Machine Learning

Supervised Learning- Decision Trees

Part One

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

- Introduction Decision Tree
- Constructing Decision Tree
- Decision Tree ID3 Algorithm

Decision Trees: In Nutshell

- Supervised Learning Technique.
- It is an an eager learning method.
- It works for classification problems and regression problems
- It works for binary and multiclass classification tasks.
- It works for both categorical and continuous features.
- It converts the training data (table or matrix structure) into a tree structure that we can use to classify or estimate new data.

Decision Trees: In Nutshell

- The key idea is to split the data into two or more homogeneous sets based on the features set.
- To predict the class or the continuous value of any new data sample we follow one path in the tree from the root to any leaf node .
- The decision tree does not make strong assumptions about the characteristics of the training data. It is a nonparametric classification method.

Decision Tree: Key Terminologies

- Root Node: The entire dataset (samples) are grouped by the root node.
- Decision Node: non-leaf nodes where the dataset are splitted into sub-trees.
- Terminal Node: is a leaf node that does not split the dataset and it has the final decision (label or continuous value).
- Splitting: is the process of dividing the dataset into two or more subsets.
- Pruning: is the process of removing one or more decision node from the tree.

Decision Tree: Pros and Cons

Pros:

- Easy to interpret; it easy to trace and understand the output of the decision tree.
- Noise Tolerance; can handle noise and outliers to a certain degree.
- Works for classification and regression problems
- Features are not limited to specific types (categorical or numerical)
- Nonparametric method.
- Support both binary and multiclass classification tasks.

Decision Tree: Pros and Cons

Cons:

- Over fitting: It can easily result in overfitting the training data. Certain techniques must be applied to avoid overfitting.
- Precision Issue: while it works with continuous features, it loses information when it categorizes continuous variables

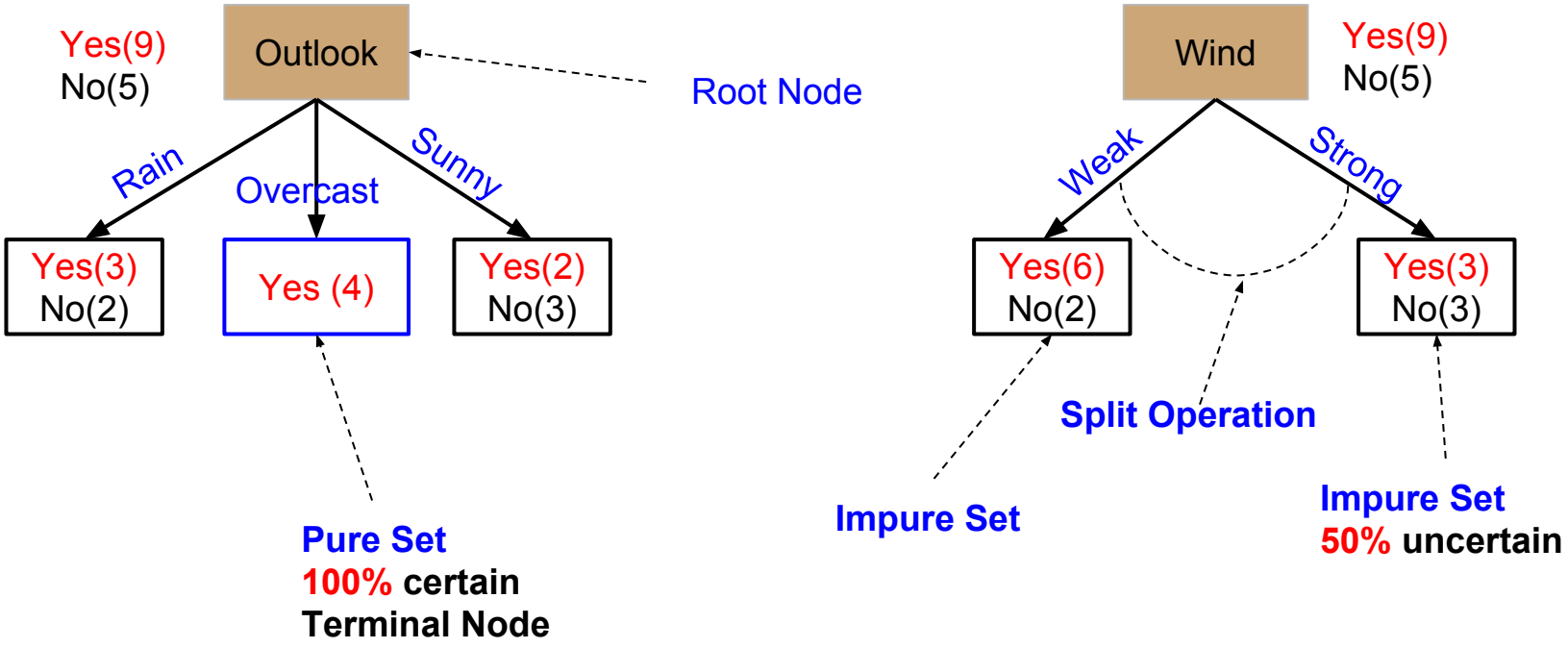
Decision Trees: Tennis Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

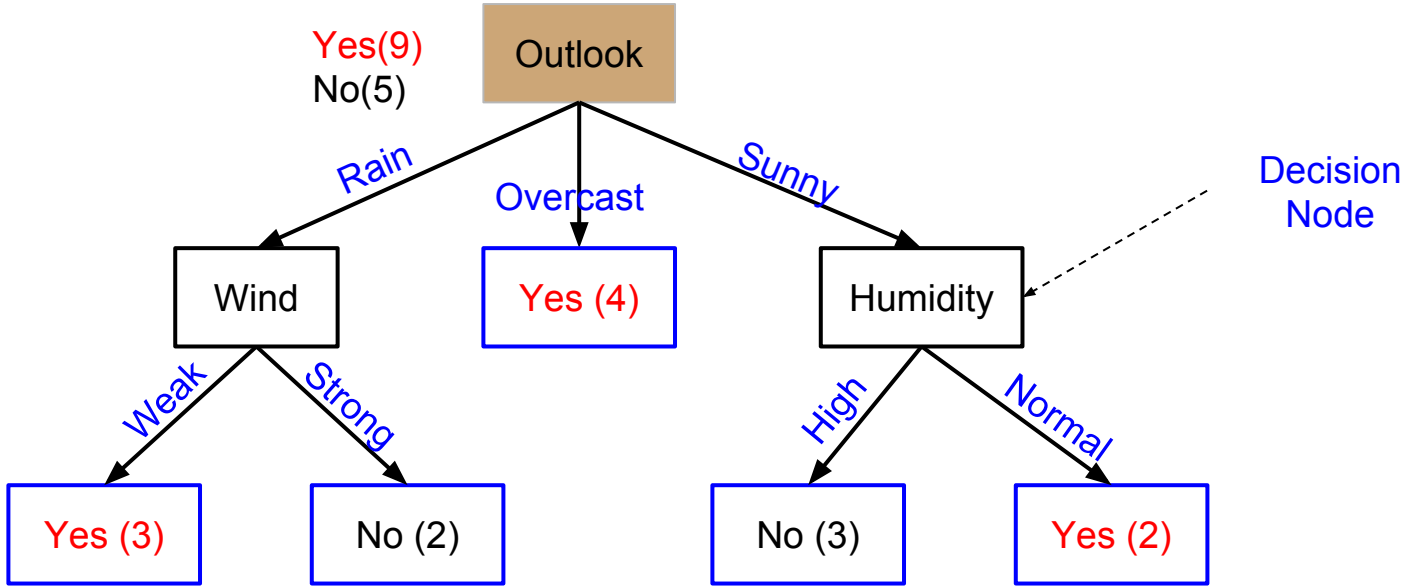
Decision Trees: Tennis Example

- Training Dataset: 14 samples
- Features: Outlook, Temperature, Humidity, Wind
- Decision: Play Tennis → {Yes, No}
- Features Values:
  - Outlook = {Sunny, Overcast, Rain}
  - Temperature = {Hot, Mild, Cold}
  - Humidity = {Normal, High}
  - Wind = {Weak, Strong}

Decision Trees: Tennis Example

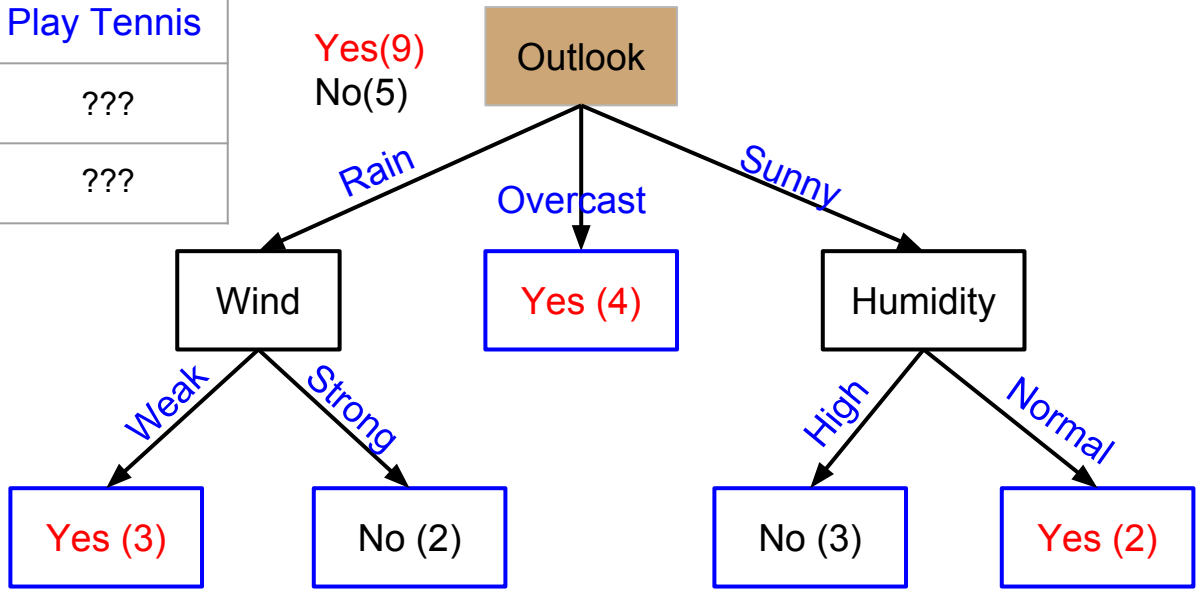


Decision Trees: Tennis Example



Decision Trees: Tennis Example

Outlook	Temp	Humidity	Wind	Play Tennis
Sunny	Cool	High	Strong	???
Overcast	Mild	High	Weak	???



In-Class Exercise

Given the play tennis example pick any other feature (not the outlook) as the root of the decision tree, build a decision tree.

Decision Trees: How it works

Which features we should use to split the training data and in what order to build the decision tree?

- We should use the feature that results in a good split

What is a good split?

- A good split increase the ***purity*** of the sets, increase certaining and decrease uncertainty.

How we measure the purity of the split?

Decision Tree: Learning & Feature Selection

The learning objective of a decision tree aims at finding the smallest tree structure.

We achieve that by recursively choosing the most significant feature as the root of the tree and every subtree.

The most significant feature is the feature that increase the purity of the split.

There are several techniques to select the most significant feature. Different techniques works for different target variables and features.

Decision Tree: Learning & Feature Selection

The learning objective of a decision tree aims at finding the smallest tree structure.

We achieve that by recursively choosing the most significant feature as the root of the tree and every subtree.

The most significant feature is the feature that increase the purity of the split.

There are several techniques to select the most significant feature. Different techniques works for different target variables and features.

Decision Tree: Learning & Feature Selection

How do we measure purity or quantify uncertainty?

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

Information Gain: Entropy

Entropy: measure the amount of uncertainty in a probability distribution.

Entropy or Information content refers to disorder or disorganization in a system.

If the set is pure then the entropy is zero and if the set is impure and the samples in the set equally divided (50% true 50% false) the entropy is **1**

Entropy is highest when the uncertainty is greatest

Information Gain: Entropy

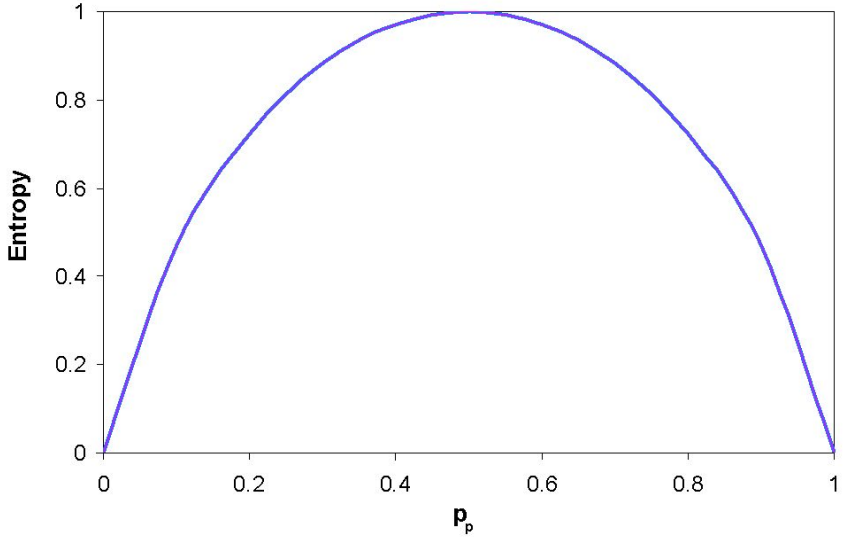
Entropy is calculated using the formula

$$Entropy = -p \cdot log_2 \cdot p - q \cdot log_2 \cdot q$$

P: is the probability of success

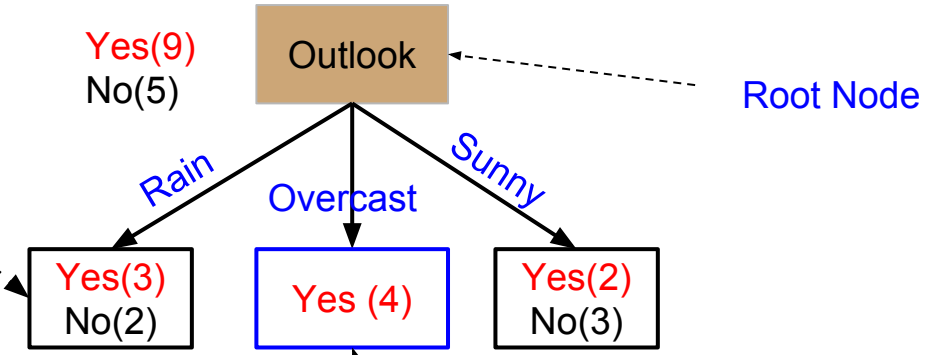
Q: is the probability of failure

We need to select the split which has the lowest entropy



Information Gain: Entropy

$$Entropy = -\frac{3}{5} \cdot log_2 \cdot \frac{3}{5} - \frac{2}{5} \cdot log_2 \cdot \frac{2}{5}$$



$$Entropy = -\frac{4}{4} \cdot log_2 \cdot \frac{4}{4} - \frac{0}{4} \cdot log_2 \cdot \frac{0}{4}$$

Pure Set
100% certain
Terminal Node

Information Gain: Entropy

Entropy tell us how pure or impure each subset after splitting.

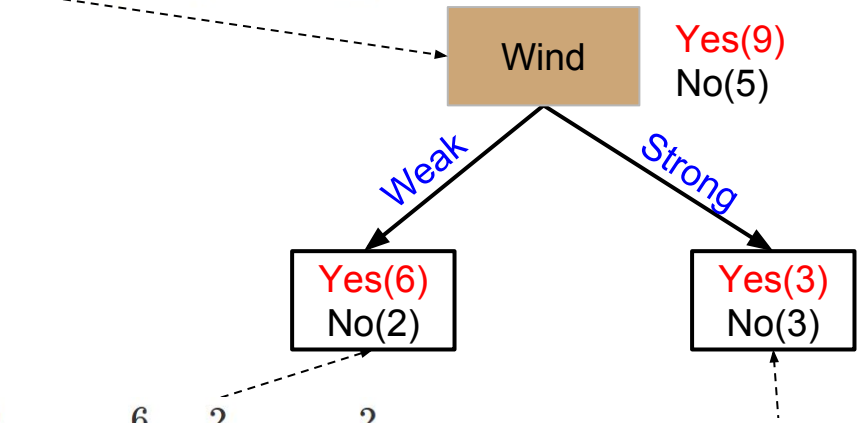
We want to use this information to decide the significance of the feature. We can do that be calculating the information gain.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{s} \frac{|s_i|}{|S|} Entropy(s_i)$$

The more gain the better the feature we use to split the data

Information Gain: Entropy

$$Entropy(wind) = -\frac{9}{14} \cdot log_2 \cdot \frac{9}{14} - \frac{5}{14} \cdot log_2 \cdot \frac{5}{14} = 0.94$$



$$Entropy(Weak) = -\frac{6}{8} \cdot log_2 \cdot \frac{6}{8} - \frac{2}{8} \cdot log_2 \cdot \frac{2}{8} = 0.81$$

$$Entropy(Strong) = -\frac{3}{6} \cdot log_2 \cdot \frac{3}{6} - \frac{3}{6} \cdot log_2 \cdot \frac{3}{6} = 1$$

$$Gain(S, Wind) = Entropy(Wind) - \frac{8}{14} Entropy(weak) - \frac{6}{14} Entropy(Strong) = 0.049$$

Decision Tree: ID3 Learning Algorithm

1. Create a root node
2. If all samples are positive, return the root node with label = +
3. If all samples are negative, return the root node with label = -
4. If the number of predicting features is empty then return the root node with label = the most common value of the target attribute in the training set.

Decision Tree: ID3 Learning Algorithm

5. Calculate the entropy of very feature using the training dataset ***S***
6. Split the training set S into subset based on the most significant feature, ***f*** where the entropy after splitting is the minimum or the information gain is the maximum.
7. For each value v in f:
    a. create a new edge
    b. Subset the data based on v and add
    c. If the subset is pure, create a terminal node and assign to it the label of the pure set
    d. If the subset is not pure then split (node, subset of samples)

Decision Tree: ID3 Learning Algorithm

Does the ID3 algorithm use uninformed search or informed search?

Is it optimal or not and why?

Questions