Homework 1 ID3 Decision Tree

Matthew Jensen

2024-09-14

Part 1

Problem 1.a

Split 1

Step 1.i Calculate the Entropy of the Target variable

$$\text{Entropy} = H(S) = -\sum_{i}^{n} p_i \cdot \log_2(p_i) \text{ where p is the proportion of class } i$$

$$\text{In the dataset, the proportion of labels when } y = 1 \text{ is } \frac{5}{7} \text{ and the proportion of lables when } y = 0 \text{ is } \frac{2}{7}$$

$$\text{Thus, Entropy } = -(\frac{5}{7} \cdot \log_2(\frac{5}{7}) + \frac{2}{7} \cdot \log_2(\frac{2}{7})) = 0.8632$$

Step 1.ii Information Gain for Each Attribute

The formula for Information Gain is:

$$\text{Information Gain } = IG(S, A) = H(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

$\boldsymbol{x_1}$:

For the given proportions and entropy calculations:

When $x_1 = 0$, the proportions are: - Proportion of $y = 1$ is $\frac{1}{5}$ - Proportion of $y = 0$ is $\frac{4}{5}$

Entropy for $x_1 = 0$ is:

$$H(Y_{x_1=0}) = -\left(\frac{4}{5}\log_2\left(\frac{4}{5}\right) + \frac{1}{5}\log_2\left(\frac{1}{5}\right)\right) \approx 0.7219$$

When $x_1 = 1$, the proportions are: * Proportion of $y = 1$ is $\frac{1}{2}$ * Proportion of $y = 0$ is $\frac{1}{2}$

Entropy for $x_1 = 1$ is:

$$H(Y_{x_1=1}) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$$

Loading [MathJax]/jax/output/HTML-CSS/jax.js

Information Gain for x_1 is:

$$IG(Y, x_1) = H(Y) - \left(\frac{4}{5}H(Y_{x_1=0}) + \frac{1}{5}H(Y_{x_1=1})\right)$$

$$IG(Y, x_1) = 0.8632 - \left(\frac{4}{5} \cdot 0.7219 + \frac{1}{5} \cdot 1\right) \approx 0.062$$

x_2 :

When $x_2 = 0$, the proportions are: - Proportion of $y = 1$ is $\frac{2}{3}$ - Proportion of $y = 0$ is $\frac{1}{3}$

Entropy for $x_2 = 0$ is:

$$H(Y_{x_2=0}) = -\left(\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) \approx .9183$$

When $x_2 = 1$, the proportions are: * Proportion of $y = 1$ is $\frac{4}{4}$ * Proportion of $y = 0$ is $\frac{0}{4}$

Entropy for $x_2 = 1$ is:

$$H(Y_{x_2=1}) = -\left(\frac{4}{4}\log_2\left(\frac{4}{4}\right) + \frac{0}{4}\log_2\left(\frac{0}{4}\right)\right) = 0$$

Information Gain for x_2 is:

$$IG(Y, x_2) = H(Y) - \left(\frac{3}{7}H(Y_{x_2=0}) + \frac{4}{7}H(Y_{x_2=1})\right)$$

$$IG(x_2) = 0.862 - 0.393 = 0.469$$

x_3 :

When $x_3 = 0$, the proportions are: - Proportion of $y = 1$ is $\frac{1}{4}$ - Proportion of $y = 0$ is $\frac{3}{4}$

Entropy for $x_3 = 0$ is:

$$H(Y_{x_3=0}) = -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right) \approx .8113$$

When $x_3 = 0$, the proportions are: - Proportion of $y = 1$ is $\frac{1}{4}$ - Proportion of $y = 0$ is $\frac{3}{4}$

Entropy for $x_3 = 0$ is:

$$H(Y_{x_3=0}) = -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right) \approx 0.811$$

When $x_3 = 1$, the proportions are: * Proportion of $y = 1$ is $\frac{1}{3}$ * Proportion of $y = 0$ is $\frac{2}{3}$

Entropy for $x_3 = 1$ is:

$$H(Y_{x_3=1}) = -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) \approx 0.918$$

Information Gain for x_3 is:

$$IG(Y, x_3) = H(Y) - \left(\frac{4}{7}H(Y_{x_3=0}) + \frac{3}{7}H(Y_{x_3=1})\right)$$

$$IG(x_3) = 0.862 - 0.856 = 0.006$$

x_4 :

When $x_4 = 0$, the proportions are: - Proportion of $y = 1$ is $\frac{0}{4}$ - Proportion of $y = 0$ is $\frac{4}{4}$

Entropy for $x_4 = 0$ is:

$$H(Y_{x_4=0}) = -\left(\frac{0}{4}\log_2\left(\frac{0}{4}\right) + \frac{4}{4}\log_2\left(\frac{4}{4}\right)\right) = 0$$

When $x_4 = 1$, the proportions are: * Proportion of $y = 1$ is $\frac{2}{3}$ * Proportion of $y = 0$ is $\frac{1}{3}$

Entropy for $x_4 = 1$ is:

$$H(Y_{x_4=1}) = -\left(\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) \approx 0.918$$

Information Gain for x_4 is:

$$IG(Y, x_4) = H(Y) - \left(\frac{4}{7}H(Y_{x_4=0}) + \frac{3}{7}H(Y_{x_4=1})\right)$$

$$IG(x_4) = 0.862 - 0.393 = 0.469$$

Split 2.1 - Left Side of Tree

Step 2.i Entropy Calculation

We know will split the dataset on x_2

Dataset when x_2 = 0

Let's start with the table x_2 = 0

Loading [MathJax]/jax/output/HTML-CSS/jax.js

x_1	x_2	x_3	x_4	y
0	0	1	0	0
0	0	1	1	1
1	0	0	1	1

When $x_2 = 0$, the proportions are: * Proportion of $y = 1$ is $\frac{2}{3}$ * Proportion of $y = 0$ is $\frac{1}{3}$

Entropy for Y|$x_2 = 0$ =

$$-(\frac{2}{3} \cdot \log_2(\frac{2}{3}) + \frac{1}{3} \cdot \log_2(\frac{1}{3})) = .9183$$

Step 2.ii Information Gain for Each Attribute

$\boldsymbol{x_1 | x_2 = 0}$:

When $x_1 = 0 | x_2 = 0$, the proportions are: - Proportion of $y = 1$ is $\frac{1}{2}$ - Proportion of $y = 0$ is $\frac{1}{2}$

Entropy for $x_1 = 0 | x_2 = 0$ is:

$$H(Y_{x_1=0|x_2=1}) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$$

When $x_1 = 1 | x_2 = 0$, the proportions are: * Proportion of $y = 1$ is $\frac{1}{1}$ * Proportion of $y = 0$ is $\frac{0}{1}$

Entropy for $x_1 = 1 | x_2 = 0$ is:

$$H(Y_{x_1=1|x_2=0}) = -\left(\frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{1}\log_2\left(\frac{1}{1}\right)\right) = 0$$

Information Gain for $x_1 | x_2 = 0$ is:

$$IG(Y, x_1 | x_2 = 0) = H(Y | x_2 = 0) - \left(\frac{2}{3}H(Y_{x_1=0|x2=0}) + \frac{1}{3}H(Y_{x_1=1|x_2=0})\right)$$

$$IG(Y, x_1 | x_2 = 0) = .9183 - \left(\frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 0\right) \approx 0.252$$

$\boldsymbol{x_3 | x_2 = 0}$:

For the given proportions and entropy calculations:

When $x_3 = 0 | x_2 = 0$, the proportions are: - Proportion of $y = 1$ is $\frac{1}{1}$ - Proportion of $y = 0$ is $\frac{0}{1}$

Entropy for $x_3 = 0 | x_2 = 0$ is:

Loading [MathJax]/jax/output/HTML-CSS/jax.js

$$H(Y_{x_3=0\,|\,x_2=0}) = -\left(\frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) = 0$$

When $x_3 = 1\,|\,x_2 = 0$, the proportions are: - Proportion of $y = 1$ is $\frac{1}{1}$ - Proportion of $y = 0$ is $\frac{0}{1}$

Entropy for $x_3 = 1\,|\,x_2 = 0$ is:

$$H(Y_{x_3=1\,|\,x_2=0}) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$$

Information Gain for $x_1\,|\,x_2 = 0$ is:

$$IG(Y, x_3\,|\,x_2 = 0) = H(Y|\,x_2 = 0) - \left(\frac{2}{3}H(Y_{x_3=0\,|\,x2=0}) + \frac{1}{3}H(Y_{x_3=1\,|\,x_2=0})\right)$$

$$IG(Y, x_3\,|\,x_2 = 0) = .9183 - \left(\frac{2}{3}\cdot 1 + \frac{1}{3}\cdot 0\right) \approx 0.252$$

(x_4 | x2= 0) :

For the given proportions and entropy calculations:

When $x_4 = 0\,|\,x_2 = 0$, the proportions are: - Proportion of $y = 1$ is $\frac{0}{1}$ - Proportion of $y = 0$ is $\frac{1}{1}$

Entropy for $x_4 = 0\,|\,x_2 = 0$ is:

0

When $x_4 = 1\,|\,x_2 = 0$, the proportions are: - Proportion of $y = 1$ is $\frac{1}{1}$ - Proportion of $y = 0$ is $\frac{0}{1}$

Entropy for $x_4 = 1\,|\,x_2 = 0$ is:

0

Information Gain for (x_4 | x_2 =0) is :

.918 - 0 = .918

Thus x_4 has the highest Information Gain and will be the next split. Because the node is pure, we can make leaf nodes where if x_4 = 0 then y = 0 and if x_4 = 1 then y=1}

Split 2.2 Right side of the tree.

We will now go back and look at the case when x_2 = 1.

ataset when x_2 = 1

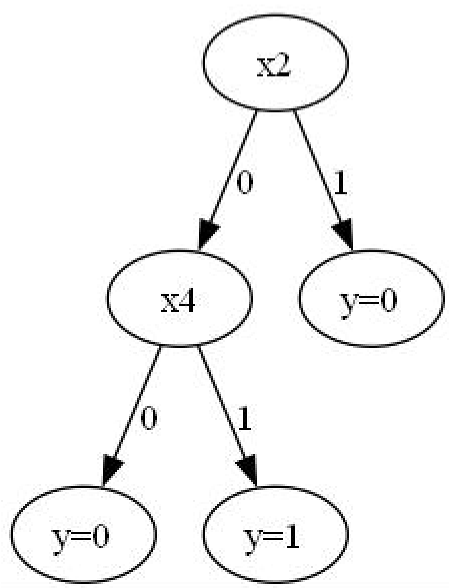
x_1	x_2	x_3	x_4	y
0	1	0	0	0
0	1	1	0	0
1	1	0	0	0
0	1	0	1	0

When $x_2 = 1$, the proportions are: * Proportion of $y = 1$ is $\frac{0}{4}$ * Proportion of $y = 0$ is $\frac{4}{4}$

Entropy for Y|$x_2 = 0$ =

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

The Entropy is 0, meaning that this is a pure node and if x_2 =1 then y = 1 and if x_2 =0 then y = 0



Picture of Final Decision Tree

Problem 1.b Boolean Function Table

x_1	x_2	x_3	x_4	y
$-$	0	$-$	0	0
$-$	0	$-$	1	1
$-$	1	$-$	$-$	0

If $x_2 = 0$ and $x_4 = 0$ then $y = 0$, If $x_2 = 0$ and $x_4 = 1$ then $y = 1$, If $x_2 = 1$ then $y = 0$

Problem 2

Loading [MathJax]/jax/output/HTML-CSS/jax.js

2.a Majority Error on Play Tennis Dataset

Outlook	Temperature	Humidity	Wind	PlayTennis
S	H	H	W	−
S	H	H	S	−
O	H	H	W	+
R	M	H	W	+
R	C	N	W	+
R	C	N	S	−
O	C	N	S	+
S	M	H	W	−
S	C	N	W	+
R	M	N	W	+
S	M	N	S	+
O	M	H	S	+
O	H	N	W	+
R	M	H	S	−

Step i. Calculate the Majority Error of Data Set

$$ME(S) = 1 - \max(p_+, p_-)$$

In dataset, the proportion of positives is $\frac{9}{14}$ and the proporiton of negatives is $\frac{5}{14}$

Thus the majority error is 1 - $\frac{9}{14}$ = $\frac{5}{14}$ = .3571

Step ii. Calculate the ME for each attribute

Outlook

When Outlook = Sunny, PlayTennis = [-,-,-,+,+] ME(Outlook = Sunny) = 1 - $\frac{3}{5}$ = $\frac{2}{5}$

When Outlook = Overcast, PlayTennis = [+,+,+,+] ME(Outlook = Overcast) = 1 - $\frac{4}{4}$ = 0

When Outlook = Rainy, PlayTennis = [+,+,-,+,-] ME(Outlook = Rainy) = 1 - $\frac{3}{5}$ = $\frac{2}{5}$

$$IG(\text{Outlook}) = \frac{5}{14} - \left(\left(\frac{5}{14} \cdot \frac{2}{5}\right) + \left(\frac{4}{14} \cdot 0\right) + \left(\frac{5}{14} \cdot \frac{2}{5}\right)\right) = \frac{5}{14} - \frac{4}{14} = .0714$$

Temperature

When Temperature = Hot, Playtennis = [-,-,+,+] ME(Temperature = Hot) = 1 - $\frac{1}{2}$ = $\frac{1}{2}$

When Temperature = Medium, Playtennis = [+,-,+,+,+,-] ME(Temperature = Medium) = 1 - $\frac{4}{6}$ = $\frac{2}{6}$

When Temperature = Cool, Playtennis = [+,-,+,+] ME(Temperature = Cool) = 1 - $\frac{3}{4}$ = $\frac{1}{4}$

$$IG(\text{Temperature}) \ = \frac{5}{14} - ((\frac{4}{14} \cdot \frac{1}{2}) + (\frac{6}{14} \cdot \frac{2}{6}) + (\frac{4}{14} \cdot \frac{1}{4})) = \frac{5}{14} - \frac{5}{14} = 0$$

Humidity

When Humidity = High, Playtennis = [-,-,+,+,-,+,-] ME(Humidity = Hight) = 1 - $\frac{4}{7}$ = $\frac{3}{7}$

When Humidity = Normal, Playtennis = [+,-,+,+,+,+,+] ME(Humidity = Normal) = 1 - $\frac{6}{7}$ = $\frac{1}{7}$

$$IG(\text{Humidity}) \ = \frac{5}{14} - ((\frac{7}{14} \cdot \frac{3}{7}) + (\frac{7}{14} \cdot \frac{1}{7}) = \frac{5}{14} - \frac{4}{14} = 0.0714$$

Wind

When Wind = Strong, Playtennis = [-,-,+,+,+,-] ME(Wind = Strong) = 1 - $\frac{3}{6}$ = $\frac{3}{6}$

When Wind = Weak, Playtennis = [-,+,+,+,-,+,+,+] ME(Wind = Weak) = 1 - $\frac{6}{8}$ = $\frac{2}{8}$

$$IG(\text{Wind}) \ = \frac{5}{14} - ((\frac{6}{14} \cdot \frac{3}{6}) + (\frac{8}{14} \cdot \frac{2}{8}) = \frac{5}{14} - \frac{5}{14} = 0$$

Outlook and Humidity have the biggest Information Gain, so let’s split on Outlook

Branch 1. Outlook = Sunny

Outlook	Temperature	Humidity	Wind	PlayTennis
S	H	H	W	−
S	H	H	S	−
S	M	H	W	−
S	C	N	W	+
S	M	N	S	+

Given Outlook = Sunny, the Majority Error is 1 - $\frac{3}{5}$ = $\frac{2}{5}$

Humidity

Given Given Outlook = Sunny, When Humidity = High, Playtennis = [-,-,-] ME(H=H|O=S) = 1 - $\frac{3}{3}$ = 0

Given Given Outlook = Sunny, When Humidity = Normal, Playtennis = [+,+] ME(H=H|O=S) = 1 - $\frac{2}{2}$ = 0

$$IG(\text{Huminity}|\ \text{O} = \text{S}) \ = \frac{2}{5} - ((\frac{3}{3} \cdot 0) + (\frac{2}{2} \cdot 0) = \frac{2}{5} - 0 = \frac{2}{5}$$

Splitting on Humidity is a pure node so we can create leaf nodes here. When H=High Playtennis = -, and when H= Normal Playtennis = +

Loading [MathJax]/jax/output/HTML-CSS/jax.js

Branch 2. Outlook = Overcast

Outlook	Temperature	Humidity	Wind	PlayTennis
O	H	H	W	+
O	C	N	S	+
O	M	H	S	+
O	H	N	W	+

Given Outlook=Overcast , the Majority Error = 1 - $\frac{4}{4}$ = 0, meaning that this node is pure and Playtennis = +.

Branch 3. Outlook = Rainy

Outlook	Temperature	Humidity	Wind	PlayTennis
R	M	H	W	+
R	C	N	W	+
R	C	N	S	−
R	M	N	W	+
R	M	H	S	−

Give Outlook = Rainy, the Majority Error = 1 - $\frac{3}{5}$ = $\frac{2}{5}$

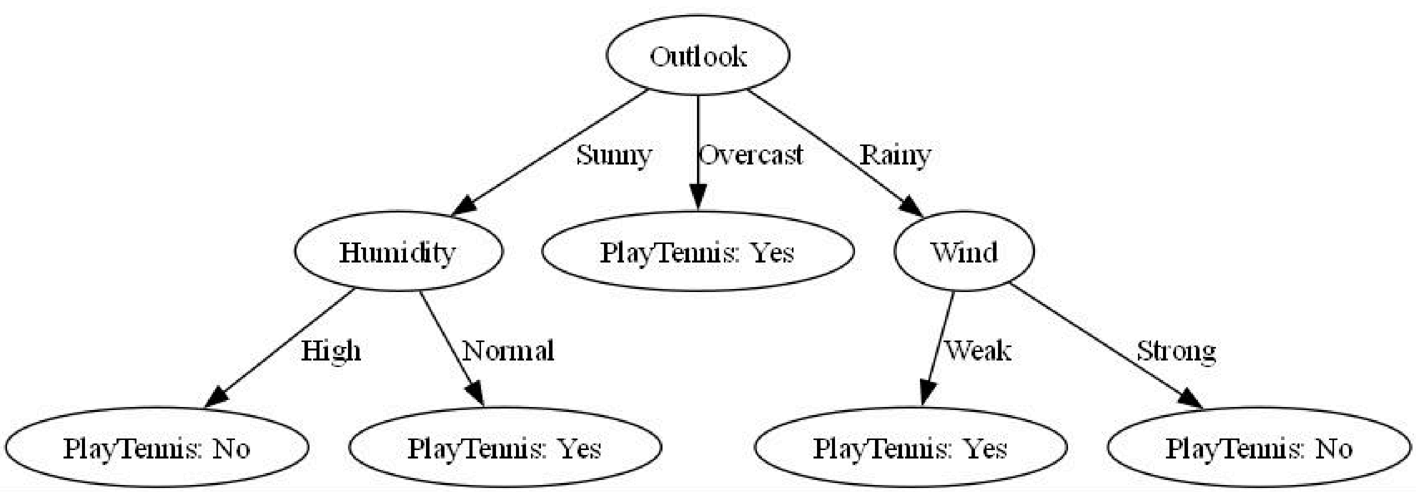
Wind

Given Outlook = Rainy , when Wind = Weak, Playtennis = [+,+,+] ME(W=W|O=R) = 1 - $\frac{3}{3}$ = 0

Given Outlook = Rainy , when Wind = Strong, Playtennis = [-,-] ME(W=W|O=R) = 1 - $\frac{2}{2}$ = 0

$$IG(\text{Wind}|\ \text{O} = \text{R}) \ = \frac{2}{5} - ((\frac{3}{3} \cdot 0) + (\frac{2}{2} \cdot 0) = \frac{2}{5} - 0 = \frac{2}{5}$$

Splitting on Wind is a pure node so we can create leaf nodes here. When Wind=Strong Playtennis = -, and when Wind = Weak Playtennis = +



Picture of PlayTennis Decision Tree Based On Majority Error

Loading [MathJax]/jax/output/HTML-CSS/jax.js

2b. Decision Tree using Gini Index

$$GI(Y) = 1 - (p_+^2 + p_-^2)$$

$$GI(Y) = 1 - ((\frac{9}{14})^2 + (\frac{5}{14})^2) = 1 - \frac{106}{196} = \frac{90}{196} = .46$$

Step ii. GI for Each Attribute

Outlook

When Outlook = Sunny, PlayTennis = [-,-,-,+,+] GI(Outlook = Sunny) = 1 - $(\frac{3}{5})^2$ + $(\frac{2}{5})^2$ = 1 - $\frac{13}{25}$ = $(\frac{12}{25})$

When Outlook = Overcast, PlayTennis = [+,+,+,+] GI(Outlook = Overcast) = 1 - $(\frac{4}{4})^2$ + $(\frac{0}{0})^2$ = 1 - 1 = 0

When Outlook = Rainy, PlayTennis = [+,+,-,+,-] GI(Outlook = Rainy) = 1 - $(\frac{3}{5})^2$ + $(\frac{2}{5})^2$ = 1 - $\frac{13}{25}$ = $(\frac{12}{25})$

$$GI(Outlook) = \frac{5}{14} \cdot \frac{12}{25} + \frac{4}{14} \cdot 0 + \frac{5}{14} \cdot \frac{12}{25} = \frac{24}{70} = .342$$

Temperature

When Temperature = Hot, Playtennis = [-,-,+,+] GI(Temperature = Hot) = 1 - $(\frac{1}{2})^2$ + $(\frac{1}{2})^2$ = 1 - $(\frac{1}{2})$ = $(\frac{1}{2})$

When Temperature = Medium, Playtennis = [+,-,+,+,+,-] GI(Temperature = Medium) = 1 - $(\frac{4}{6})^2$ + $(\frac{2}{6})^2$ = 1 - $(\frac{20}{36})$ = $(\frac{16}{36})$

When Temperature = Cool, Playtennis = [+,-,+,+] GI(Temperature = Hot) = 1 - $(\frac{3}{4})^2$ + $(\frac{1}{4})^2$ = 1 - $(\frac{10}{16})$ = $(\frac{6}{16})$

$$GI(Temperature) = \frac{4}{14} \cdot \frac{1}{2} + \frac{6}{14} \cdot \frac{16}{36} + \frac{4}{14} \cdot \frac{6}{16} = .44$$

Humidity

When Humidity = High, Playtennis = [-,-,+,+,-,+,-] GI(Humidity = High) = 1 - $(\frac{3}{7})^2$ + $(\frac{4}{7})^2$ = 1 - $(\frac{25}{49})$ = $(\frac{24}{49})$

When Humidity = Normal, Playtennis = [+,-,+,+,+,+,+] GI(Humidity = Normal) = 1 - $(\frac{6}{7})^2$ + $(\frac{1}{7})^2$ = 1 - $(\frac{37}{49})$ = $(\frac{12}{49})$

$$GI(Humidity) = \frac{7}{14} \cdot \frac{24}{49} + \frac{7}{14} \cdot \frac{12}{49} = .368$$

Wind

When Wind = Strong, Playtennis = [-,-,+,+,+,-] GI(Wind = Strong) = 1 - $(\frac{1}{2})^2$ + $(\frac{1}{2})^2$ = 1 - $(\frac{1}{2})$ = $(\frac{1}{2})$

When Wind = Weak, Playtennis = [-,+,+,+,-,+,+,+] ME(Wind = Weak) = 1 - $(\frac{6}{8})^2$ + $(\frac{2}{8})^2$ = 1 - $(\frac{40}{64})$ = $(\frac{24}{64})$

$$GI(Wind) = \frac{6}{14} \cdot \frac{1}{2} + \frac{8}{14} \cdot \frac{24}{64} = .43$$

Outlook has the Smallest Gini Index, so we will split on Outlook

Branch 1. Outlook = Sunny

Humidity

Given Outlook = Sunny:

When Humidity = High PlayTennis = [-,-,-], GI(Humidity = High | O=S) = 1 - $(\frac{0}{0})^2$ + $(\frac{3}{3})^2$ = 1 - 1 = 0

When Humidity = Normal PlayTennis = [+,+], GI(Humidity = High | O=S) = 1 - $(\frac{2}{2})^2$ + $(\frac{0}{0})^2$ = 1 - 1 = 0

$$Gini(\text{Humidity} \mid \text{Outlook} = \text{Sunny}) = \frac{3}{5} \times 0 + \frac{2}{5} \times 0 = 0$$

$$IG(\text{Humidity} \mid \text{Outlook} = \text{Sunny}) = 0.48 - 0 = 0.48$$

Given Outlook = Sunny, the Gini Index for Humidity is 0, meaning that this is a pure node and we can create leaf nodes. When Humidity = High, Playtennis = +, When Humidity = Normal, Playtennis = -

Branch 2. Outlook = Overcast

Give Outlook = Rainy, the Gini Index is 1 - 1 - $(\frac{4}{4})^2$ + $(\frac{0}{0})^2$ = 1 - 1 = 0

Thus this node is pure and can we can create a leaf node. Given Outlook = Overcast, Playtennis = +

Branch 3. Outlook = Rainy

Give Outlook = Rainy, the Majority Error = 1 - $\frac{3}{5}$ = $\frac{2}{5}$

Wind

Given Outlook = Rainy , when Wind = Weak, Playtennis = [+,+,+] ME(W=W|O=R) = 1 - $\frac{3}{3}$ = 0

Given Outlook = Rainy , when Wind = Strong, Playtennis = [-,-] ME(W=W|O=R) = 1 - $\frac{2}{2}$ = 0

$$IG(\text{Wind}|\text{ O} = \text{R}) \ = \frac{2}{5} - ((\frac{3}{3} \cdot 0) + (\frac{2}{2} \cdot 0) = \frac{2}{5} - 0 = \frac{2}{5}$$

Wind

Given Outlook = Rainy:

When Wind = Strong PlayTennis = [-,-], GI(Wind = Strong | O=R) = 1 - $(\frac{0}{2})^2 + (\frac{2}{2})^2$ = 1 - 1 = 0

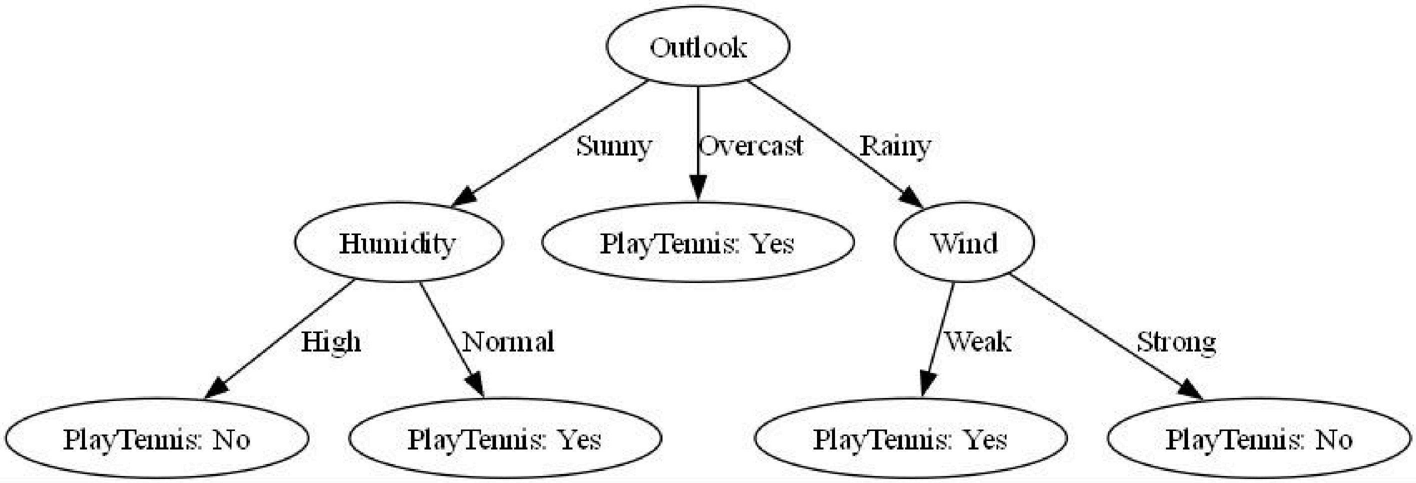
When Wind = Weak PlayTennis = [+,+,+], GI(Wind = Weak | O=R) = 1 - $(\frac{3}{3})^2 + (\frac{0}{3})^2$ = 1 - 1 = 0

$$Gini(\text{Wind} \mid \text{Outlook} = \text{Rainy}) = \frac{2}{5} \times 0 + \frac{3}{5} \times 0 = 0$$

$$IG(\text{Wind} \mid \text{Outlook} = \text{Rainy}) = 0.48 - 0 = 0.48$$

Thus this node is pure and can we can create a leaf node. Given Outlook = Rainy, When Windy = Weak Playtennis = +, and wehn Windy = Strong Playtennis = -

This tree will be the same as the Majority Error Tree.



PlayTennis Decision Tree Based On Gini Index

2C. Decision Tree Similarity

All 3 Decision Trees are the same. Some reasons include that we are training on the same subset of training data, ME and GI are both methods to mathematically choose the best feature to split on and got similar results, and I handled ties the same. For the ME tree, there was a tie between Outlook and Humidity, but I decided to split on Outlook rather than Humidity, but I could've made a different tree.

Problem 3

3.a Impute NA Value Based On Most Common Value

Most common Outlook Value is Sunny and Rain, I will impute the missing Value with Sunny.

New Dataset:

Outlook	Temperature	Humidity	Wind	PlayTennis
S	H	H	W	−
S	H	H	S	−
O	H	H	W	+
R	M	H	W	+
R	C	N	W	+
R	C	N	S	−
O	C	N	S	+
S	M	H	W	−
S	C	N	W	+
R	M	N	W	+
S	M	N	S	+
O	M	H	S	+
O	H	N	W	+
R	M	H	S	−
S	**M**	**N**	**W**	+

Loading [MathJax]/jax/output/HTML-CSS/jax.js

$$\text{Entropy(Y)} \;=\; -(\frac{10}{15}\log_2(\frac{10}{15})+\frac{5}{15}\log_2(\frac{5}{15})) = .918$$

3a.ii. Attribute Entropy Values and Information Gain

Outlook

$$\text{Entropy(Outlook = Sunny)} \;=\; -(\frac{1}{2}\log_2(\frac{1}{2})+\frac{1}{2}\log_2(\frac{1}{2})) = 1$$

$$\text{Entropy(Outlook = Overcast)} \;=\; -(\frac{4}{4}\log_2(\frac{4}{4})+\frac{0}{0}\log_2(\frac{0}{0})) = 0$$

$$\text{Entropy(Outlook = Rainy)} \;=\; -(\frac{3}{5}\log_2(\frac{3}{5})+\frac{2}{5}\log_2(\frac{2}{5})) = .971$$

$$\text{Entropy(Outlook)} = \frac{6}{15}\cdot 1+\frac{4}{15}\cdot 0+\frac{5}{15}\cdot .971 \approx .723$$

$$\textbf{IG(Outlook)} \;=\; .918-.723 = \mathbf{.195}$$

Temperature

$$\text{Entropy(Temperature = Hot)} \;=\; -(\frac{1}{2}\log_2(\frac{1}{2})+\frac{1}{2}\log_2(\frac{1}{2})) = 1$$

$$\text{Entropy(Temperature = Mild)} \;=\; -(\frac{5}{7}\log_2(\frac{5}{7})+\frac{2}{7}\log_2(\frac{2}{7})) = .863$$

$$\text{Entropy(Temperature = Cold)} \;=\; -(\frac{3}{4}\log_2(\frac{3}{4})+\frac{1}{4}\log_2(\frac{1}{4})) = .811$$

$$\text{Entropy(Temperature)} = \frac{4}{15}\cdot 1+\frac{7}{15}\cdot .863+\frac{4}{15}\cdot .811 \approx .901$$

$$\textbf{IG(Temperature)} \;=\; .918-.901 = \mathbf{.017}$$

Humidity

$$\text{Entropy(Humidity = High)} \;=\; -(\frac{3}{7}\log_2(\frac{3}{7})+\frac{4}{7}\log_2(\frac{4}{7})) = .985$$

$$\text{Entropy(Humidity = Normal)} \;=\; -(\frac{7}{8}\log_2(\frac{7}{8})+\frac{1}{8}\log_2(\frac{1}{8})) = .543$$

$$\text{Entropy(Humidity)} = \frac{7}{15}\cdot .985+\frac{8}{15}\cdot .543 \approx .749$$

$$\textbf{IG(Humidity)} \;=\; .918-.749 = \mathbf{.169}$$

Wind

$$\text{Entropy(Wind = Strong)} \;=\; -(\frac{1}{2}\log_2(\frac{1}{2})+\frac{1}{2}\log_2(\frac{1}{2})) = 1$$

$$\text{Entropy(Wind = Weak)} \;=\; -(\frac{7}{9}\log_2(\frac{7}{9}) + \frac{2}{9}\log_2(\frac{2}{9})) = .764$$

$$\text{Entropy(Wind)} = \frac{6}{15} \cdot 1 + \frac{9}{15} \cdot .764 \approx .858$$

$$\textbf{IG(Wind)} \;=\; .918 - .858 = \mathbf{.06}$$

Based on Entropy, the best attribute to split on is Outlook

3b. Impute NA Value Based On Most Common Among Label

Among Positive labels, the most common value for Outlook is Overcast

Outlook	Temperature	Humidity	Wind	PlayTennis
S	H	H	W	−
S	H	H	S	−
O	H	H	W	+
R	M	H	W	+
R	C	N	W	+
R	C	N	S	−
O	C	N	S	+
S	M	H	W	−
S	C	N	W	+
R	M	N	W	+
S	M	N	S	+
O	M	H	S	+
O	H	N	W	+
R	M	H	S	−
O	**M**	**N**	**W**	+

The Entropy for this dataset is the same as 3a, **.918**

3b.ii Attribute Entropy and Information Gain

Outlook

$$\text{Entropy(Outlook = Sunny)} \;=\; -(\frac{2}{5}\log_2(\frac{2}{5}) + \frac{3}{5}\log_2(\frac{3}{5})) = .971$$

$$\text{Entropy(Outlook = Overcast)} \;=\; -(\frac{5}{5}\log_2(\frac{5}{5}) + \frac{0}{0}\log_2(\frac{0}{0})) = 0$$

$$\text{Entropy(Outlook = Rainy)} \;=\; -(\frac{3}{5}\log_2(\frac{3}{5}) + \frac{2}{5}\log_2(\frac{2}{5})) = .971$$

$$\text{Entropy(Outlook)} = \frac{5}{15} \cdot .971 + \frac{5}{15} \cdot 0 + \frac{5}{15} \cdot .971 \approx .647$$

Loading [MathJax]/jax/output/HTML-CSS/jax.js

$$\textbf{IG(Outlook)} \;=\; .918 - .647 = \mathbf{.271}$$

Temperature

Temperature Entropy and Information Gain will be the same here as in 3a. **IG = .017**

Humidity

Humidity Entropy and Information Gain will be the same here as in 3a. **IG = .169**

Wind

Wind Entropy and Information Gain will be the same here as in 3a. **IG = .06**

Thus, the highest Information Gain is still Outlook, which actually increased from 3a

3C. Impute NA Value Based On Fractional Counts

Outlook	Temperature	Humidity	Wind	PlayTennis
S	H	H	W	−
S	H	H	S	−
O	H	H	W	+
R	M	H	W	+
R	C	N	W	+
R	C	N	S	−
O	C	N	S	+
S	M	H	W	−
S	C	N	W	+
R	M	N	W	+
S	M	N	S	+
O	M	H	S	+
O	H	N	W	+
R	M	H	S	−
NA	**M**	**N**	**W**	**+**

3c.i Entropy

$$Entropy(Y) = -(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}) = \mathbf{.94}$$

3c.ii Attribute Entropy

Outlook

Because we know that the label is positive, we add to the positive label count, and the total size increase as well.

Outlook = Sunny

Positive Label = 2 + $\frac{5}{14}$ = $\frac{33}{14}$ Negative Label = 3 Size = 5 + $\frac{5}{14}$ = $\frac{75}{14}$

$$\text{Entropy(Sunny)} = -(\frac{\frac{33}{14}}{\frac{75}{14}} \cdot \log_2(\frac{\frac{33}{14}}{\frac{75}{14}}) + \frac{3}{\frac{75}{14}} \cdot \log_2(\frac{3}{\frac{75}{14}})) = .989$$

Loading [MathJax]/jax/output/HTML-CSS/jax.js

Outlook = Overcast

Positive Label = 4 + $\frac{4}{14}$ = $\frac{60}{14}$ Negative Label = 0 Size = 4 + $\frac{4}{14}$ = $\frac{60}{14}$

$$\text{Entropy(Overcast)} \ = 0$$

Outlook = Rainy

Positive Label = 3 + $\frac{5}{14}$ = $\frac{47}{14}$ Negative Label = 2 Size = 5 + $\frac{5}{14}$ = $\frac{75}{14}$

$$\text{Entropy(Rainy)} \ = \ -(\frac{\frac{47}{14}}{\frac{75}{14}} \cdot \log_2(\frac{\frac{47}{14}}{\frac{75}{14}}) + \frac{2}{\frac{75}{14}} \cdot \log_2(\frac{2}{\frac{75}{14}})) = .953$$

$$\text{Entropy(Outlook)} \ = \ \frac{5}{14} \cdot .989 + \frac{5}{14} \cdot .953 = \mathbf{.694}$$

$$IG(Outlook) = .94 - .694 = \mathbf{.246}$$

Temperature

Temperature = Hot

Positive Label = 2 + $\frac{4}{14}$ = $\frac{32}{14}$ Negative Label = 2 Size = 4 + $\frac{4}{14}$ = $\frac{60}{14}$

$$\text{Entropy(Hot)} \ = \ -(\frac{\frac{32}{14}}{\frac{60}{14}} \cdot \log_2(\frac{\frac{32}{14}}{\frac{60}{14}}) + \frac{2}{\frac{60}{14}} \cdot \log_2(\frac{2}{\frac{60}{14}})) = .997$$

Temperature = Medium

Positive Label = 4 + $\frac{6}{14}$ = $\frac{62}{14}$ Negative Label = 2 Size = 6 + $\frac{6}{14}$ = $\frac{90}{14}$

$$\text{Entropy(Medium)} \ = \ -(\frac{\frac{62}{14}}{\frac{90}{14}} \cdot \log_2(\frac{\frac{62}{14}}{\frac{90}{14}}) + \frac{2}{\frac{90}{14}} \cdot \log_2(\frac{2}{\frac{90}{14}})) = .894$$

Temperature = Cold

Positive Label = 3 + $\frac{4}{14}$ = $\frac{46}{14}$ Negative Label = 1 Size = 6 + $\frac{4}{14}$ = $\frac{84}{14}$

$$\text{Entropy(Medium)} \ = \ -(\frac{\frac{46}{14}}{\frac{84}{14}} \cdot \log_2(\frac{\frac{46}{14}}{\frac{84}{14}}) + \frac{1}{\frac{84}{14}} \cdot \log_2(\frac{1}{\frac{84}{14}})) = .907$$

$$\text{Entropy(Temperature)} \ = \ \frac{4}{14} \cdot .997 + \frac{6}{14} \cdot .894 + \frac{4}{14} \cdot .907 = \mathbf{.927}$$

$$IG(Temperature) = .94 - .927 = \mathbf{.013}$$

Humidity

Loading [MathJax]/jax/output/HTML-CSS/jax.js

Humidity = High

Positive Label = 3 + $\frac{7}{14}$ = $\frac{49}{14}$ Negative Label = 4 Size = 7 + $\frac{7}{14}$ = $\frac{105}{14}$

$$\text{Entropy(High)} = -(\frac{\frac{49}{14}}{\frac{105}{14}} \cdot \log_2(\frac{\frac{49}{14}}{\frac{105}{14}}) + \frac{7}{\frac{105}{14}} \cdot \log_2(\frac{7}{\frac{105}{14}})) = .606$$

Humidity = Normal

Positive Label = 6 + $\frac{7}{14}$ = $\frac{91}{14}$ Negative Label = 1 Size = 7 + $\frac{7}{14}$ = $\frac{105}{14}$

$$\text{Entropy(Normal)} = -(\frac{\frac{91}{14}}{\frac{105}{14}} \cdot \log_2(\frac{\frac{91}{14}}{\frac{105}{14}}) + \frac{1}{\frac{105}{14}} \cdot \log_2(\frac{1}{\frac{105}{14}})) = .567$$

$$\text{Entropy(Humidity)} = \frac{7}{14} \cdot .606 + \frac{7}{14} \cdot .567 = \mathbf{.5865}$$

$$IG(Humidity) = .94 - .5865 = \mathbf{.3535}$$

Wind

Wind = Strong

Positive Label = 3 + $\frac{6}{14}$ = $\frac{48}{14}$ Negative Label = 3 Size = 6 + $\frac{6}{14}$ = $\frac{90}{14}$

$$\text{Entropy(Strong)} = -(\frac{\frac{48}{14}}{\frac{90}{14}} \cdot \log_2(\frac{\frac{48}{14}}{\frac{90}{14}}) + \frac{3}{\frac{90}{14}} \cdot \log_2(\frac{3}{\frac{90}{14}})) = .997$$

Wind = Weak

Positive Label = 6 + $\frac{8}{14}$ = $\frac{92}{14}$ Negative Label = 2 Size = 8 + $\frac{8}{14}$ = $\frac{120}{14}$

$$\text{Entropy(Strong)} = -(\frac{\frac{92}{14}}{\frac{120}{14}} \cdot \log_2(\frac{\frac{92}{14}}{\frac{120}{14}}) + \frac{2}{\frac{120}{14}} \cdot \log_2(\frac{2}{\frac{120}{14}})) = .784$$

$$\text{Entropy(Wind)} = \frac{6}{14} \cdot .997 + \frac{8}{14} \cdot .784 = \mathbf{.875}$$

$$IG(Wind) = .94 - .875 = \mathbf{.065}$$

Based on Fractional Counts, the best Attribute to split on is Humidity

4. Prove that Information Gain is Always Non-

Loading [MathJax]/jax/output/HTML-CSS/jax.js

Negative

Entropy

The entropy of a dataset S is defined as:

$$H(S) = -\sum_{i=1}^{k} p_i \log(p_i)$$

where p_i is the proportion of instances in class i in S, and k is the number of classes.

Information Gain

Information Gain for an attribute A is defined as:

$$IG(S, A) = H(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

where S_v is the subset where A takes value v, and $\frac{|S_v|}{|S|}$ is the proportion of elements in S_v.

Proof

Entropy H is concave, meaning the weighted average entropy of subsets is less than or equal to the entropy of the whole set:

$$H(S) \geq \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

Thus, Information Gain is:

$$IG(S, A) = H(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} H(S_v) \geq 0$$

Conclusion

Since splitting the data reduces or maintains entropy, Information Gain is always nonnegative.

5. Regression Tree Gain

Because this is a Regression problem, we want to minimize the variance in the tree's predictions. We can use the Information Gain formula but replace entropy with variance.

$$Var(Y) = \frac{1}{n} \sum_{i}^{n} (y_i - \hat{y})^2$$

After splitting, calculate the variance of the target variable within each split (subset of data) and weight it by the proportion of data points in that split:

Loading [MathJax]/jax/output/HTML-CSS/jax.js

$$Var(Attribute) = \sum_{i}^{k} \frac{N_k}{N} Var(Y_k)$$

K is the number of subsets after the split (for continuous variables, this could be binary splits, while for categorical variables, it could be more).

N_k is the number of data points in the k-th subset.

Var(Y_k) is the variance of the target values in the k-th subset.

$$IG(Y, k) = Var(Y) - \sum_{i}^{k} \frac{N_k}{N} Var(Y_k)$$

You would want to choose the Attribute that outputs the highest IG value to split on that attribute