Decision Tree

Weather Prediction

Bikash Singh

Roll No: 21MA60R02

Masters in Technology

in

Computer Science and Data Processing

at

Indian Institute of Technology Kharagpur

Department of Mathematics
Indian Institute of Technology Kharagpur
West Bengal - 721302, India

December, 2021

Abstract

This is a report on Decision Tree Algorithm. At the very beginning we will know about Decision Tree and its structures. After that we will see why we need Decision Tree and how to build it. After that we will introduce entropy, Information Gain and Gini Index. Then we will see two algorithms to build Decision Tree. Here is one example to explain the algorithms. Here we will implement this in python code. Finally we will see some drawbacks of decision Tree.

Contents

Chapter 1

Introduction

In this chapter we will see definition of Decision Tree, motivation of Decision tree and how to build a Decision Tree.

Definition

Decision Tree is like a tree structure in which each **internal node** represents a test on features and each branch represents an outcome of the test and each **leaf node** holds a label.

So mainly Decision Tree look like tree structure and it works like flowchart.

Root Node: It is the node present in the starting of a Decision Tree.From this the whole data set starts diving depending on the features.

Internal node: The nodes we get after splitting the the root nodes are called Internal Nodes. This is also known as Decision Nodes.

Leaf Node: The nodes where the further splitting is not possible are called Leaf Nodes. This is also known as Terminal Node.

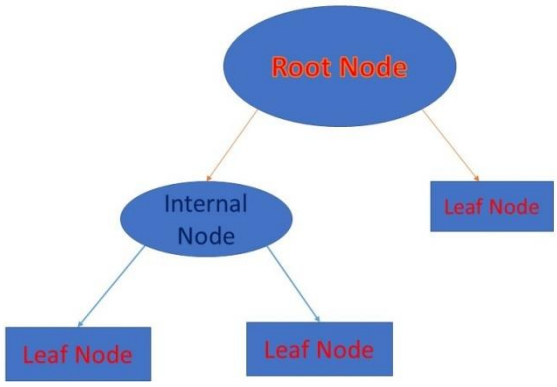

Figure 1.1: Decision Tree

Motivation

Decision Tree is a Supervised Learning Technique that can be used for both Classification and Regression problems. But mostly it is preferred to solve Classification problems.

Classification: Classification is the process of dividing the data sets into different categories by adding some labels.

Example: If we want to divide a whole class into two labels - M.Sc. Mathematics and B.Tech.

Why use Decision Tree?

We use Decision Trees for some following reasons-

1. Decision Tree is easy to explain
2. Decision Tree requires less effort for data preperation during pre-processing.

3. Decision Tree can handle multidimensional data
4. Missing values in the data also do not effect the process of Decision Tree

How to build a Decision Tree

So in this section we will know the process of making a decision tree. The steps are as follows -

Step 1: Begin with a root node which contains the whole data set

Step 2: Find the best attribute in the data set

Step 3: Divide the root node into subsets depending on the possible values of the best attribute

Step 4: Generate the Decision Tree node which contains the possible values of the best attribute

Step 5: Now take all these decision tree nodes as root node and make the corresponding decision trees.Do this until no further division is possible.

So now the important thing is that how can we determine that which attribute will be the best attribute.

This problem can be solved using a technique known as Attribute Selection Measure (ASM). By this measurement we can easily select the best attribute for the nodes of the tree. There are two popular techniques of ASM -

1. Information Gain
2. Gini Index

Before knowing these two techniques we should know Entropy.

Entropy

Entropy is a metric to measure the uncertainty in a data set. Mainly entropy specifies the randomness in data.

Mathematical Expression: Let S be a data set and C is the set of attributes of S then the Entropy of S,

$$H(S) = -\sum_{c \in C} p(c) \log_2 p(c)$$

where p(c) is the probability of c.

When we use node in a Decision Tree to partition the training set of instances into smaller subsets then Entropy changes. Information Gain is a measure of this change in Entropy.

Information gain: Information gain of A is a measure of change in entropy from before to after a set S is split on the attribute A .

$$I.G(S, A) = H(S) - \sum_{S_i} p(S_i) H(S_i)$$

where S_i are the partitions of S after splitting it on attribute A.

Gini Index

Gini Index is also a measure of uncertainty in a data set. It also specifies the randomness of the data set.

Mathematical Expression: If target variables take N different labels and $p(i)$ are the probability of the label i in the set S then Gini Index of S,

$$G(S) = 1 - \sum_{i=1}^{N} p(i)^2$$

Chapter 2

Algorithms

In this chapter we will give two algorithms to construct a Decision Tree. Here we will study the ID3 algorithm and the CART algorithm and we will see the application of the algorithms by one example and drawbacks of decision tree.

ID3 Algorithm

The full name of this algorithm is Iterative Dichotomiser 3.This algorithm uses Entropy and Information Gain as metric. This algorithm decides which attribute to be used to classify the current subset of data using Information Gain.

Algorithm:

1. Calculate H(S)
2. For every attribute A calculate I.G(S,A)
3. Choose the highest I.G attribute as best attribute

4. Repeat

CART Algorithm

The full name of this algorithm is Classification and Regression Tree. This algorithm uses Gini Index as metric. This algorithm decides which attribute to be used to classify the current subset of data using Average Gini Index.

Algorithm:

1. Calculate G(S)
2. For every attribute calculate average Gini Index
3. Choose the lowest average Gini Index attribute as best attribute
4. Repeat

Example

In this section we will see how can we construct Decision Trees using these two Algorithms. Here is one data of weathers with corresponding status of playing cricket. Now using the algorithms on this data we will construct a decision tree which will help us to predict the status of playing cricket for a new data on weathers.

SL. No.	Fever	Breathing Issue	Cough	Covid
1	Yes	Yes	Yes	Yes
2	No	No	No	No
3	Yes	Yes	No	Yes
4	Yes	No	Yes	No
5	Yes	Yes	Yes	Yes
6	No	No	Yes	No
7	Yes	Yes	No	Yes
8	Yes	Yes	No	Yes
9	No	Yes	Yes	Yes
10	Yes	No	Yes	Yes
11	No	No	Yes	No
12	No	Yes	Yes	Yes
13	Yes	No	Yes	No
14	No	Yes	Yes	No

Figure 2.1: Weather Data
[**Data from "Machine Learning" by T. Mitchell**]

ID3

Here S=whole weather data set and C={Yes,No}

So entropy of the whole data set,

$$H(S) = -\Sigma_{c \in C}\, p(c) \log_2 p(c)$$

$$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$$

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

$$= 0.94$$

Divide the data set based on outlook

Entropy of sunny outlook is

$H(S_1) = H(outlook = sunny) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{2}{5} \log_2(\frac{2}{5}) - \frac{3}{5} \log_2(\frac{3}{5}) = 0.971$

And probability, $p(S_1) = \frac{5}{14}$

Entropy of overcast outlook is

$H(S_2) = H(outlook = overcast) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{4}{4} \log_2(\frac{4}{4}) - 0 \log_2(0)$

$= 0$

And probability, $p(S_2) = \frac{4}{14}$

Entropy of rainy outlook is

$H(S_3) = H(outlook = rainy) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{3}{5} \log_2(\frac{3}{5}) - \frac{2}{5} \log_2(\frac{2}{5})$

$= 0.971$

And probability, $p(S_3) = \frac{5}{14}$

Average entropy of outlook is

$\frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971 = 0.693$

Information Gain of the attribute outlook is

$I.G = H(S) - \sum_{i=1}^{3} p(S_i)H(S_i) = 0.94 - 0.693 = 0.247$

Now divide the data set based on temp

Entropy of hot temp is

$H(S_1) = H(temp = hot) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{1}{2} \log_2(\frac{1}{2}) - \frac{1}{2} \log_2(\frac{1}{2}) = 1$

And probability, $p(S_1) = \frac{4}{14}$

Entropy of mild temp is

$H(S_2) = H(temp = mild) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{2}{3} \log_2(\frac{2}{3}) - \frac{1}{3} \log_2(\frac{1}{3}) = 0.9183$

And probability, $p(S_2) = \frac{6}{14}$

Entropy of cool temp is

$H(S_3) = H(temp = cool) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{3}{4} \log_2(\frac{3}{4}) - \frac{1}{4} \log_2(\frac{1}{4}) = 0.8113$

And probability, $p(S_3) = \frac{4}{14}$

Average entropy of temp is

$\frac{4}{14} \times 1 + \frac{6}{14} \times 0.9183 + \frac{4}{14} \times 0.8113 = 0.9111$

Information Gain of the attribute temp is

$I.G = H(S) - \sum_{i=1}^{3} p(S_i)H(S_i) = 0.94 - 0.911 = 0.029$

Now divide the data set based on humidity

Entropy of high humidity is

$H(S_1) = H(humidity = high) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{3}{7} \log_2(\frac{3}{7}) - \frac{4}{7} \log_2(\frac{4}{7})$

And probability, $p(S_1) = \frac{1}{2}$

Entropy of normal humidity is

$H(S_2) = H(humidity = normal) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{1}{7} \log_2(\frac{1}{7}) - \frac{6}{7} \log_2(\frac{6}{7})$

And probability, $p(S_2) = \frac{1}{2}$

Average entropy of humidity is 0.788

Information Gain of the attribute humidity is

$I.G = H(S) - \sum_{i=1}^{2} p(S_i)H(S_i) = 0.94 - 0.788 = 0.152$

Now divide the data set based on windy

Entropy of false windy is

$H(S_1) = H(windy = false) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{6}{8} \log_2(\frac{6}{8}) - \frac{2}{8} \log_2(\frac{2}{8})$

And probability, $p(S_1) = \frac{8}{14}$

Entropy of true windy is

$H(S_1) = H(windy = true) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{3}{6} \log_2(\frac{3}{6}) - \frac{3}{6} \log_2(\frac{3}{6})$

And probability, $p(S_2) = \frac{6}{14}$

Average entropy of windy is 0.892

Information Gain of the attribute windy is

$I.G = H(S) - \sum_{i=1}^{2} p(S_i)H(S_i) = 0.94 - 0.892 = 0.048$

So from here we can see that Information Gain is maximum for the attribute outlook.

Therefore outlook will be the root node of the decision tree.

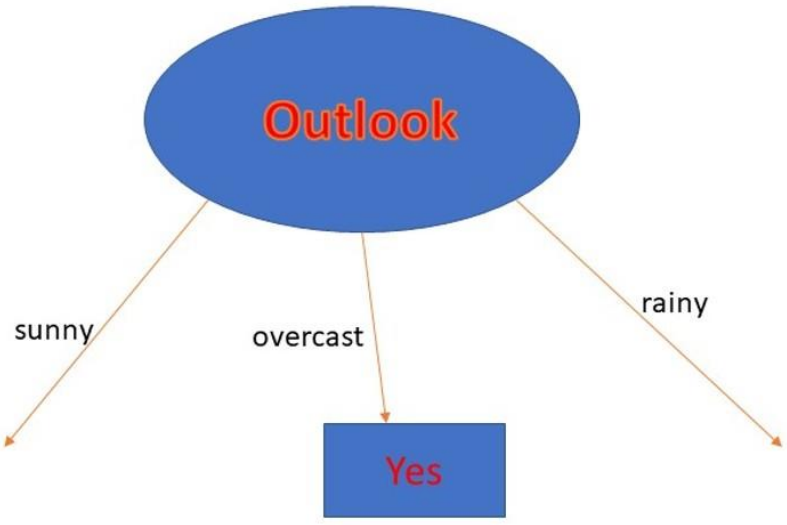

Figure 2.2: Decision Tree based on outlook

If we see the overcast outlook from the data set, then it always follow that Yes we can play cricket. So in the decision tree it is also going only to Yes. Now we will calculate for the branch sunny. At first we will reduce the data set to sunny outlook. At first divide the data set based on temp

outlook	temp	humidity	windy	play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Sunny	Mild	Normal	True	Yes

Figure 2.3: Data based on Sunny

Entropy of hot temp is

$H(S_1) = H(temp = hot|sunny) = -\sum_{c\in C} p(c)\log_2 p(c)$

$= -p(Yes)\log_2 p(Yes) - p(No)\log_2 p(No)$

$= -1\log_2(1) = 0$

And probability, $p(S_1) = \frac{2}{5}$

Entropy of mild temp is

$H(S_2) = H(temp = mild|sunny) = -\sum_{c\in C} p(c)\log_2 p(c)$

$= -p(Yes)\log_2 p(Yes) - p(No)\log_2 p(No)$

$= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1$

And probability, $p(S_2) = \frac{2}{5}$

Entropy of cool temp is

$H(S_3) = H(temp = cool|sunny) = -\sum_{c\in C} p(c)\log_2 p(c)$

$= -p(Yes)\log_2 p(Yes) - p(No)\log_2 p(No)$

$= -1 \log_2(1) = 0$

And probability, $p(S_3) = \frac{1}{5}$

Average entropy of temp is

$\frac{2}{5} \times 0 + \frac{2}{5} \times 1 + \frac{1}{5} \times 0 = 0.4$

Information Gain of the attribute temp is

$I.G = H(S) - \sum_{i=1}^{3} p(S_i)H(S_i) = 0.971 - 0.4 = 0.571$

Now divide the data set based on humidity

Entropy of high humidity is

$H(S_1) = H(humidity = high|sunny) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{3}{3} \log_2(\frac{3}{3})$

And probability, $p(S_1) = \frac{3}{5}$

Entropy of normal humidity is

$H(S_2) = H(humidity = normal|sunny) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{2}{2} \log_2(\frac{2}{2})$

And probability, $p(S_2) = \frac{2}{5}$

Average entropy of humidity is 0

Information Gain of the attribute humidity is

$I.G = H(S) - \sum_{i=1}^{2} p(S_i)H(S_i) = 0.971 - 0 = 0.971$

Now divide the data set based on windy

Entropy of true windy is

$H(S_1) = H(windy = true|sunny) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{1}{2} \log_2(\frac{1}{2}) - \frac{1}{2} \log_2(\frac{1}{2}) = 1$

And probability, $p(S_1) = \frac{2}{5}$

Entropy of false windy is

$H(S_2) = H(windy = false|sunny) = -\sum_{c \in C} p(c) \log_2 p(c)$

$= -p(Yes) \log_2 p(Yes) - p(No) \log_2 p(No)$

$= -\frac{2}{3} \log_2(\frac{2}{3}) - \frac{1}{3} \log_2(\frac{1}{3}) = 0.9183$

And probability, $p(S_2) = \frac{3}{5}$

Average entropy of windy is $\frac{2}{5} \times 1 + \frac{3}{5} \times 0.9183 = 0.951$

Information Gain of the attribute windy is

$I.G = H(S) - \sum_{i=1}^{2} p(S_i) H(S_i) = 0.971 - 0.951 = 0.020$

So from here we can see that Information Gain is maximum for the attribute humidity.

Therefore humidity will be the internal node of the decision tree.

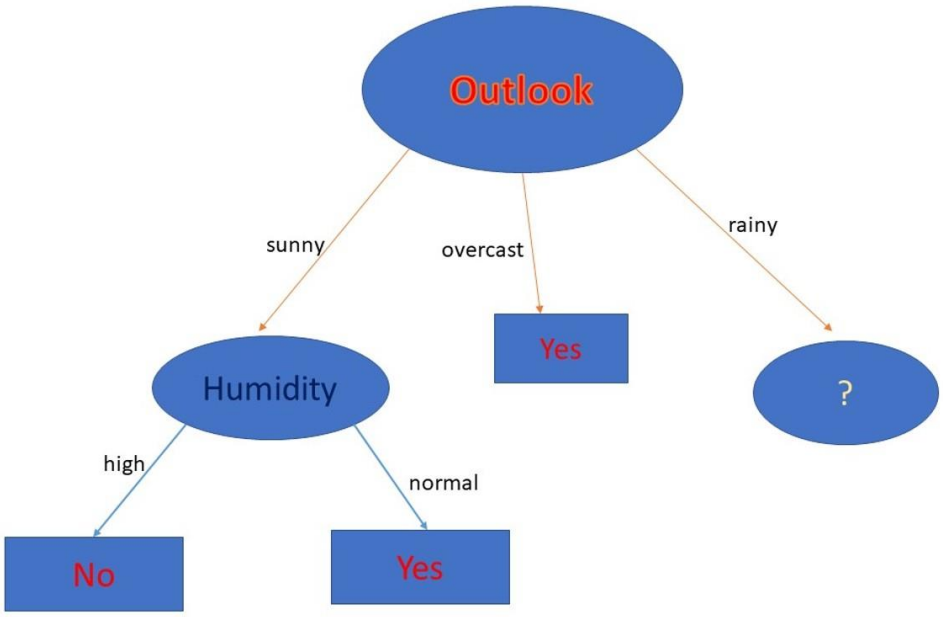

Figure 2.4: Decision Tree based on Sunny

Now we will calculate for the branch rainy. At first we will reduce the data set to rainy outlook.

outlook	temp	humidity	windy	play
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Rainy	Mild	Normal	False	Yes
Rainy	Mild	High	True	No

Figure 2.5: Data based on Rainy

At first divide the data set based on temp

Entropy of mild temp is

$$H(S_1) = H(temp = mild|rainy) = -\Sigma_{c\in C}\, p(c)\log_2 p(c)$$

$$= -p(Yes)\log_2 p(Yes) - p(No)\log_2 p(No)$$

$$= -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = 0.9183$$

And probability, $p(S_1) = \frac{3}{5}$

Entropy of cool temp is

$H(S_2) = H(temp = cool|rainy) = -\Sigma_{c\in C}\, p(c)\log_2 p(c)$

$= -p(Yes)\log_2 p(Yes) - p(No)\log_2 p(No)$

$= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1$

And probability, $p(S_2) = \frac{2}{5}$

Average entropy of temp is

$\frac{3}{5} \times 0.9183 + \frac{2}{5} \times 1 = 0.951$

Information Gain of the attribute temp is

$I.G = H(S) - \Sigma_{i=1}^{2}\, p(S_i)H(S_i) = 0.971 - 0.951 = 0.020$

Now divide the data set based on windy

Entropy of true windy is

$H(S_1) = H(windy = true|rainy) = -\Sigma_{c\in C}\, p(c)\log_2 p(c)$

$= -p(Yes)\log_2 p(Yes) - p(No)\log_2 p(No)$

$= -1\log_2(1) = 0$

And probability, $p(S_1) = \frac{3}{5}$

Entropy of false windy is

$H(S_2) = H(windy = false|rainy) = -\Sigma_{c\in C}\, p(c)\log_2 p(c)$

$= -p(Yes)\log_2 p(Yes) - p(No)\log_2 p(No)$

$= -1\log_2(1) = 0$

And probability, $p(S_2) = \frac{2}{5}$

Average entropy of windy is 0

Information Gain of the attribute windy is

$$I.G = H(S) - \Sigma_{i=1}^{2} p(S_i)H(S_i) = 0.971 - 0 = 0.971$$

So here is no need to calculate for humidity.

So from here we can see that Information Gain is maximum for the attribute humidity.

Therefore humidity will be the internal node of the decision tree.

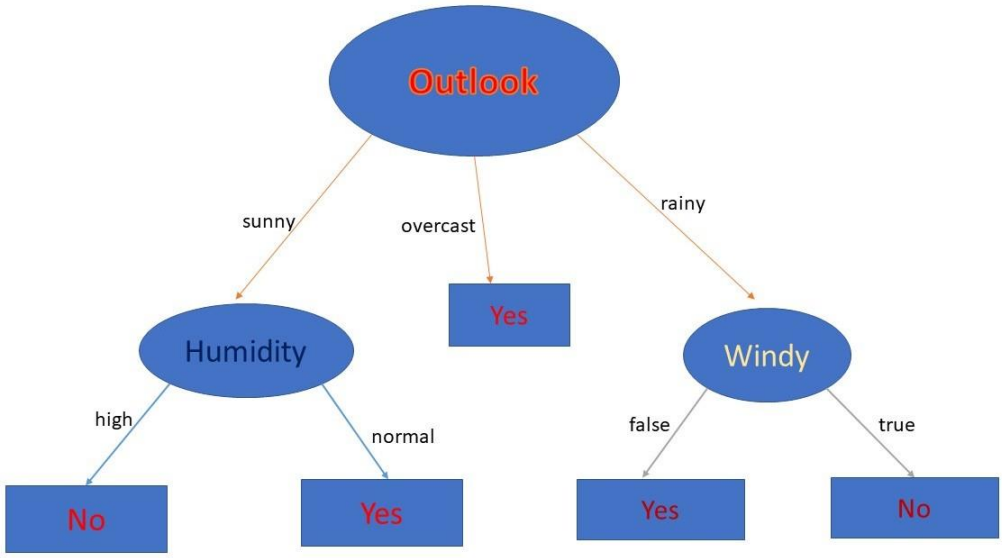

Figure 2.6: Decision Tree based on Whole Data

CART

Here S=whole weather data set and C={Yes,No}

So Gini Index of the whole data set is

$$G(S) = 1 - \Sigma_{c \in C}\, p(c)^2$$

$$= 1 - p(Yes)^2 - p(No)^2$$

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

$$= 0.46$$

Now divide the data set based on outlook

Gini Index of sunny outlook is

$$G(\text{outlook=sunny}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = \frac{12}{25}$$

$$p(\text{outlook=sunny}) = \frac{5}{14}$$

Gini Index of overcast outlook is

$$G(\text{outlook=overcast}) = 1 - \left(\frac{5}{5}\right)^2 = 0$$

$$p(\text{outlook=overcast}) = \frac{4}{14}$$

Gini Index of rainy outlook is

$$G(\text{outlook=rainy}) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = \frac{12}{25}$$

$$p(\text{outlook=rainy}) = \frac{5}{14}$$

$$\text{Average Gini} = \frac{5}{14} \times \frac{12}{25} + 0 \times \frac{4}{14} + \frac{5}{14} \times \frac{12}{25} = 0.343$$

Now divide the data set based on temp

Gini Index of hot temp is

$$G(\text{temp=hot}) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = \frac{1}{2}$$

p(temp=hot)=$\frac{4}{14}$

Gini Index of mild temp is

G(temp=mild)=$1-(\frac{4}{6})^2-(\frac{2}{6})^2=\frac{4}{9}$

p(temp=mild)=$\frac{6}{14}$

Gini Index of cool temp is

G(temp=cool)=$1-(\frac{3}{4})^2-(\frac{1}{4})^2=\frac{6}{16}$

p(temp=cool)=$\frac{4}{14}$

Average Gini=$\frac{4}{14}\times\frac{1}{2}+\frac{4}{9}\times\frac{6}{14}+\frac{4}{14}\times\frac{6}{16}$=0.44

Now divide the data set based on humidity

Gini Index of high humidity is

G(humidity=high)=$1-(\frac{3}{7})^2-(\frac{4}{7})^2=\frac{24}{49}$

p(humidity=high)=$\frac{1}{2}$

Gini Index of normal humidity is

G(humidity=normal)=$1-(\frac{1}{7})^2-(\frac{6}{7})^2=\frac{12}{49}$

p(humidity=normal)=$\frac{1}{2}$

Average Gini=$\frac{1}{2}\times\frac{24}{49}+\frac{1}{2}\times\frac{12}{49}$=0.37

Now divide the data set based on windy

Gini Index of false windy is

G(windy=false)=$1-(\frac{2}{8})^2-(\frac{6}{8})^2=\frac{6}{16}$

p(windy=false)=$\frac{8}{14}$

Gini Index of true windy is

G(windy=true)=$1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = \frac{1}{2}$

p(windy=true)=$\frac{6}{14}$

Average Gini=$\frac{6}{16} \times \frac{8}{14} + \frac{1}{2} \times \frac{6}{14}$=0.429

So the Average Gini Index is minimum for the attribute outlook. Therefore outlook will be the root node for the decision tree.

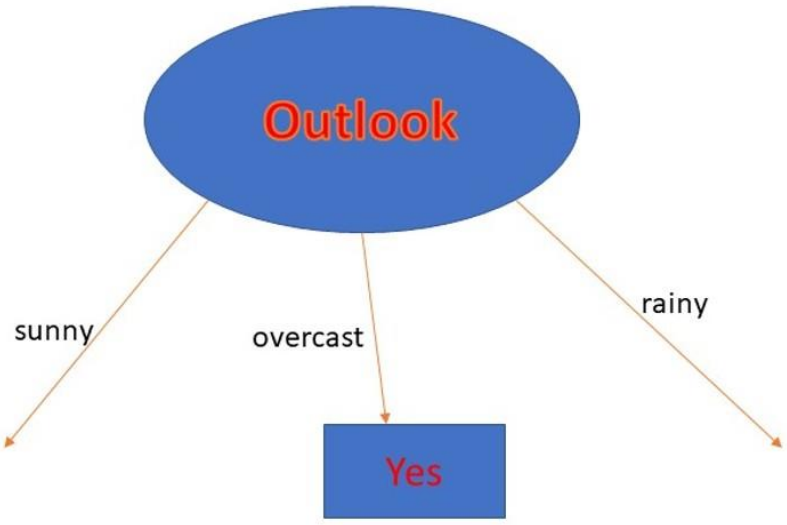

Figure 2.7: Decision Tree based on outlook

If we see the overcast outlook from the data set, then it always follow that Yes we can play cricket. So in the decision tree it is also going only to Yes. Now we will calculate for the branch sunny. At first we will reduce the data set to sunny outlook.

outlook	temp	humidity	windy	play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Sunny	Mild	Normal	True	Yes

Figure 2.8: Data based on Sunny

At first divide the data set based on temp

Gini Index of hot temp is

$G(\text{temp=hot}|\text{sunny})=1-(1)^2=0$

$p(\text{temp=hot}|\text{sunny})=\frac{2}{5}$

Gini Index of mild temp is

$G(\text{temp=mild}|\text{sunny})=1-(\frac{1}{2})^2-(\frac{1}{2})^2=\frac{1}{2}$

$p(\text{temp=mild}|\text{sunny})=\frac{2}{5}$

Gini Index of cool temp is

$G(\text{temp=cool}|\text{sunny})=1-(1)^2=0$

$p(\text{temp=cool}|\text{sunny})=\frac{1}{5}$

Average Gini$=\frac{2}{5}\times\frac{1}{2}=0.2$

Now divide the data set based on humidity

Gini Index of high humidity is

$G(\text{humidity=high}|\text{sunny})=1-(1)^2=0$

p(humidity=high|sunny)=$\frac{3}{5}$

Gini Index of normal humidity is

G(humidity=normal|sunny)=$1-(1)^2=0$

p(humidity=normal|sunny)=$\frac{1}{2}$

Average Gini=0

Since 0 is the possible minimum value of gini Index, so there is no need to calculate for windy.

So Average Gini Index is minimum for the attribute humidity.

Therefore humidity will be the internal node of the decision tree.

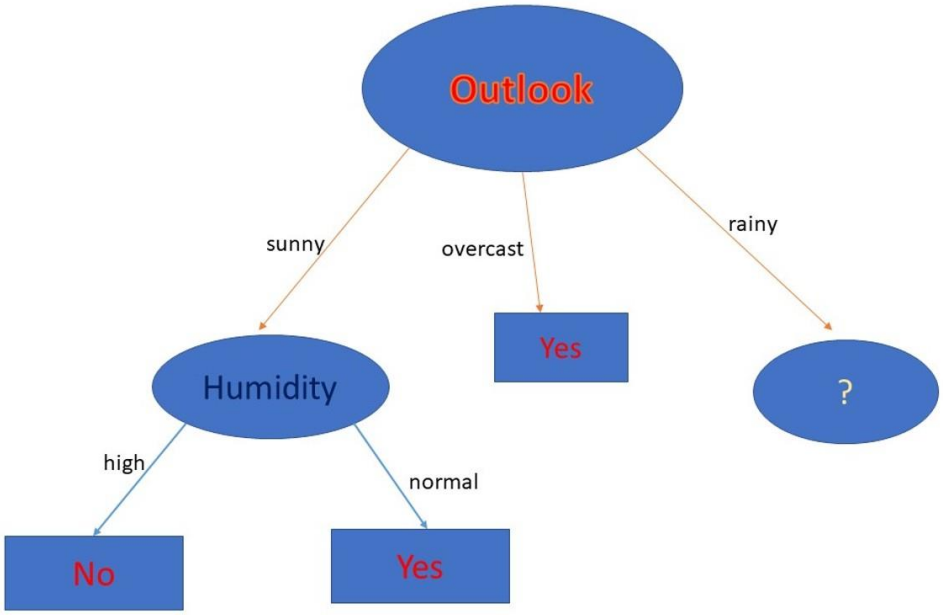

Figure 2.9: Decision Tree based on Sunny

Now we will calculate for the branch rainy. At first we will reduce the data set to rainy outlook.

outlook	temp	humidity	windy	play
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Rainy	Mild	Normal	False	Yes
Rainy	Mild	High	True	No

Figure 2.10: Data based on Rainy

At first divide the data set based on temp

Gini Index of mild temp is

$G(temp=mild|sunny)=1-(\frac{1}{3})^2-(\frac{2}{3})^2=\frac{4}{9}$

$p(temp=mild|rainy)=\frac{3}{5}$

Gini Index of cool temp is

$G(temp=cool|rainy)=1-(\frac{1}{2})^2-(\frac{1}{2})^2=\frac{1}{2}$

$p(temp=cool|rainy)=\frac{2}{5}$

$Average\ Gini=\frac{3}{5}\times\frac{4}{9}+\frac{2}{5}\times\frac{1}{2}=\frac{21}{45}$

Now divide the data set based on humidity

Gini Index of high humidity is

$G(humidity=high|rainy)=1-(\frac{1}{2})^2-(\frac{1}{2})^2=\frac{1}{2}$

$p(humidity=high|rainy)=\frac{2}{5}$

Gini Index of normal humidity is

$G(\text{humidity=normal}|\text{rainy})=1-(\frac{1}{3})^2-(\frac{2}{3})^2=\frac{4}{9}$

$p(\text{humidity=normal}|\text{sunny})=\frac{3}{5}$

$\text{Average Gini}=\frac{2}{5}\times\frac{1}{2}+\frac{3}{5}\times\frac{4}{9}=\frac{21}{45}$

Now divide the data set based on windy

Gini Index of true windy is

$G(\text{windy=true}|\text{rainy})=1-(1)^2=0$

$p(\text{windy=true}|\text{rainy})=\frac{2}{5}$

Gini Index of false windy is

$G(\text{windy=false}|\text{rainy})=1-(1)^2=0$

$p(\text{windy=false}|\text{rainy})=\frac{3}{5}$

Average Gini=0

So Average Gini Index is minimum for the attribute windy.

Therefore windy will be the internal node of the decision tree.

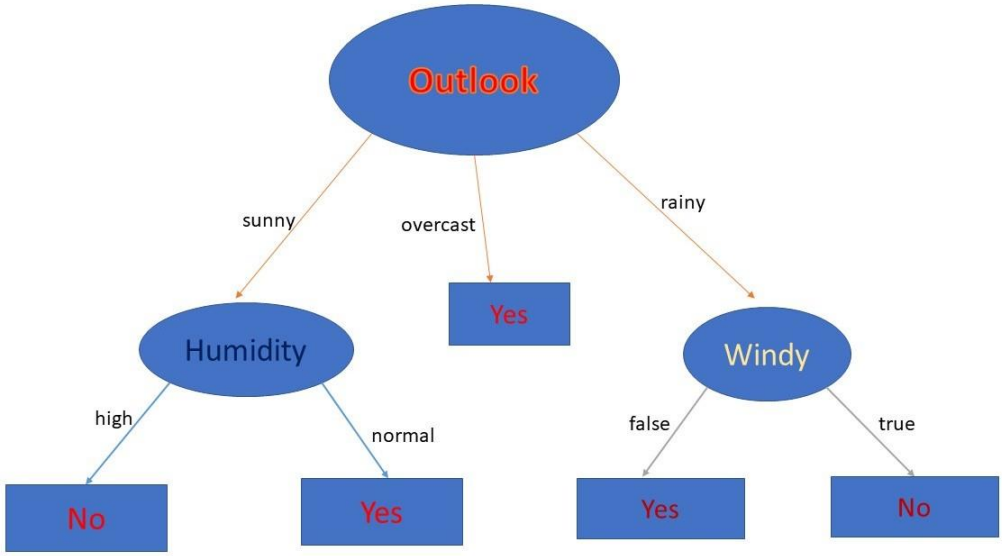

Figure 2.11: Decision Tree based on Whole Data

So now we have the decision tree based on the data. Now we can predict the status of playing cricket depending on some new weather data

outlook	temp	humidity	windy	play
Rainy	Cool	Normal	False	?

Figure 2.12: New Data

If we use this new data and implement it in our decision tree, then the predicted status of playing cricket will be Yes.

Drawbacks

But decision tree has some drawbacks which are as follows -

1. A small change in the data can make a large change in the structure of the decision tree, then the decision tree can be unstable
2. For decision tree sometimes calculations can be more complex compared to others
3. Decision Tree is not so good for applying regression
4. Sometimes it may have overfitting issue

Chapter 3

Implement in Python

In this chapter we will implement our decision tree algorithms in python and also visualize the decision tree

3.1 Python code for Decision Tree

At first we will import some python in-build libraries- numpy and pandas.

```
#importing libraries
import numpy as np
import pandas as pd
df = pd.read_csv('data.csv')
```

Now we will make some duplicate columns to get the numerical equivalents of the columns

```
#Get dummy columns for equivalent numerical values
df_getdummy=pd.get_dummies(data=df, columns=['temp', 'outlook', 'windy','humidity'])
```

Now we will import train-test function from sklearn to split into training

and testing data set. In X we take only the feature variables and in y we take the response variable.

```
#import train_test fuction from sklearn library
from sklearn.model_selection import train_test_split
X = df_getdummy.drop('play',axis=1)
y = df_getdummy['play']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
```

Now we will import decision tree classifier and fit our training data wih this model.

```
#importing decision tree classifier from sklearn and fit the model
from sklearn.tree import DecisionTreeClassifier
dtree = DecisionTreeClassifier(max_depth=3)
dtree.fit(X_train,y_train)
predictions = dtree.predict(X_test)
```

Now we will import plot-tree to visualize our Decision Tree

Code for Entropy

Codes to get the decision tree based on entropy are given below

```
#visualize the decision tree
from sklearn.tree import DecisionTreeClassifier
dtree = DecisionTreeClassifier(criterion='entropy',max_depth=3)
dtree.fit(X_train,y_train)
predictions = dtree.predict(X_test)
from sklearn.tree import plot_tree
import matplotlib.pyplot as plt
fig = plt.figure(figsize=(16,12))
a = plot_tree(dtree, feature_names=df_getdummy.columns, fontsize=12, filled=True,
class_names=['Not play', 'play'])
```

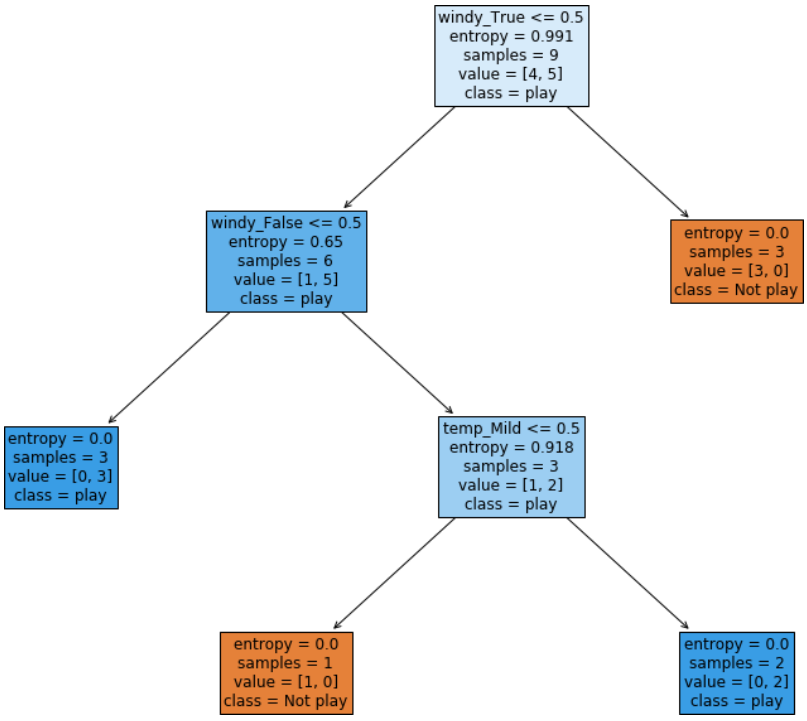

Figure 3.1: Decision Tree based on Entropy

Code for Gini Index

Codes to get the decision tree based on gini index are given below

```
from sklearn import tree
clf = tree.DecisionTreeClassifier(criterion = 'gini')
clf = clf.fit(X_train, y_train)
import graphviz
dot_data = tree.export_graphviz(clf, out_file=None)
graph = graphviz.Source(dot_data)
graph
```

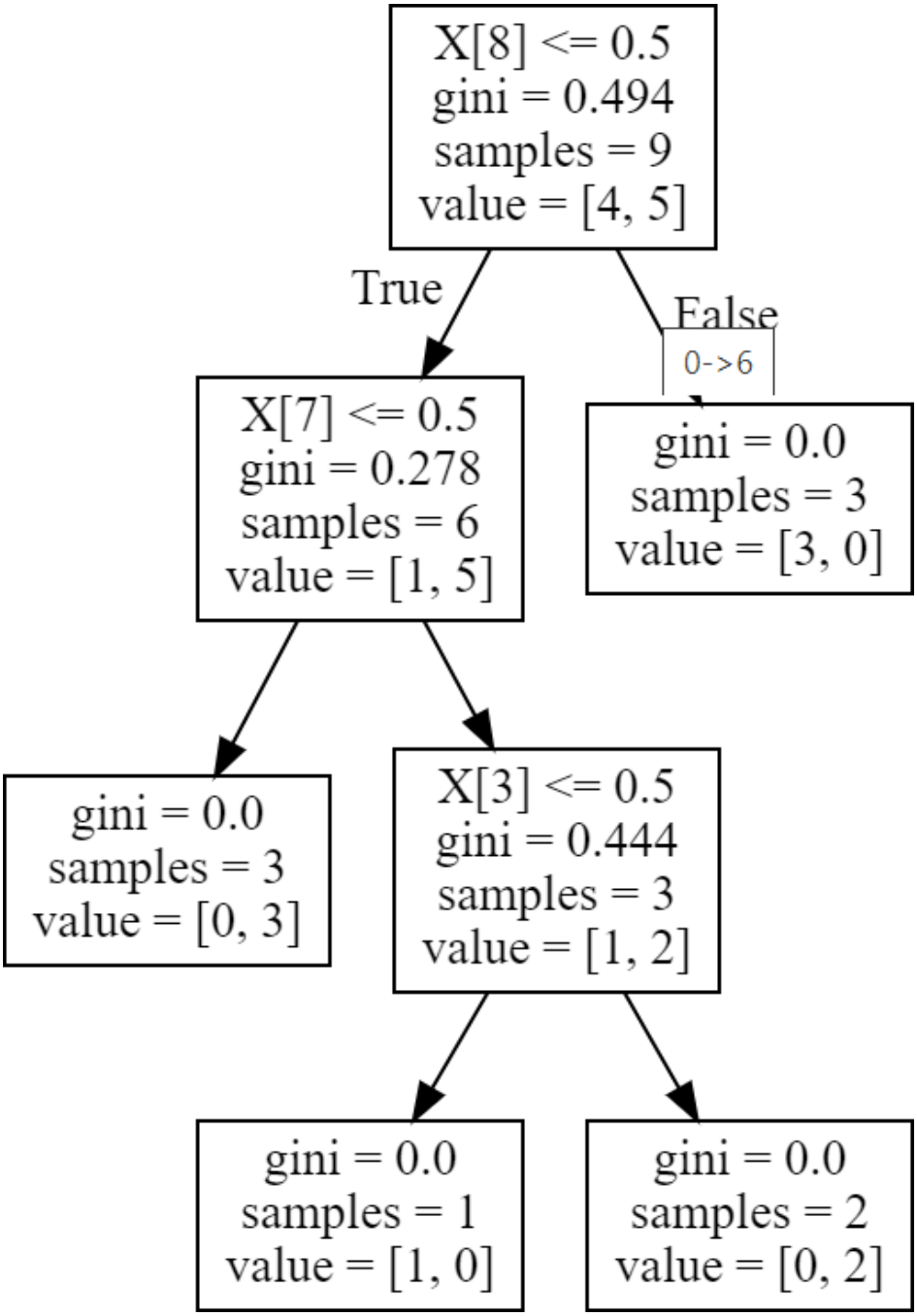

Figure 3.2: Decision Tree based on Gini Index

Bibliography

[1] Machine Learning, Tom Mitchell McGraw Hill, 1997

[2] Data Classification Algorithms and Applications, C.C.Aggarwal CRC Press