SCS 3201

Machine Learning and Neural Computing

Kasun Gunawardana

University of Colombo School of Computing

Decision Trees

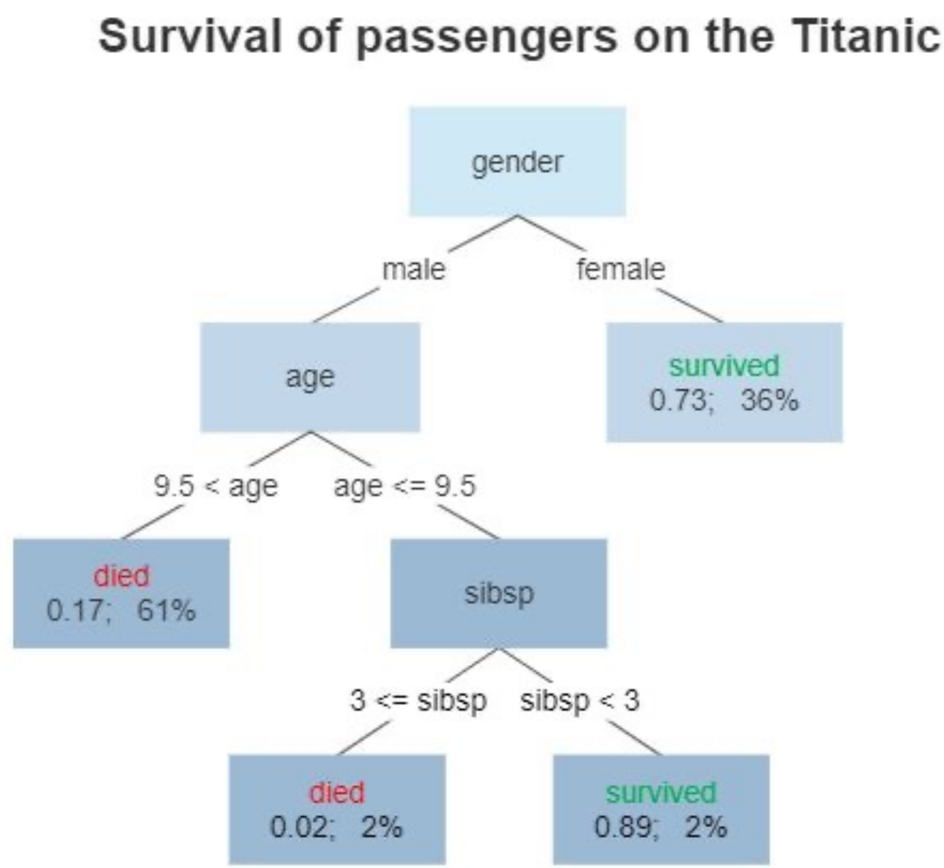
Decision Tree

- Supervised Learning algorithm
- Often used for classification tasks

Decision Tree

- The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

- “sibsp" is the number of spouses or siblings aboard



Source - https://en.wikipedia.org/wiki/Decision_tree_learning

Decision Tree: Why?

- Easy to understand
- Easy to explain the outcome
- Applicable for Numerical and Categorical data
- Easy to build
- Efficient and Scalable
- Non-parametric
- Can handle missing data well

Two Main Types

1. **Categorical Variable Decision Tree**:

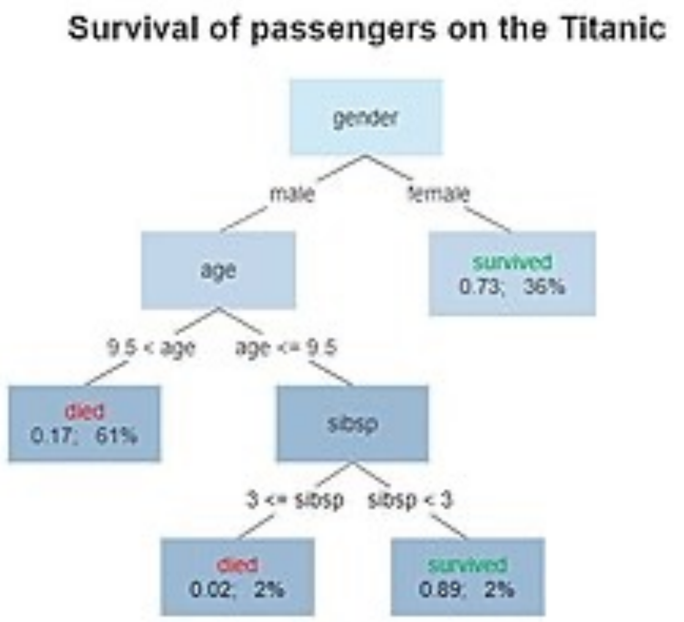
   To deal with categorical target variable.

2. **Continuous Variable Decision Tree**:

   To deal with a continuous target variable.

Key Concepts

- Root Node:
- Decision Node: A node branch out
- Leaf Node: Nodes do not split
- Branch: A sub tree of the tree
- Split: Division of a node into brnaches



Decision Tree: Basics

- At the beginning, the entire training dataset is regarded as the Root Node.
- Data points are consumed iteratively and distributed over the tree based on attribute values.
- Selection order of attributes at different levels on the decision tree is decided using statistical measures.

Decision Tree: Basics (Cont.)

- Decision trees classify a data point by sorting it down the tree from the root to a leaf node.
- Leaf node provides the classification of the example.
- Each node in the tree acts as a test case for some attribute.
- Each edge descending from the node corresponds to the possible answers to the test case.

Decision Tree: Basics (Cont.)

- A split increases the homogeneity of resultant sub-nodes (purity of the node increases).
- The way that a split decision is made affects the overall accuracy of the decision tree.
- Several algorithms are available for this task.
  - To decide to split a node into sub-nodes.
- Type of the target variable also influences on the algorithm selection.

Challenge

- Identification of the most effective order of attributes for splits.

Challenge (Cont.)

- Identification of the most effective order of attributes for splits.

- Rank attributes using a measurement that quantifies the effectiveness of each attribute for a split.

Attribute Selection Measures

- Entropy
- Information gain
- Gini index
- Gain Ratio
- Chi-Square
- Reduction in Variance

Entropy

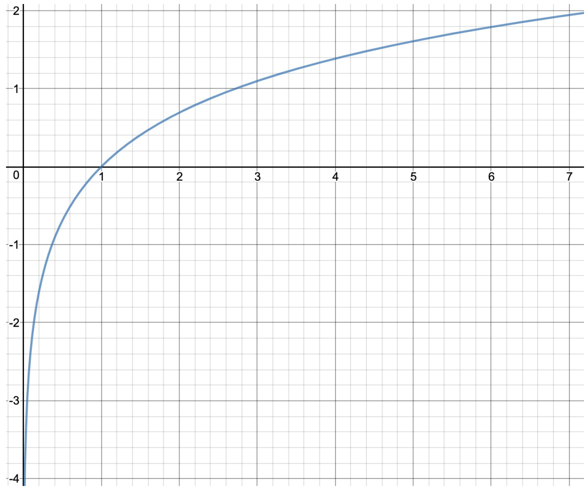
- Measure of uncertainty of a random variable
- Characterizes the impurity of an arbitrary collection of examples.
- The higher the entropy more the information.

Entropy

- Entropy for a dataset with C classes is defined as;

$$H(S) = -\sum_{i}^{C} p_i . \log_2 p_i$$

- $H(S)$: Entropy of dataset S
- p_i: Probability of randomly picking a data of class i.



Typical $ln(x)$ graph

Example

#	Outlook	Temperature	Humidity	Wind	Play Golf ?
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cold	Normal	False	Yes
6	Rainy	Cold	Normal	True	No
7	Overcast	Cold	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cold	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Entropy - Example

- Let's find the Entropy for the dataset considering the target variable
  - 14 – Instances in two classes (Yes/ No)
  - 9 – Yes
  - 5 – No

$$H(S) = -\left(\frac{9}{14} \times \log_2 \frac{9}{14}\right) - \left(\frac{5}{14} \times \log_2 \frac{5}{14}\right)$$

$$H(S) = 0.94$$

Entropy (Given a Feature)

- If we split the dataset using the feature X, what would be the new Entropy?
- Entropy of S given X;

$$H(S|X) = \sum_{c \in X} P(c).H(c)$$

Entropy (Given a Feature)

- If we split the dataset using the feature 'Outlook', what would be the new Entropy?

$$H(Play|Outlook) = ??$$

		Play Golf ?	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Entropy (Given a Feature)

		Play Golf ?	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

- $H(Play|Outlook) = P(sunny).H(sunny) + P(overcast).H(overcast) + P(rainy).H(rainy)$

- $H(Play|Outlook) = \frac{5}{14}\left(-\frac{3}{5} \times \log_2 \frac{3}{5} - \frac{2}{5} \times \log_2 \frac{2}{5}\right) + \frac{4}{14}\left(-\frac{4}{4} \times \log_2 \frac{4}{4} - \frac{0}{4} \times \log_2 \frac{0}{4}\right) + \frac{5}{14}\left(-\frac{2}{5} \times \log_2 \frac{2}{5} - \frac{3}{5} \times \log_2 \frac{3}{5}\right)$

		Play Golf ?	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

- $H(Play|Outlook) = \frac{5}{14}\left(-\frac{3}{5}\times\log_2\frac{3}{5} - \frac{2}{5}\times\log_2\frac{2}{5}\right) + \frac{4}{14}\left(-\frac{4}{4}\times\log_2\frac{4}{4} - \frac{0}{4}\times\log_2\frac{0}{4}\right) + \frac{5}{14}\left(-\frac{2}{5}\times\log_2\frac{2}{5} - \frac{3}{5}\times\log_2\frac{3}{5}\right)$

- $H(Play|Outlook) = \frac{5}{14}\times 0.971 + \frac{4}{14}\times 0.0 + \frac{5}{14}\times 0.971$

- $H(Play|Outlook)$ =0.693

Entropy for Splits

- $H(Play|Outlook) = 0.693$
- $H(Play|Temperature) = 0.911$
- $H(Play|Humidity) = 0.788$
- $H(Play|Windy) = 0.892$

Entropy for Splits

- $H(Play|Outlook) = 0.693$
- $H(Play|Temperature) = 0.911$
- $H(Play|Humidity) = 0.788$
- $H(Play|Windy) = 0.892$

- Which split gives the highest Entropy reduction ?

Information Gain

- Measures the difference in Entropy between before and after split using an attribute.
- Degree of uncertainity reduced by a split

$$IG(S,X) = H(S) - H(S|X)$$

$IG(S,X)$- Information Gain by split	$H(S\|X)$- Entropy after split by feature X
$H(S)$- Entropy before split	

Information Gain

- Measures the difference in Entropy between before and after split using an attribute.
- Degree of uncertainity reduced by a split

$$IG(S, X) = H(S) - \sum_{c \in X} P(c).H(c)$$

$H(S)$ - Entropy of set S, before split	$P(c)$ - proportion of elements in subset c
X – Feature X based subsets	$H(c)$ – Entropy of subset c

Information Gain - Example

#	Outlook	Temperature	Humidity	Wind	Play Golf ?
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cold	Normal	False	Yes
6	Rainy	Cold	Normal	True	No
7	Overcast	Cold	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cold	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Information Gain

- $IG(Play|Outlook) = 0.94 - 0.693$
- $IG(Play|Temperature) = 0.94 - 0.911$
- $IG(Play|Humidity) = 0.94 - 0.788$
- $IG(Play|Windy) = 0.94 - 0.892$

Information Gain

- $IG(Play|Outlook) = 0.247$
- $IG(Play|Temperature) = 0.029$
- $IG(Play|Humidity) = 0.152$
- $IG(Play|Windy) = 0.048$

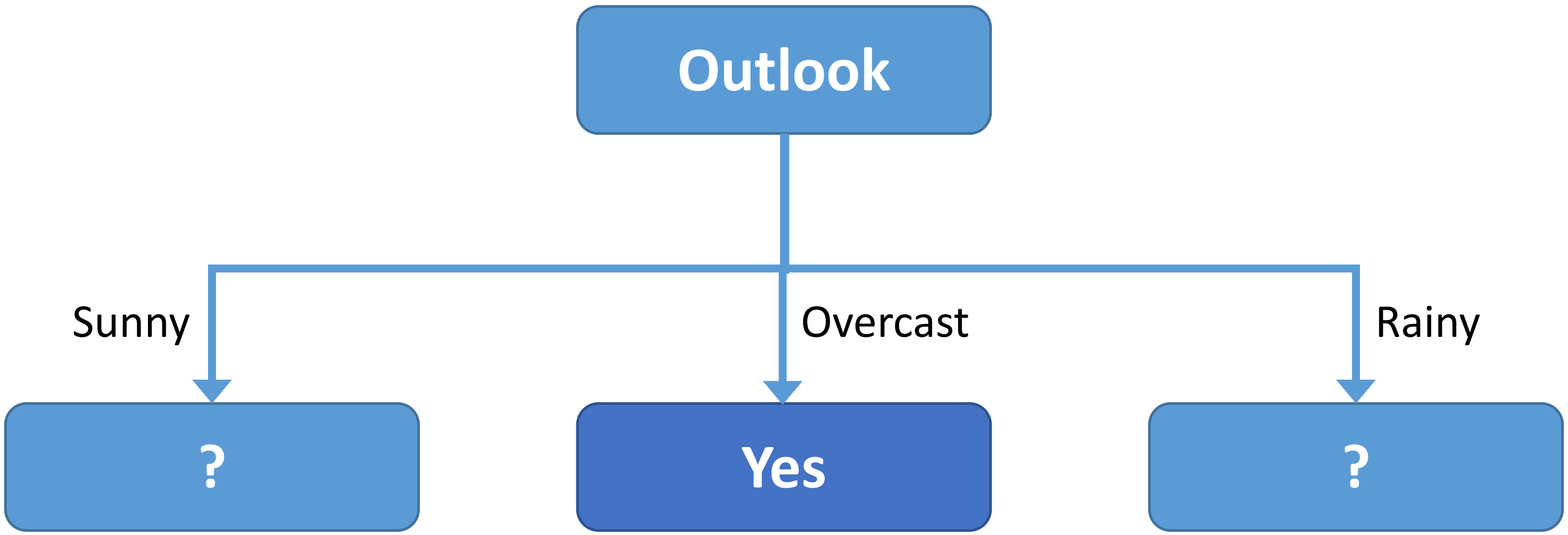
- Which split gives the highest **Information Gain** ?

Information Gain

- $IG(Play|Outlook) = 0.247$
- $IG(Play|Temperature) = 0.029$
- $IG(Play|Humidity) = 0.152$
- $IG(Play|Windy) = 0.048$

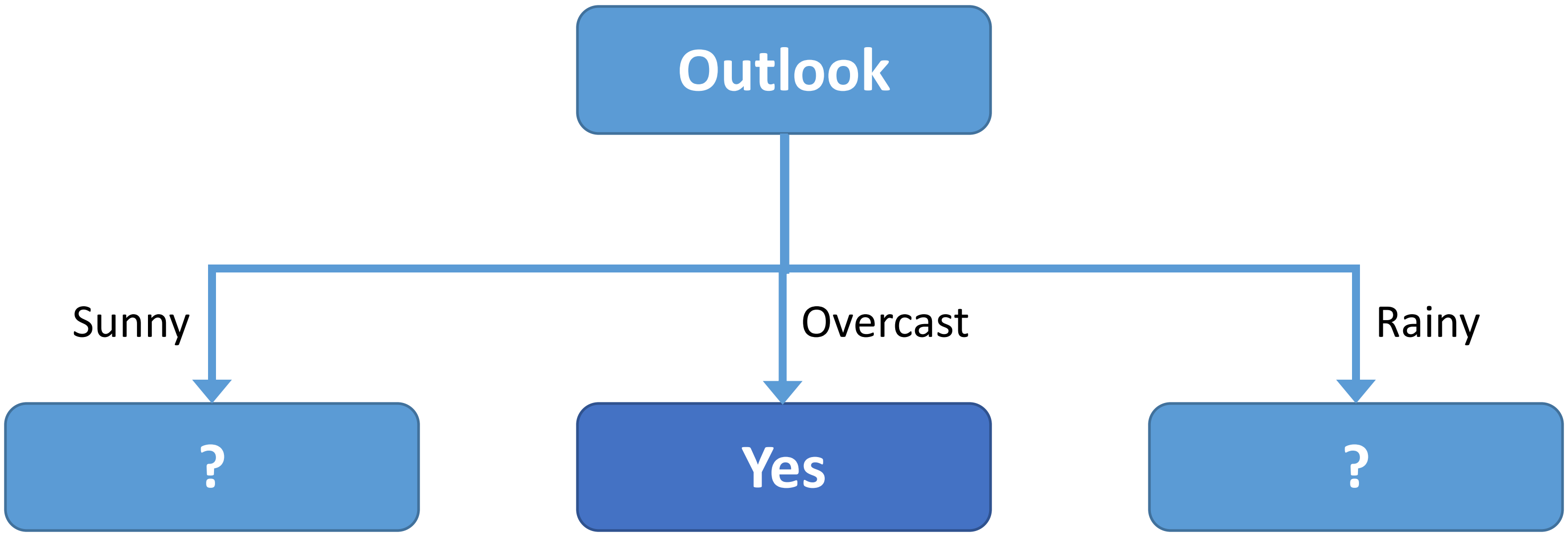
- Therefore, our root node is **Outlook**

Tree after 1st Split



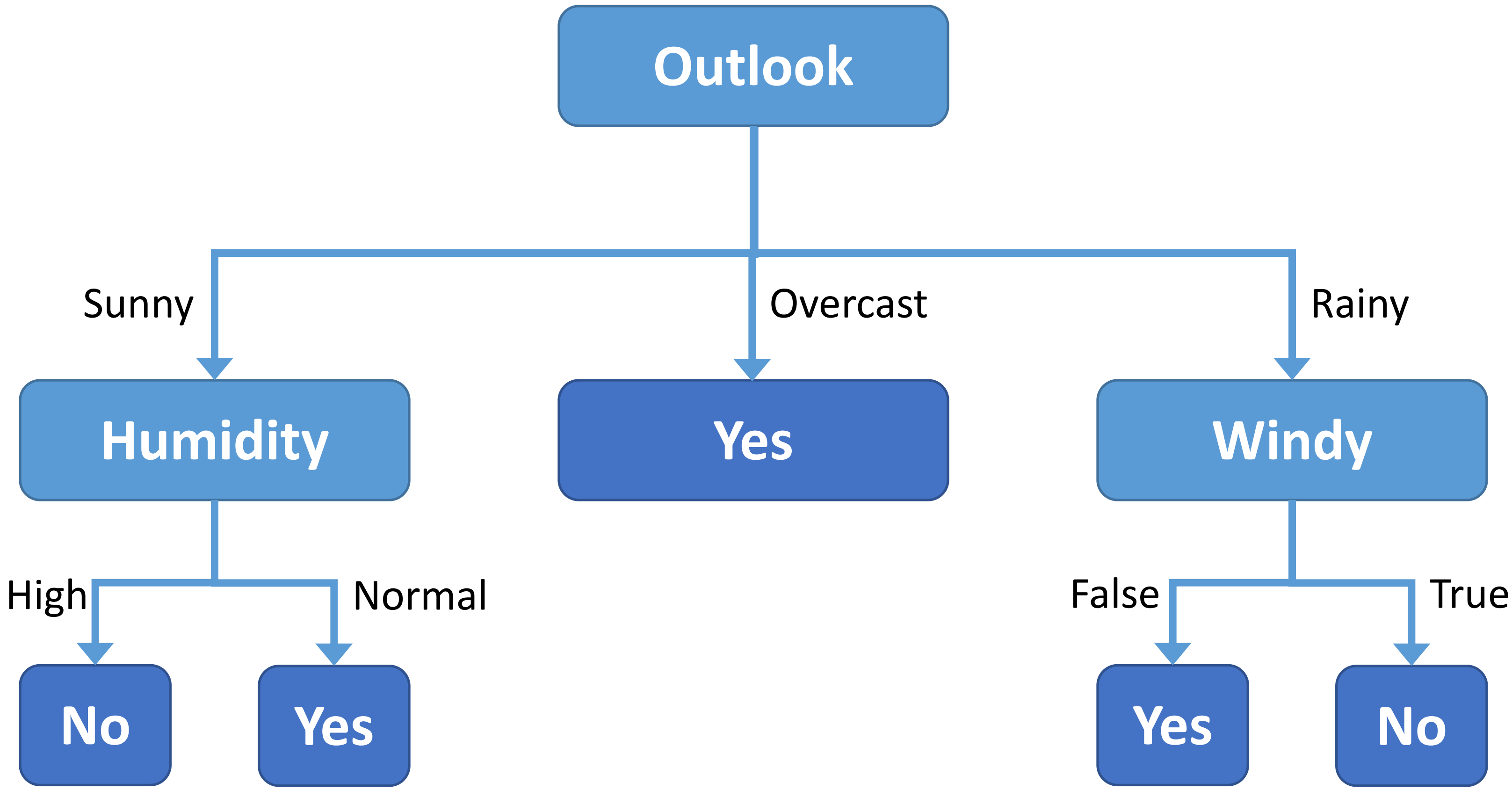
		Play Golf ?	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Tree after 1st Split



Now we have to repeat the same process for all non-leaf nodes.

Final Result



Reference

1. https://www.youtube.com/watch?v=YFHtj5pkvQw
2. https://www.youtube.com/watch?v=tu_TclzJIeM

Gini Impurity

- Gini Impurity measures the impurity of a set and the calculated value will be always reside between 0 and 1.
- Purity of a given set of items is considered 1, if we found two randomly picked items from the given set, are in the same class and probability for the occurrence of this event is 1.

Gini Impurity

- Gini impurity G, Number of classes c,
- $P(i)$ - probability of picking a datapoint with class i

$$G = 1 - \sum_{i=1}^{c} P(i)^2$$

or

$$G = \sum_{i=1}^{c} P(i)(1 - P(i))$$

Gini Impurity - Example

#	Outlook	Temperature	Humidity	Wind	Play Golf ?
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cold	Normal	False	Yes
6	Rainy	Cold	Normal	True	No
7	Overcast	Cold	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cold	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Gini Impurity – Before Splitting

- Gini Impurity before split

$$G = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2$$

$$G = 0.46$$

Gini Impurity: Split by Outlook

$$G_{Sunny} = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

$$G_{Overcast} = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0$$

$$G_{Rainy} = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$

		Play Golf ?	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

$$G(Play|Outlook) = \left(\frac{5}{14}\right)\times 0.48 + \left(\frac{4}{14}\right)\times 0 + \left(\frac{5}{14}\right)\times 0.48$$

$$G(Play|Outlook) = 0.34$$

Gini Impurity: Split by Temperature

$$G_{Hot} = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$$

$$G_{Mild} = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.44$$

$$G_{Cold} = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$$

		Play Golf ?	
		Yes	No
Temp	Hot	2	2
	Mild	4	2
	Cold	3	1

$$G(Temp.) = \left(\frac{4}{14}\right)\times 0.5 + \left(\frac{6}{14}\right)\times 0.44 + \left(\frac{4}{14}\right)\times 0.375$$

$$G(Temp.) = 0.44$$

Gini Impurity: Split by Humidity

$$G_{High} = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.49$$

$$G_{Normal} = 1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 = 0.24$$

		Play Golf ?	
		Yes	No
Humidity	High	3	4
	Normal	6	1

$$G(Humid.) = \left(\frac{7}{14}\right) \times 0.49 + \left(\frac{7}{14}\right) \times 0.24$$

$$G(Humid.) = 0.36$$

Gini Impurity: Split by Windy

$$G_{True} = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$$

$$G_{False} = 1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2 = 0.375$$

		Play Golf ?	
		Yes	No
Windy	True	3	3
	False	6	2

$$G(Windy) = \left(\frac{6}{14}\right)\times 0.5 + \left(\frac{8}{14}\right)\times 0.375$$

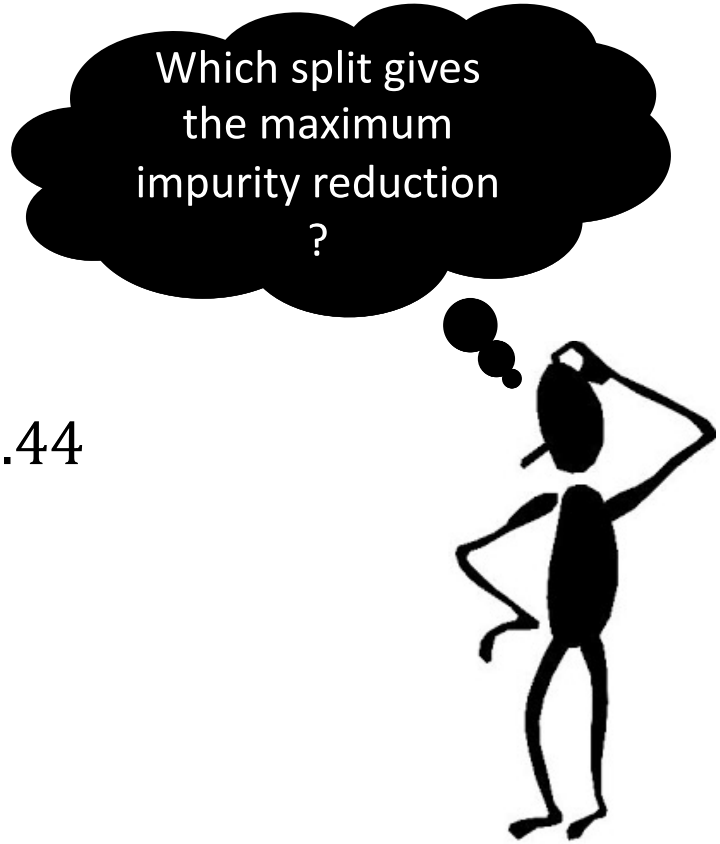
$$G(Windy) = 0.43$$

Gini Impurity Results

- Before Split
  - $G(Play) = 0.46$
- After Split
  - $G(Play|Outlook) = 0.34$
  - $G(Play|Temperature) = 0.44$
  - $G(Play|Humidity) = 0.36$
  - $G(Play|Windy) = 0.43$

Gini Impurity Results

- Before Split
  - $G(Play) = 0.46$
- After Split
  - $G(Play|Outlook) = 0.34$
  - $G(Play|Temperature) = 0.44$
  - $G(Play|Humidity) = 0.36$
  - $G(Play|Windy) = 0.43$



Reduction in Variance

- Employed when the target variable is Continuous (Maybe for regression problems)
- The standard way of variance calculation is used to determine the best split.
  - Calculate the variance for parent node before the split
  - Calculate the variance for each child node after the split.
  - Calculate the weighted average of variance for all child nodes after the split.
  - Compare variance reductions

Problem of Overfitting

A Decision Tree h, is said overfit with the training examples if there is a hypothesis h', that fits the same training examples less well, but performs better over the entire distribution of instances.

Problem of Overfitting (Cont.)

- Random noice or error
- Too many independent variables
- Too few samples are in the leaf nodes

Handling Overfittnig

- Pre-pruning
  - Tree is not allowed to fully grown. It stops the tree building before it produces leaves with few samples
  - May lead to underfit
- Post-pruning
  - Tree is allowed to fully grown and then it removes subtrees based on criteria.

Pre-pruning Strategies

- Specify minimum samples for a node
- Specify maximum depth for the tree
- Specify maximum leaf nodes

Post-Pruning

- The idea of pruning is to eliminate subtrees that do not make significant contribution for the final result.
- A tree that is too large, may become overfit
- Objective is to reduce the size of the tree without reducing the classification accuracy.

Common Approaches for Pruning

- Reduced Error Pruning

  Starting with leaf nodes, each node is replaced with its most popular class. If the prediction accuracy remains unchanged then the change will be made permanent.

- Cost Complexity Pruning

  Generates a series of trees where a preceding tree is created by removing a subtree of the succeeding tree based on an error measurement and replacing with a leaf node.

Cost Complexity Pruning

- It generates **a series of trees** T_0 to T_m where T_0 is the initial tree and T_m is the root node alone.
- At step i, the tree is created by removing a subtree from tree at step $i-1$ and replacing it with a leaf node with value chosen as in the tree building algorithm.

1. Define the error rate of tree T over dataset S as $E(T,S)$
2. The subtree that minimizes following function is chosen for removal

$$= \frac{E(prune(T,t),S) - E(T,S)}{|leaves(T)| - |leaves(prune(T,t))|}$$

- $prune(T,t)$ - The tree after pruning subtree t from the tree T

External References

- Decision Trees
  - https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/
- Pruning
  - https://www.youtube.com/watch?v=u4kbPtiVVB8
- Cost Complexity Pruning
  - http://mlwiki.org/index.php/Cost-Complexity_Pruning

Other Pruning Approaches

- Critical Value Pruning
- Minimum Error Pruning
- Pessimistic Error Pruning

Continuous Variables

- When the variable X is continuous, determining the split point will be a challenge.

Variable X	25	30	5	23.2	33	12.5	40	17	37	7
Target Y	High	Low	Low	High	Low	Medium	High	High	High	Medium

- Answer – Define several threshold levels based on the corresponding class labels

Continuous Variable Conversion

1. Order the dataset based on the values of X

Variable X	5	7	12.5	17	23.2	25	30	33	37	40
Target Y	Low	Medium	Medium	High	High	High	Low	Low	High	High

2. Mark the Target variable transition on X

Variable X	5	7	12.5	17	23.2	25	30	33	37	40
Target Y	Low	Medium	Medium	High	High	High	Low	Low	High	High

3. Take the average of the range of X where the transition is;

X’	6	14.75	27.5	35

$The\ potential\ candidates\ for\ the\ split\ can\ be\ 6, 14.75, 27.5, and\ 35.$

Missing Values

- Disregard all the instances with missing values
- Use the most common (highest frequency) category of the variable to the missing value.
- Use the most common (highest frequency) category of the variable among all the observations that have the same class of the target variable.
- Treat missing value as another category

Class Imbalance

- What if a dataset contains 95% records from the class A and rest from the class B?
- No split might be possible
- Better choice would be to estimate class A for every new record.

Class Imbalance (Cont.)

- Prior Probabilities (Priors) – Most of the time algorithms assume prior probabilities of classess are reflected by the data distribution.
  - Prior Probability of Class A – 0.95
  - Prior Probability of Class B – 0.05
- Some tree alogorithms allow to adjust these priors and minimize the affect of class imbalance
  - Set Prior Probabilities of Class A and B to 0.5

Class Imbalance (Cont.)

- Misclassification Cost – Most of the time algorithms assume misclassification cost of classes are equal.
- But we can inform the algorithm that the cost of misclassifcation of a class is different from other.
- By doing so, we can increase the influence even for an underrepresented class.

Class Imbalance (Cont.)

- Class A – 95% and Class B – 5%
- Ratio = 95:5 = 19:1
- We can say the misclassification of Class B record is 19 times cost expensive than Class A

	Class A	Class B
Class A	x	1
Class B	19	x

Misclassification Cost Matrix

Decision Tree Algorithms

Algorithm	Splitting Criterion	Input variables	Split Type	Complexity Regularization
ID3	Information Gain (Entropy)	Categorical or Continuous	Multi	
C5.0	Gain Ration (Entropy)	Categorical or Continuous	Multi	Pruning
CART	Gini Impurity	Categorical or Continuous	Binary	Pruning
CHAID	Chai-Squire Test	Categorical	Multi	Pre-Pruning

Decision Tree - Controls

- Maximum Depth: The maximum number of levels that the tree can grow.
- Minimum Samples in a Leaf Node: The minimum number of instances for terminal nodel.
- Minimum Samples in a Parent Node:

General Information

- Since decision trees adopt greedy approach, it might end up in a suboptimal structure.
- If variables do not work well with splitting then the entire tree will be ineffective.
- Due to the fact that a small change in training data can produce a significant difference, trees are considered weak learners or unstable models.

General Information (Cont.)

- In general, SVMs and ANNs produce better accuracy than a single tree.
- Ensemble of trees may produce better accuracy.

Assignment 2

- Concise **report** on "Random Forest"
  - How it works (Theoritical Background)
  - An application with real world dataset
  - Source Code

- Strict Penalties for Plagiarism

Thank You…!