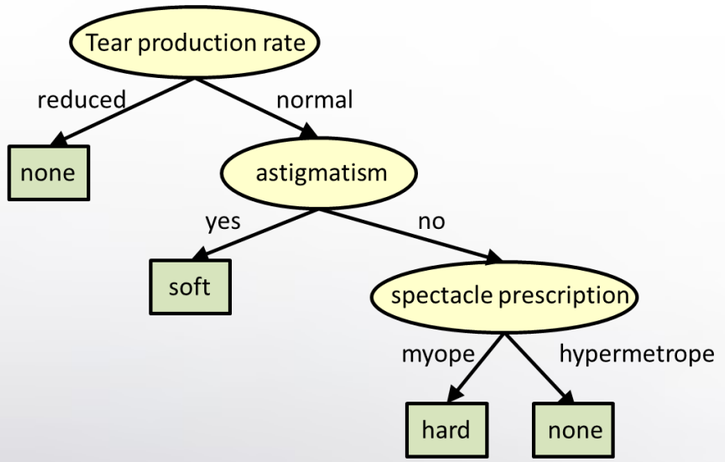




Pruning

- Pruning is a technique used to reduce overfitting by removing or merging branches. Pre-pruning stops tree growth during construction, while post-pruning simplifies the tree after it's fully grown.





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

Application and Advantages of Decision Trees

- Classification tasks: Widely used in fields like healthcare, finance, and marketing for categorising data into predefined classes.
- Regression tasks: Applied in predicting continuous variables, such as house prices or customer lifetime value.
- Performs well on large datasets
- Simple to understand and interpret
- Handles missing values well

Introduction to Ensemble Learning

Definition

Ensemble learning brings together multiple models to make decisions.

Goal

Combine basic models to create a strong collective learner with higher accuracy for classification and regression tasks.

Process

Train a group of individual learners and then combine them through some strategy to make final predictions.

Ensemble Learning Workflow

1 Train Base Learners

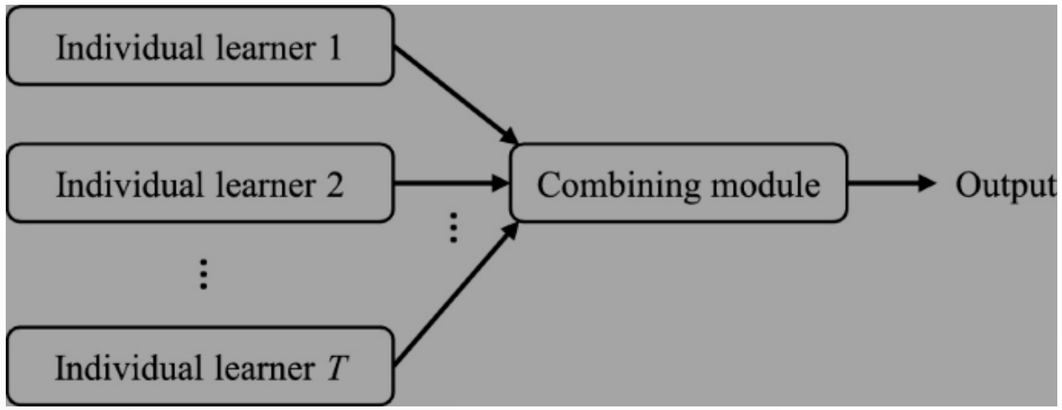
First, train a group of individual learners using the base learning algorithm.

2 Combine Learners

Next, combine the trained base learners through some strategy to create the ensemble.

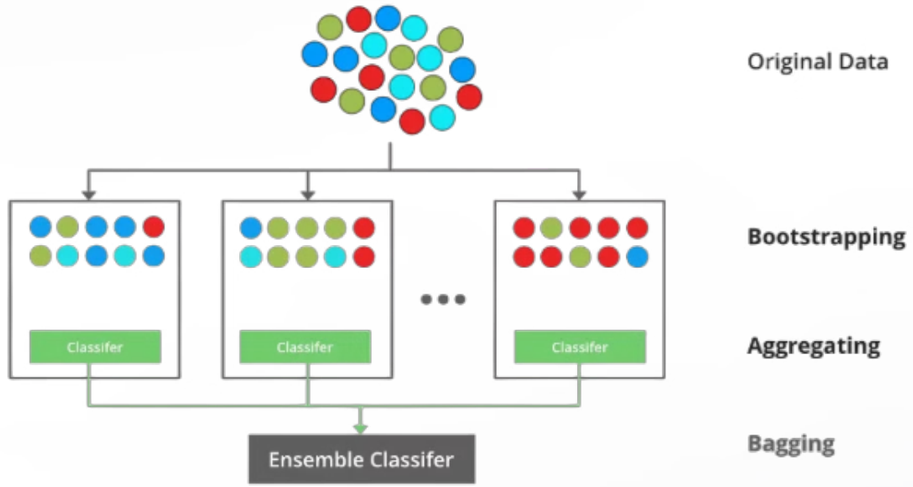
3 Make Predictions

Finally, use the combined ensemble to make predictions on new data, leveraging the collective knowledge of multiple models.



Zhi-Hua Zhou and Zhi-Hua Zhou. *Ensemble learning*. Springer, 2021.

Bootstrap Aggregating (Bagging)



https://medium.com/@hemaanushatangellamudi/bootstrapped-aggregation-bagging-481f4812e3ea

Definition

Bootstrap Aggregating, or bagging, is an ensemble algorithm that improves accuracy of machine learning algorithms.

Purpose

Bagging aims to prevent overfitting and reduce variance.

Key Datasets

Bagging uses original, bootstrap, and out-of-bag datasets.

Algorithm 8.2 Bagging

Input: Training set: $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
Base learning algorithm $\mathcal{L}$;
Number of training rounds T.

Process:

1: **for** $t = 1, 2, \ldots, T$ **do**
2: $h_t = \mathcal{L}(D, \mathcal{D}_{bs})$.
3: **end for**

Output: $H(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} \sum_{t=1}^{T} \mathbb{I}(h_t(\boldsymbol{x}) = y)$.

Bagging Algorithm Implementation

The bagging algorithm provides more stable and reliable predictions by averaging multiple models, reducing overfitting and improving accuracy through diverse training datasets.

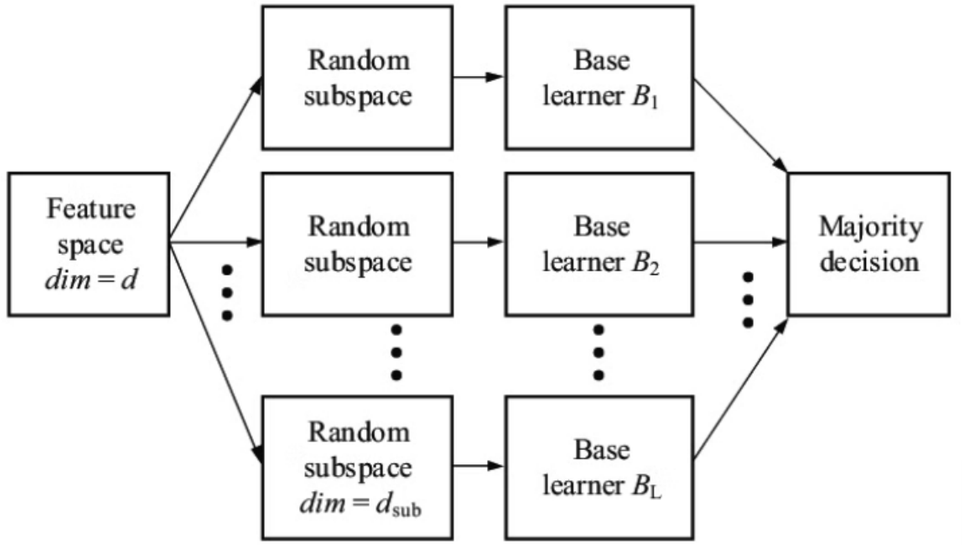
Random Subspace Method

Overview

The random subspace method, introduced by Tin Kam Ho in 1998, selects random subsets of features for each base learner.

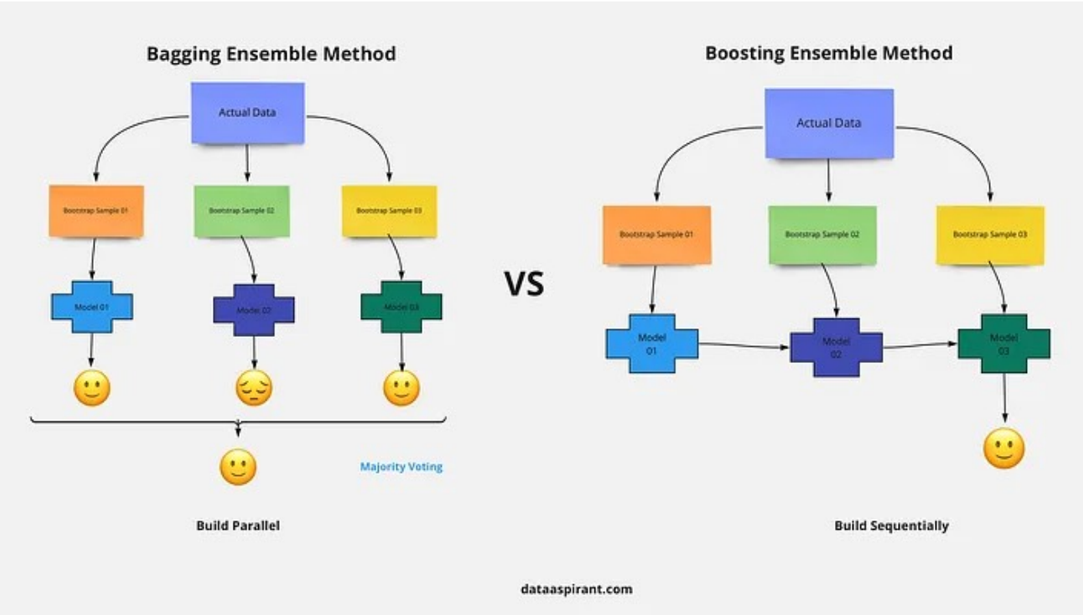
Advantages

Increases diversity among base learners, improves generalization, reduces overfitting, and is useful for high-dimensional datasets.



https://www.researchgate.net/figure/Scheme-of-Ensemble-Learning-Random-subspaces-are-obtained-from-feature-space-dim-and_fig1_317828745

Boosting Methods (AdaBoost, XGBoost)



https://medium.com/geekculture/xgboost-versus-random-forest-898e42870f30

Definition of Boosting

- Boosting is a technique that builds a strong model by gradually improving a series of simple models, with each one learning from the mistakes of the last.

Boosting vs. Bagging

- Boosting: corrects errors from previous models
- Bagging: combines their results

Random Decision Forests

1

Concept

Random forests combine multiple decision trees into an ensemble, producing more accurate results.

2

Development

Extension of Bagging idea

3

Functionality

Builds multiple decision trees using Bagging and Random subspace.

4

Prediction

Majority Voting to determine final prediction

Voting in Random Forests

- **Simple Majority Voting**

Each decision tree uses a single vote. As a result of this voting, the class that gets the most votes becomes the final prediction.

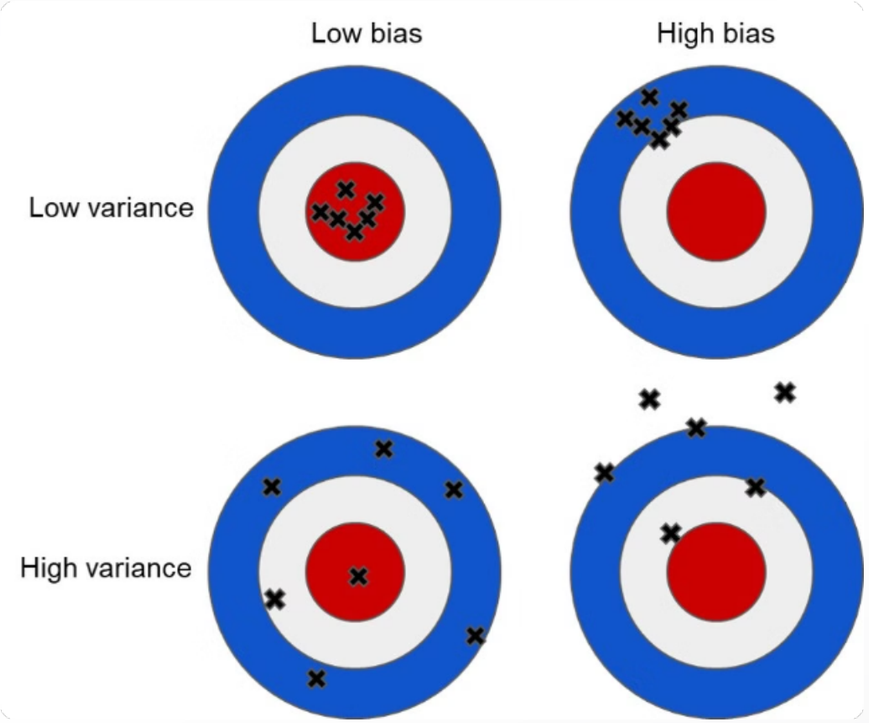
$$\max_{1 \leq j \leq k} \sum_{t=1}^{n} d_{t,j}$$

- **Weighted Majority Voting**

The prediction with the highest weighted vote becomes the final output.

$$\max_{1 \leq j \leq k} \sum_{t=1}^{n} w_t d_{t,j}$$

Generalization Error and Bias-Variance Decomposition



https://corporate.jasoncollins.blog/heuristics-and-the-bias-variance-trade-off

Generalization Error

Measures model errors on unseen data, calculated differently for regression and classification problems.

Bias-Variance Tradeoff

Balancing model complexity to minimize both bias (underfitting) and variance (overfitting) errors.

$$\text{Bias}^2 = \sum_i \left(f(x_i) - \mathbb{E}_{\mathcal{D}} \left[f(x_i; \hat{\theta}_{\mathcal{D}}) \right] \right)^2,$$

$$\text{Variance} = \sum_i \mathbb{E}_{\mathcal{D}} \left(\left(f(x_i; \hat{\theta}_{\mathcal{D}}) - \mathbb{E}_{\mathcal{D}} \left[f(x_i; \hat{\theta}_{\mathcal{D}}) \right] \right)^2 \right),$$

$$\text{Error} = \text{Bias}^2 + \text{Variance} + \sum_i \sigma_\epsilon^2,$$

Cross-Validation

Leave-P-Out Cross-Validation

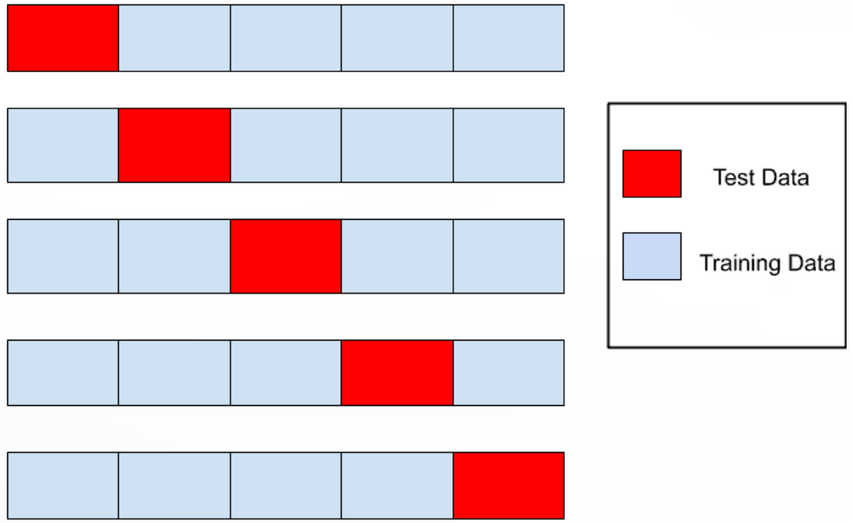
Exhaustive method that tests all possible combinations of p observations removed from the dataset.

Leave-One-Out Cross-Validation

Special case of Leave-P-Out where p=1, less computationally intensive but still costly for large datasets.

K-Fold Cross-Validation

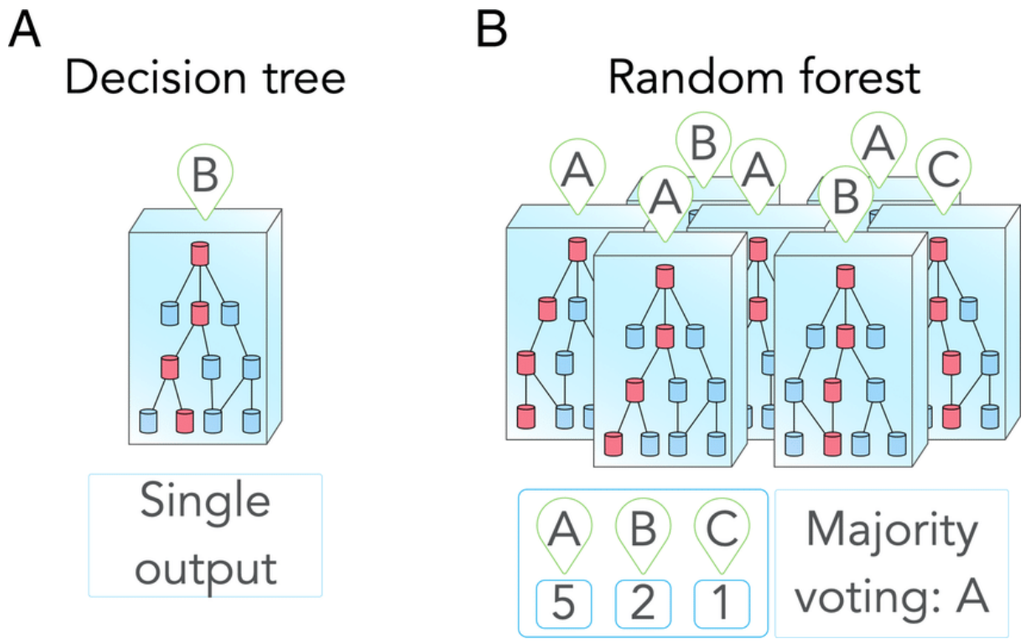
Divides data into k subsets, using each subset once as validation set and the rest for training, then averages results.



https://www.mltut.com/k-fold-cross-validation-in-machine-learning-how-does-k-fold-work/

Conclusion

- Decision trees offer straightforward modelling but have limitations
- Random forests enhance predictive performance by combining multiple models
- Random forest solves some problems of decision trees such as overfitting
- Random forest is a valuable tool in machine learning that aims increased accuracy and robustness



https://www.researchgate.net/figure/Random-forests-are-collections-of-randomised-decision-trees-A-A-single-decision-tree_fig1_339279807