Machine Learning

Supervised Learning- Decision Trees

Part Two

Dr. Sherif Saad

Learning Objectives

Introduce the students to machine learning concepts

Explain the main three types of learning and ML terminology

Understand the building blocks for successfully designing machine learning systems.

How to apply machine learning algorithms (not really how to create them)

Outlines

- Data Splitting Techniques
- Regression Decision Trees
- Decision Trees Overfitting
- Continuous Features

Selecting feature to Split the Data

- Information Gain
- Gini Index
- Chi-Square
- Reduction in Variance

Gini Index

Idea: If we randomly selected any two samples from the data, then they must belong to the same class, and the probability for that to happen is 1 if the data is pure (only one class exist in the data)

The Gini index works with binary decisions (Yes|No)

It can only do binary split

A higher Gini index means a better split (less uncertainty)

The CART (Classification And Regression Tree) algorithm uses Gini index to split the data

Gini Index

To calculate the gini index for a split:

- Calculate Gini for subsets, using formula sum of square of probability for yes and no ***(p^2+q^2)***.
- Calculate Gini for split using weighted Gini score of each subset of that split.

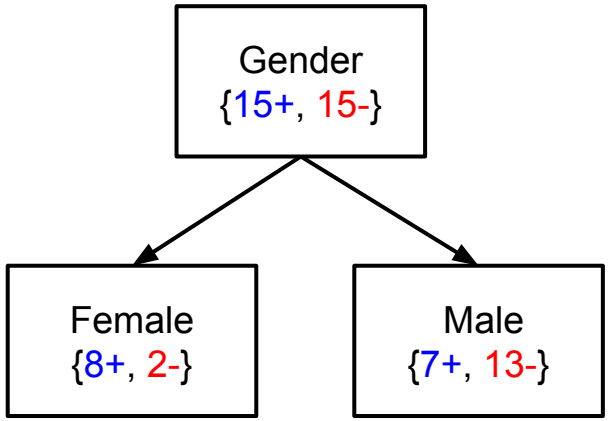
Gini Index: Titanic Dataset Example

Let us consider the Titanic dataset and assume we have set of 30 passengers. Only 15 passengers survived. Let us say we can split the dataset based on the gender or based on the age of majority.

Based on gender we get two subsets Male and Female. The male has 20 passengers, and the female has 10 passengers. 2 females did not survive, and 13 males did not survive

Based on the age of majority, we get to subsets above the age of majority and less than the age of majority. There are 6 out of the 14 passengers in the less than the age of majority did not survive. There are 9 out of the 16 passengers above the majority age did not survive

Gini Index: Titanic Dataset Example

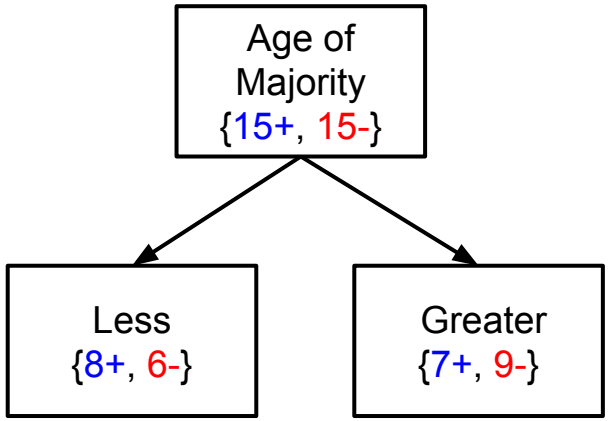


$0.8^2 + 0.2^2 = 0.68$

$0.35^2 + 0.65^2 = 0.55$

Gini for Split Gender

10/30 * 0.68 + 20/30 * 0.55 = 0.59



$0.57^2 + 0.43^2 = 0.51$

$0.44^2 + 0.65^2 = 0.51$

Gini index for Age of Majority

14/30 * 0.51 + 16/30 * 0.51 = 0.51

The higher the Gini index the better the split

Chi-Square

Work for binary classification and can generate more than two split.

Calculate the Chi-Square test for each candidate split (subset) using the equation

$$ChiSquare = \sqrt{\frac{(Actual - Expected)^2}{Expected}}$$

The higher the Chi-Square the better the split (the feature is more significant)

The CHAID (chi-square automatic interaction detection) decision tree algorithm

Reduction in Variance

The reduction in variance is a technique to build regression decision trees.

We calculate the variance of the decision variable (feature) and based on the variance we decide if the split is good or not (pure or not pure)

$$variance = \frac{\sum_1^n (x - \mu)}{n}$$

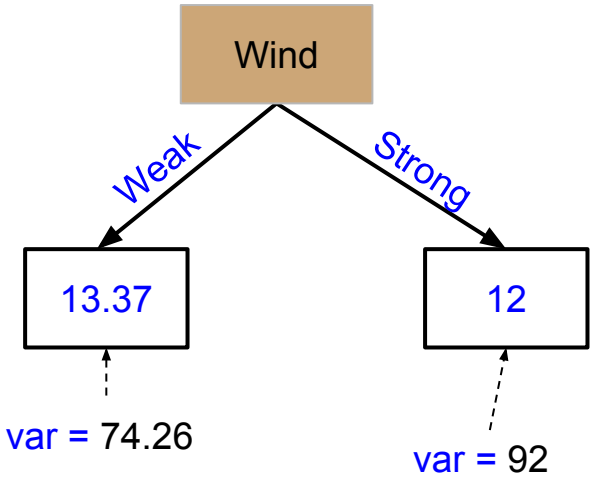
The split result in (lower | higher)? variance is selected as the best split.

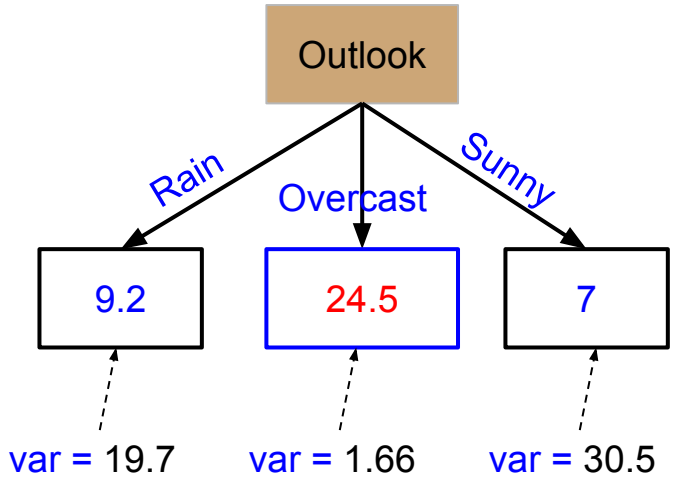
For each subset we calculate the variance and for candidate split we calculate the weighted average of each subset variance.

Regression Decision Tree

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	2h
D2	Sunny	Hot	High	Strong	3h
D3	Overcast	Hot	High	Weak	26h
D4	Rain	Mild	High	Weak	12h
D5	Rain	Cool	Normal	Weak	14h
D6	Rain	Cool	Normal	Strong	5h
D7	Overcast	Cool	Normal	Strong	24h
D8	Sunny	Mild	High	Weak	4h
D9	Sunny	Cool	Normal	Weak	13h
D10	Rain	Mild	Normal	Weak	11h
D11	Sunny	Mild	Normal	Strong	13h
D12	Overcast	Mild	High	Strong	23h
D13	Overcast	Hot	Normal	Weak	25h
D14	Rain	Mild	High	Strong	4h

Regression Decision Tree



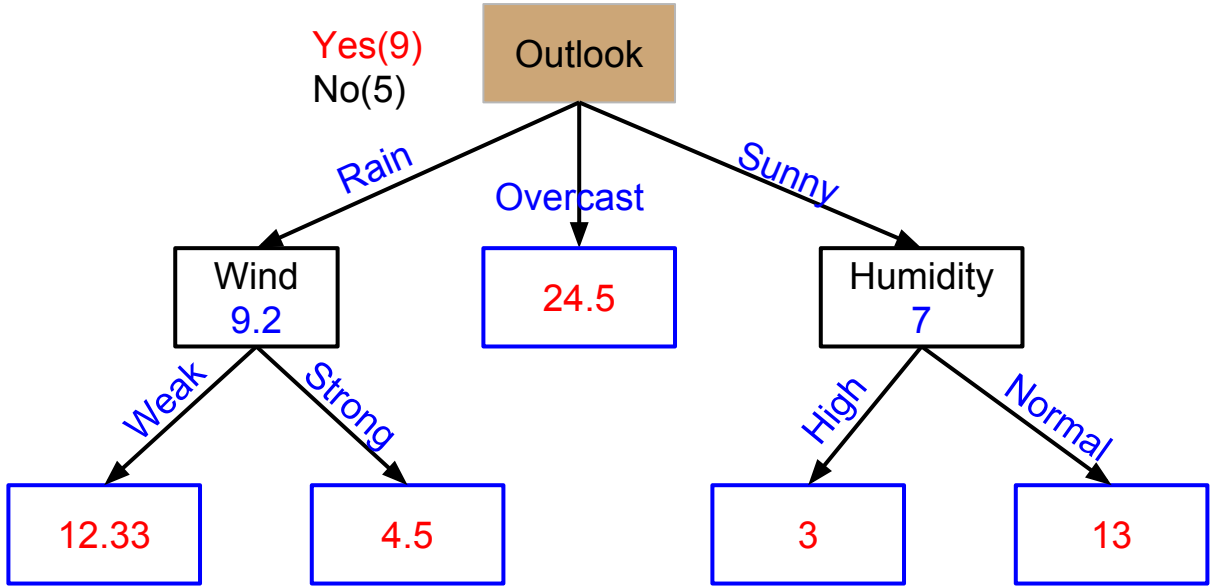


Rot Node

var= 74.26 * 8/14 + 92 * 6/14 = 81.8

var = 19.7 * 5/14 + 1.66 * 4/14 + 30.5 * 5/14 = 18.4

Decision Trees: Tennis Example

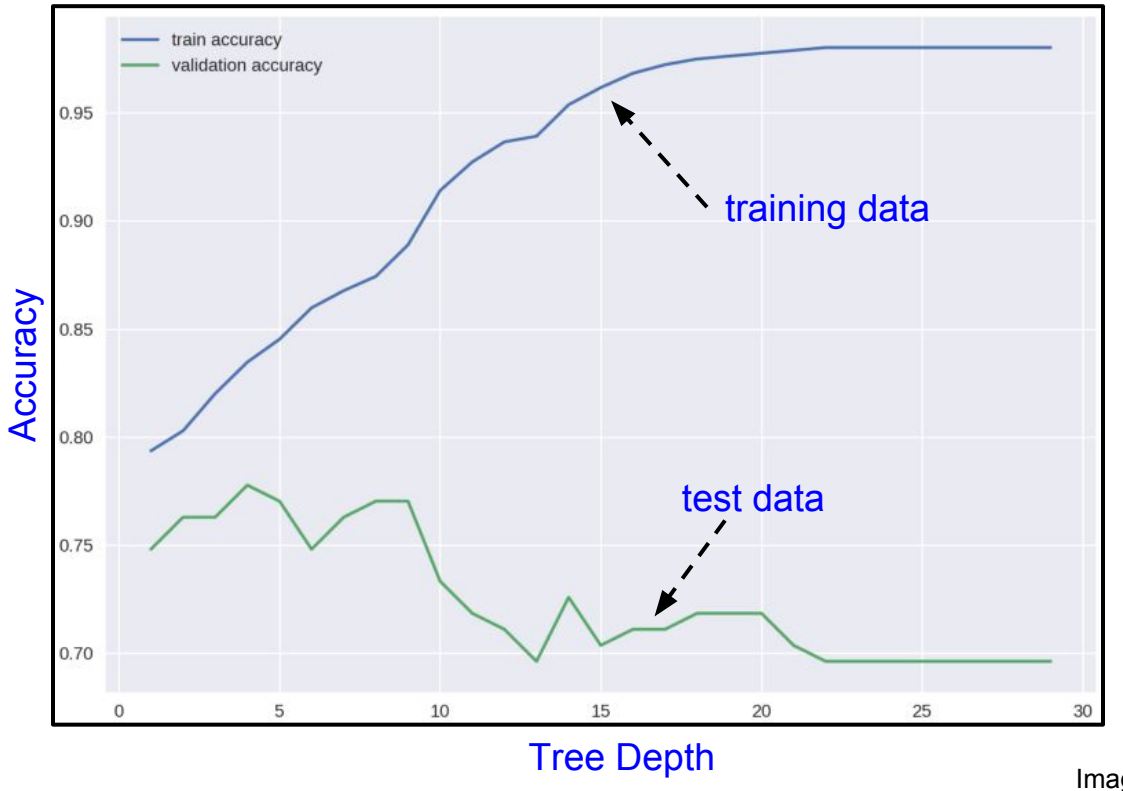


Decision Trees Overfitting

What is overfitting in machine learning?

- When the machine learning model tightly fits the given training data to the point it will be inaccurate when applied to new data (not observed during the training phase)
- Think of buying an Off-the-rack suit and tailored suit in term of cost and fitting
- In decision tree overfitting occurs when the decision tree is constructed in a way that perfectly fit all the samples in the training set?

Decision Trees Overfitting



Titanic Data Set: **Accuracy for the validation set drops significantly as tree depth increases, suggests overfitting**

Image source: https://medium.com/towards-data-science/

Decision Trees Overfitting

In general, there are two methods to avoid or mitigate the effect of overfitting in decision trees.

It is clear that when the tree size increases the chance of overfitting increase. Then one method to avoid overfitting is to set constraints on the tree size while constructing the tree

The second method is tree pruning, which removes nodes from the tree after the initial training.

Decision Trees Overfitting

Setting Constraints on Tree Size:

- Maximum Depth of the tree
  - Larger depth will result in overfitting.
- Maximum features to consider for split
  - Consider a value between the square-root to 40% of the total number features
- Maximum number of terminal nodes
  - In a binary split for a tree of ***depth n*** the maximum number of leaf node ***is 2^n***
- Minimum samples for a terminal node
- Minimum samples in a node to consider for split

To estimate the a good maximum or minimum value we use cross-validation

Decision Trees Overfitting

Tree Pruning:

- Use the training data to construct a decision tree without applying any constraints. (e.g. create a decision tree with maximum depth)
- Starting from the leaf node move towards the root node and remove decision nodes that give negative gain.

Decision Trees Overfitting

Tree Pruning (continue):

- Use subtree replacement pruning, use a validation set (testing data), remove a node from the tree and its subtrees, run the validation and check the performance
- Remove the node that results in the greatest improvement, continue applying pruning until you notice a decrease in the performance.

Decision Trees : Continuous Features

The challenge with continuous features is that we have infinitely many possible split points.

What is the minimum branching factor for a given continuous features ***F*** ?

ID	X	Y	Class
1	1.7	8	B
2	3.2	1.4	B
3	1.3	7.1	A
4	4.2	0.7	B
5	0.5	9.8	A
6	2.6	4.7	A
7	6.1	0.3	B
8	1.8	7.9	A
9	2.4	5.6	B

Decision Trees : Continuous Features

- Select a threshold ***t*** over the range of the continuous feature.
- To compute the best threshold ***t*** for a given feature ***f***:
  - Sort the samples based on the values in f
  - Move the threshold t from the smallest value to the largest value
  - Select t that give the best information gain
  - We only compute the information gain when the class label change

ID	X	Y	Class
1	1.7	8	B
2	3.2	1.4	B
3	1.3	7.1	A
4	4.2	0.7	B
5	0.5	9.8	A
6	2.6	4.7	A
7	6.1	0.3	B
8	1.8	7.9	A
9	2.4	5.6	B

Decision Trees : Continuous Features

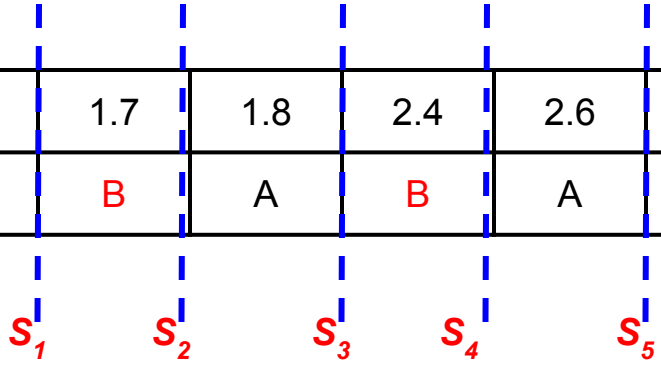
- Select a threshold ***t*** over the range of the continuous feature.
- To compute the best threshold ***t*** for a given feature ***f***:
  - Sort the samples based on the values in f
  - Move the threshold t from the smallest value to the largest value
  - Select t that give the best information gain
  - We only compute the information gain when the class label change

ID	X	Y	Class
5	0.5	9.8	A
3	1.3	7.1	A
1	1.7	8	B
8	1.8	7.9	A
9	2.4	5.6	B
6	2.6	4.7	A
2	3.2	1.4	B
4	4.2	0.7	B
7	6.1	0.3	B

Decision Trees : Continuous Features

We only need to calculate the information gain | gini index when the class label change

X	0.5	1.3	1.7	1.8	2.4	2.6	3.2	4.2	6.1
Class	A	A	B	A	B	A	B	B	B



Could you guess which point is the best split?

How many points should we choose fro splitting?

What is the information gain at t =1.8 ?

Questions