CS 5350/6350: Machine Learining Spring 2020

Homework 1

Britton Gaul
u0915408

February 9, 2020

1 Decision Tree

1.

(a)

$x_1 : S_0 = 1,2,3,5,7; S_1 = 4,6; (p_0^+, p_0^-) = (.2,.8); (p_1^+, p_1^-) = (.5,.5); H(S_0) = .72; H(S_1) = 1; H(S) = .86$; information gain = .06

$x_2 : S_0 = 1,3,4; S_1 = 2,5,6,7; (p_0^+, p_0^-) = (2/3, 1/3); (p_1^+, p_1^-) = (0,1); H(S_0) = .92; H(S_1) = 0; H(S) = .86$; information gain = .47

$x_3 : S_0 = 2,4,6,7; S_1 = 1,3,5; (p_0^+, p_0^-) = (.25,.75); (p_1^+, p_1^-) = (1/3, 2/3); H(S_0) = .81; H(S_1) = .92; H(S) = .86$; information gain = .03

$x_4 : S_0 = 1,2,5,6; S_1 = 3,4,7; (p_0^+, p_0^-) = (0,1); (p_1^+, p_1^-) = (2/3, 1/3); H(S_0) = 0; H(S_1) = .92; H(S) = .86$; information gain = .47

The first split will be on x_2, because it has the highest information gain. As a result the subset to be further split is S = 1, 3, 4 where $x_2 = 0$

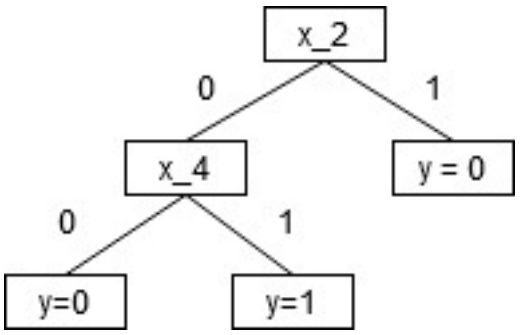
$x_1 : S_0 = 1,3; S_1 = 4; (p_0^+, p_0^-) = (.5,.5); (p_1^+, p_1^-) = (1,0); H(S_0) = 1; H(S_1) = 0; H(S) = .92$; information gain = .25

$x_3 : S_0 = 4; S_1 = 1,3; (p_0^+, p_0^-) = (1,0); (p_1^+, p_1^-) = (.5,.5); H(S_0) = 0; H(S_1) = 1; H(S) = .92$; information gain = .25

$x_4 : S_0 = 1; S_1 = 3,4; (p_0^+, p_0^-) = (0,1); (p_1^+, p_1^-) = (1,0); H(S_0) = 0; H(S_1) = 0; H(S) = .92$; information gain = .92

The second split will be on x_4 because it was the largest infomration gain in the subset. For when $x_2 = 1$ the y value will always 0.

Tree:



(b)

Boolean function: $y = x_2' \wedge x_4$

Table for function:

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

‘

2.

(a)

First split:

-Outlook:

$S_s = 1, 2, 8, 9, 11; p^+ = 2/5; MajorityError = .4$

$S_o = 3, 7, 12, 13; p^+ = 4/4 = 1; MajorityError = 0$

$S_r = 4, 5, 6, 10, 14; p^+ = 3/5; MajorityError = .4$

$InformationGain = \frac{5}{14} - \frac{5}{14} \cdot (.4 + .4) - \frac{4}{14} \cdot 0 = .07$

-Temperature:

$S_h = 1, 2, 3, 13; p^+ = 1/2; MajorityError = .5$

$S_m = 4, 8, 10, 11, 12, 14; p^+ = 2/3; MajorityError = 1/3$

$S_c = 5, 6, 7, 9; p^+ = 3/4; MajorityError = .25$
$InformationGain = \frac{5}{14} - \frac{5}{14} \cdot (.5 + .25) - \frac{6}{14} \cdot \frac{1}{3} = 0$
-Humidity:
$S_h = 1, 2, 3, 4, 8, 12, 14; p^+ = 3/7; MajorityError = 3/7$
$S_n = 5, 6, 7, 9, 10, 11, 13; p^+ = 6/7; MajorityError = 1/7$
$S_l = none$
$InformationGain = \frac{5}{14} - \frac{7}{14} \cdot (\frac{3}{7} + \frac{1}{7}) = .07$
-Wind:
$S_s = 2, 6, 7, 11, 12, 14; p^+ = 1/2; MajorityError = .5$
$S_w = 1, 3, 4, 5, 8, 9, 10, 13; p^+ = 3/4; MajorityError = .25$
$InformationGain = \frac{5}{14} - \frac{6}{14} \cdot (\frac{1}{2} - \frac{8}{14}) \cdot .25 = 0$
Outlook has the highest information gain so it will be used to split

Second split of subsets:
Sunny subset 1, 2, 8, 9, 11:
-Temperature:
$S_h = 1, 2; p^+ = 0; MajorityError = 0$
$S_m = 8, 11, 14; p^+ = 1/2; MajorityError = .5$
$S_c = 9; p^+ = 1; MajorityError = 0$
$InformationGain = .4 \cdot \frac{2}{5} \cdot .5 = .2$
-Humidity
$S_h = 1, 2, 8; p^+ = 0; MajorityError = 0$
$S_n = 9, 11; p^+ = 1; MajorityError = 0$
$S_l = none$
$InformationGain = .4$
-Wind:
$S_s = 2, 11; p^+ = 1/2; MajorityError = .5$
$S_w = 1, 8, 9; p^+ = 1/3; MajorityError = \frac{1}{3}$
$InformationGain = .4 - \frac{2}{5} \cdot .5 - \frac{3}{5} \cdot \frac{1}{3} = 0$
Humidity has the largest information gain so it is used for the second split of the sunny subset of outlook

The Humidity split has only two possible outcomes:
$S_{sh} = 1, 2, 8; p^+ = 0$
$S_{sn} = 9, 11; p^+ = 1$
so this cannot be split anymore

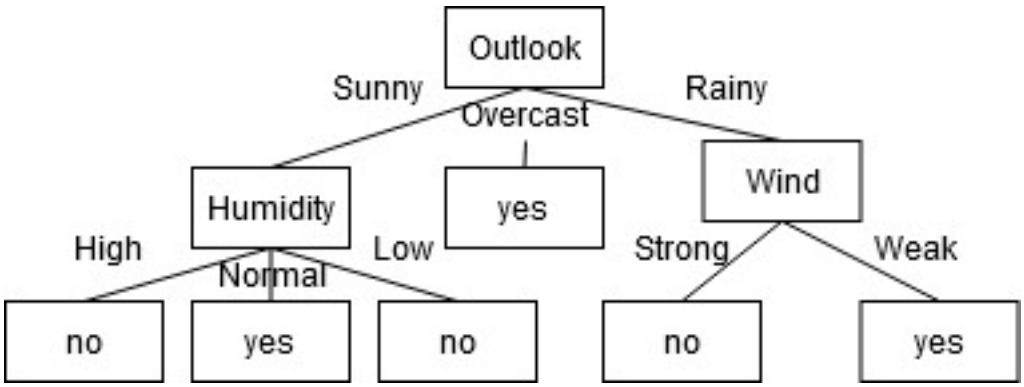
Overcast subset 3, 7, 12, 13:
For the overcast subset 3, 7, 12, 13; $p^+ = 1$ with Majority Error = 0, so this subset always results in 'yes'

Rainy Subset 4, 5, 6, 10, 14:
-Temperature:
$S_h = none$
$S_m = 4, 10, 14; p^+ = 2/3; MajorityError = .33$

$S_c = 5, 6; p^+ = 1/2; MajorityError = .5$
$InformationGain = .4 - \frac{3 \cdot .33 + 2 \cdot .5}{5} = 0$
-Humidity
$S_h = 4, 14; p^+ = 1/2; MajorityError = .5$
$S_n = 5, 6, 10; p^+ = 2/3; MajorityError = .33$
$S_l = none$
$InformationGain = .4 - frac2 \cdot .5 + 3 \cdot .335 = 0$
-Wind:
$S_s = 6, 14; p^+ = 0; MajorityError = 0$
$S_w = 4, 5, 10; p^+ = 1; MajorityError = 0$
$InformationGain = .4$
Wind has the largest information gain from the rainy subset so it will be chosen to split.

The split subsets from the Rainy subset are:
$S_{rs} = 6, 14; p^+ = 0; MajorityError = 0$
$S_{rw} = 4, 5, 10; p^+ = 1; MajorityError = 0$
So the final result can be determined from this subset

Final tree using Majority Error:



(b)

similar to part a above but using the Gini Index to calculate the information gain
First split:
-Outlook:
$S_s = 1, 2, 8, 9, 11; p^+ = 2/5; GiniIndex = .48$
$S_o = 3, 7, 12, 13; p^+ = 4/4 = 1; GiniIndex = 0$
$S_r = 4, 5, 6, 10, 14; p^+ = 3/5; GiniIndex = .48$
$InformationGain = .46 - \frac{5}{14} \cdot (.48 + .48) = .12$
-Temperature:
$S_h = 1, 2, 3, 13; p^+ = 1/2; GiniIndexr = .5$
$S_m = 4, 8, 10, 11, 12, 14; p^+ = 2/3; GiniIndex = .44$
$S_c = 5, 6, 7, 9; p^+ = 3/4; GiniIndex = .38$
$InformationGain = .46 - \frac{4 \cdot .5 + 6 \cdot .44 + 4 \cdot .38}{14} = .02$

-Humidity:
$S_h = 1, 2, 3, 4, 8, 12, 14; p^+ = 3/7; GiniIndex = .49$
$S_n = 5, 6, 7, 9, 10, 11, 13; p^+ = 6/7; GiniIndex = .24$
$S_l = none$
$InformationGain = .46 \cdot \frac{7 \cdot .49 + 7 \cdot .24}{14} = .09$
-Wind:
$S_s = 2, 6, 7, 11, 12, 14; p^+ = 1/2; GiniIndex = .5$
$S_w = 1, 3, 4, 5, 8, 9, 10, 13; p^+ = 3/4; GiniIndex = .38$
$InformationGain = .46 \cdot \frac{6 \cdot .5 + 8 \cdot .38}{14} = .03$
Outlook has the highest information gain so it will be used to split

Second split of subsets:
Sunny subset 1, 2, 8, 9, 11:
-Temperature:
$S_h = 1, 2; p^+ = 0; GiniIndex = 0$
$S_m = 8, 11, 14; p^+ = 1/2; GiniIndex = .5$
$S_c = 9; p^+ = 1; GiniIndex = 0$
$InformationGain = .48 - \frac{2}{5} \cdot .5 = .28$
-Humidity
$S_h = 1, 2, 8; p^+ = 0; GiniIndex = 0$
$S_n = 9, 11; p^+ = 1; GiniIndex = 0$
$S_l = none$
$InformationGain = .48$
-Wind:
$S_s = 2, 11; p^+ = 1/2; GiniIndex = .5$
$S_w = 1, 8, 9; p^+ = 1/3; GiniIndex = .44$
$InformationGain = .48 - \frac{2}{5} \cdot .5 - \frac{3}{5} \cdot .44 = .02$
Humidity has the largest information gain so it is used for the second split of the sunny subset of outlook

The Humidity split has only two possible outcomes:
$S_{sh} = 1, 2, 8; p^+ = 0$
$S_{sn} = 9, 11; p^+ = 1$
so this cannot be split anymore

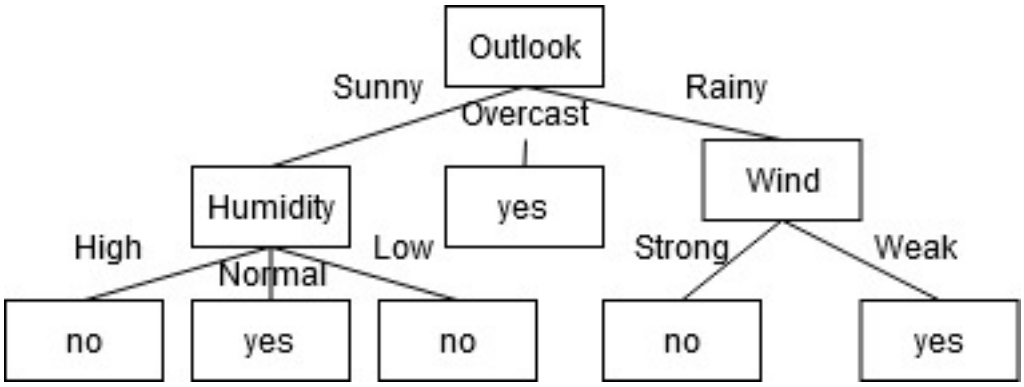
Overcast subset 3, 7, 12, 13:
For the overcast subset 3, 7, 12, 13; $p^+ = 1$ with Gini Index = 0, so this subset always results in 'yes'

Rainy Subset 4, 5, 6, 10, 14:
-Temperature:
$S_h = none$
$S_m = 4, 10, 14; p^+ = 2/3; GiniIndex = .44$
$S_c = 5, 6; p^+ = 1/2; GiniIndex = .5$
$InformationGain = .48 - .6 \cdot .44 - .4 \cdot .5 = .16$

-Humidity
$S_h = 4, 14; p^+ = 1/2; GiniIndex = .5$
$S_n = 5, 6, 10; p^+ = 2/3; GiniIndex = .44$
$S_l = none$
$InformationGain = .48 - .4 \cdot .5 - .6 \cdot .44 = .16$
-Wind:
$S_s = 6, 14; p^+ = 0; GiniIndex = 0$
$S_w = 4, 5, 10; p^+ = 1; GiniIndex = 0$
$InformationGain = .48$
Wind has the largest information gain from the rainy subset so it will be chosen to split.

The split subsets from the Rainy subset are:
$S_{rs} = 6, 14; p^+ = 0; GiniIndex = 0$
$S_{rw} = 4, 5, 10; p^+ = 1; GiniIndex = 0$
So the final result can be determined from this subset

Final tree using Gini Index (Result is the same as part a):



(c)

The trees from parts a and b are the same as the tree discussed in lecture. This is beacuse the algorithm always results in the same tree structure even if the information gain is calculated differently. Both the Majority Error and Gini Index functions work for this algorithm so the result should be the same.

3.

(a)

-Outlook:
$S_s = 1, 2, 8, 9, 11, 15; p^+ = 1/2; gain = 1$
$S_o = 3, 7, 12, 13; p^+ = 1; gain = 0$
$S_r = 4, 5, 6, 10, 14; p^+ = 3/5; gain = .97$
$InformationGain = .92 - \frac{6 \cdot 1 + 4 \cdot 0 + 5 \cdot .97}{15} = .197$
-Temperature:
$S_h = 1, 2, 3, 13; p^+ = 1/2; gain = 1$

$S_m = 4, 8, 10, 11, 12, 14, 15; p^+ = 5/7; gain = .86$
$S_c = 5, 6, 7, 9; p^+ = 3/4; gain = .81$
$InformationGain = .92 - \frac{4+7 \cdot .86+4 \cdot .81}{15} = .04$
-Humidity:
$S_h = 1, 2, 3, 4, 8, 12, 14; p^+ = 3/7; gain = .99$
$S_n = 5, 6, 7, 9, 10, 11, 13, 15; p^+ = 7/8; gain = .54$
$S_l = none$
$InformationGain = .92 - \frac{7 \cdot .99+8 \cdot .54}{14} = .17$
-Wind:
$S_s = 2, 6, 7, 11, 12, 14; p^+ = 1/2; gain = 1$
$S_w = 1, 3, 4, 5, 8, 9, 10, 13, 15; p^+ = 7/9; gain = .76$
$InformationGain = .92 - \frac{6 \cdot 1+9 \cdot .76}{15} = .06$

Outlook has the highest information gain, so it should be the chosen feature to split on.

(b)

-Outlook:
$S_s = 1, 2, 8, 9, 11; p^+ = .4; gain = .97$
$S_o = 3, 7, 12, 13, 15; p^+ = 1; gain = 0$
$S_r = 4, 5, 6, 10, 14; p^+ = 3/5; gain = .97$
$InformationGain = .92 - \frac{5 \cdot .97+5 \cdot .97}{15} = .27$
-Temperature:
$S_h = 1, 2, 3, 13; p^+ = 1/2; gain = 1$
$S_m = 4, 8, 10, 11, 12, 14, 15; p^+ = 5/7; gain = .86$
$S_c = 5, 6, 7, 9; p^+ = 3/4; gain = .81$
$InformationGain = .92 - \frac{4+7 \cdot .86+4 \cdot .81}{15} = .04$
-Humidity:
$S_h = 1, 2, 3, 4, 8, 12, 14; p^+ = 3/7; gain = .99$
$S_n = 5, 6, 7, 9, 10, 11, 13, 15; p^+ = 7/8; gain = .54$
$S_l = none$
$InformationGain = .92 - \frac{7 \cdot .99+8 \cdot .54}{14} = .17$
-Wind:
$S_s = 2, 6, 7, 11, 12, 14; p^+ = 1/2; gain = 1$
$S_w = 1, 3, 4, 5, 8, 9, 10, 13, 15; p^+ = 7/9; gain = .76$
$InformationGain = .92 - \frac{6 \cdot 1+9 \cdot .76}{15} = .06$

Outlook still has the highest information gain, so it should be the chosen feature to split on.

(c)

-Outlook:
$S_s = 1, 2, 8, 9, 11, 15(\frac{5}{15}); p^+ = \frac{2+5/14}{5+5/14} = .44; gain = .99$
$S_o = 3, 7, 12, 13, 15(\frac{4}{15}); p^+ = 1; gain = 0$
$S_r = 4, 5, 6, 10, 14, 15(\frac{5}{15}); p^+ = \frac{3+5/14}{5+5/14} = .63; gain = .95$
$InformationGain = .92 - \frac{5.36 \cdot .99+5.36 \cdot 95}{15} = .23$

-Temperature:
$S_h = 1, 2, 3, 13; p^+ = 1/2; gain = 1$
$S_m = 4, 8, 10, 11, 12, 14, 15; p^+ = 5/7; gain = .86$
$S_c = 5, 6, 7, 9; p^+ = 3/4; gain = .81$
$InformationGain = .92 - \frac{4+7\cdot.86+4\cdot.81}{15} = .04$
-Humidity:
$S_h = 1, 2, 3, 4, 8, 12, 14; p^+ = 3/7; gain = .99$
$S_n = 5, 6, 7, 9, 10, 11, 13, 15; p^+ = 7/8; gain = .54$
$S_l = none$
$InformationGain = .92 - \frac{7\cdot.99+8\cdot.54}{14} = .17$
-Wind:
$S_s = 2, 6, 7, 11, 12, 14; p^+ = 1/2; gain = 1$
$S_w = 1, 3, 4, 5, 8, 9, 10, 13, 15; p^+ = 7/9; gain = .76$
$InformationGain = .92 - \frac{6\cdot1+9\cdot.76}{15} = .06$

Outlook still has the highest information gain, so it should be the chosen feature to split on.

(d)

Starting from the subsets to continue from part c:
Sunny subset 1, 2, 8, 9, 11, 15($\frac{5}{14}$):
-Temperature:
$S_h = 1, 2; p^+ = 0; gain = 0$
$S_m = 8, 11, 15(\frac{5}{14}); p^+ = \frac{1+5/14}{2+5/14} = .58; gain = .98$
$S_c = 9; p^+ = 1; gain = 0$
$InformationGain = .99 - \frac{2.36\cdot.98}{5.36}$
-Humidity
$S_h = 1, 2, 8; p^+ = 0; gain = 0$
$S_n = 9, 11, 15(\frac{5}{14}); p^+ = 1; gain = 0$
$S_l = none$
$InformationGain = .99$
-Wind:
$S_s = 2, 11; p^+ = 1/2; gain = 1$
$S_w = 1, 8, 9, 15(\frac{5}{14}); p^+ = \frac{1+5/14}{3+5/14} = .4; gain = .97$
$InformationGain = .99 - \frac{2\cdot1+3.36\cdot.97}{5.36} = .009$
Humidity has the largest information gain so it is used for the second split of the sunny subset of outlook

The Humidity split has only two possible outcomes:
$S_{sh} = 1, 2, 8; p^+ = 0$
$S_{sn} = 9, 11; p^+ = 1$
so this cannot be split anymore

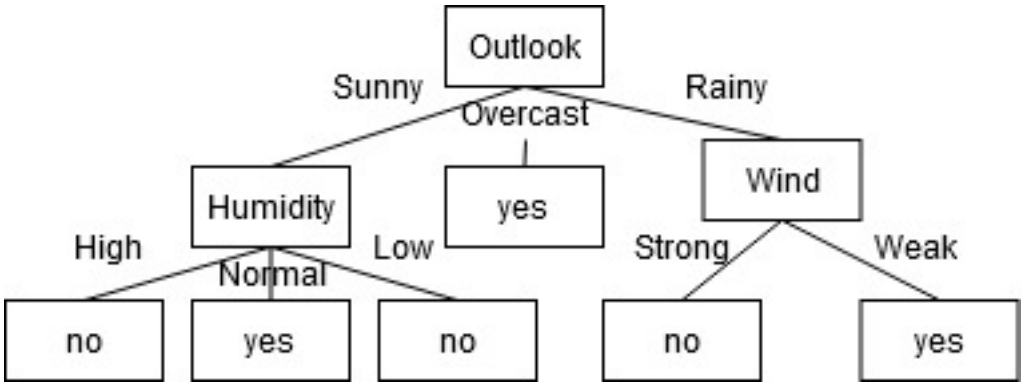
Overcast subset 3, 7, 12, 13, 15($\frac{4}{15}$):
For the overcast subset 3, 7, 12, 13, 15($\frac{4}{15}$); $p^+ = 1$ with gain = 0, so this subset

always results in 'yes'

Rainy Subset 4, 5, 6, 10, 14, 15($\frac{5}{15}$):
-Temperature:
$S_h = none$
$S_m = 4, 10, 14, 15(\frac{5}{15}); p^+ = \frac{2+5/14}{3+5/14} = .7; gain = .88$
$S_c = 5, 6; p^+ = 1/2; gain = 1$
$InformationGain = .95 - \frac{3.36 \cdot .88 + 2}{5.36} = .03$
-Humidity
$S_h = 4, 14; p^+ = 1/2; gain = 1$
$S_n = 5, 6, 10, 15(\frac{5}{14}); p^+ = \frac{2+5/14}{3+5/14} = .44; gain = .88$
$S_l = none$
$InformationGain = .95 - \frac{2+3.36 \cdot .99}{5.36} = .03$
-Wind:
$S_s = 6, 14; p^+ = 0; gain = 0$
$S_w = 4, 5, 10, 15(\frac{5}{14}); p^+ = 1; gain = 0$
$InformationGain = .95$
Wind has the largest information gain from the rainy subset so it will be chosen to split.

The split subsets from the Rainy subset are:
$S_{rs} = 6, 14; p^+ = 0; gain = 0$
$S_{rw} = 4, 5, 10; p^+ = 1; gain = 0$
So the final result can be determined from this subset

Final tree for part d:



2 Decision Tree Practice

1. GitHub Repository Link: https://github.com/BritGaul/CS5350

2.

(a)

code for decision tree is in the linked github repository under the DecisionTree directory

(b)

Errors for Training Dataset:

depth	Majority Error	Gini Index	Information Gain
1	30.2	30.2	30.2
2	30.01	22.2	22.2
3	24.7	17.59	18.10
4	21.3	8.9	8.2
5	18.01	2.7	2.7
6	17.20	0	0

Errors for Testing Dataset:

depth	Majority Error	Gini Index	Information Gain
1	29.67	29.67	29.67
2	31.59	22.25	22.25
3	26.24	18.41	19.64
4	25.96	13.74	15.11
5	22.66	8.65	8.38
6	22.66	8.52	8.38

(c)

By comparing the training errors and the testing errors, it could be said that using information gain is best approach, because it results in the lowest error for both the training datasets and the testing datasets. The Gini Index errors stay relatively close to the inforamtion gain for both datasets, but the Majoroity Error stays much higher, even when the depth increases.

3.

(a)

Errors for Training Dataset:

depth	Majority Error	Gini Index	Information Gain
1	10.88	10.88	11.92
2	10.5	10.52	10.6
3	9.76	10.1	10.22
4	8.64	8.76	8.68
5	7.84	7.38	7.14
6	7.48	5.72	5.68
7	7.28	4.5	4.52
8	7.20	4.5	4.52
9	7.18	2.94	3.2
10	7.18	2.46	2.64
11	7.18	2.24	2.34
12	7.18	2.20	2.22
13	7.18	2.20	2.22
14	7.18	2.20	2.22
15	7.18	2.20	2.22
16	7.18	2.20	2.22

Errors for Testing Dataset:

depth	Majority Error	Gini Index	Information Gain
1	12.48	11.66	12.48
2	11.02	11.53	11.26
3	11.78	10.93	10.98
4	11.58	11.64	11.78
5	11.60	12.26	12.74
6	11.96	13.26	13.64
7	11.96	14.30	14.32
8	12.04	15.48	15.24
9	12.08	16.00	16.14
10	12.14	16.24	16.88
11	12.14	16.56	16.88
12	12.14	16.52	17.08
13	12.14	16.58	17.04
14	12.14	16.58	17.04
15	12.14	16.58	17.04
16	12.14	16.58	17.04

(b)

Errors for Training Dataset:

depth	Majority Error	Gini Index	Information Gain
1	10.88	10.88	11.92
2	10.5	10.52	10.6
3	9.76	10.1	10.22
4	8.64	8.76	8.68
5	7.84	7.38	7.14
6	7.48	5.72	5.68
7	7.28	4.5	4.52
8	7.20	4.5	4.52
9	7.18	2.94	3.2
10	7.18	2.46	2.64
11	7.18	2.24	2.34
12	7.18	2.20	2.22
13	7.18	2.20	2.22
14	7.18	2.20	2.22
15	7.18	2.20	2.22
16	7.18	2.20	2.22

Errors for Testing Dataset:

depth	Majority Error	Gini Index	Information Gain
1	11.60	11.66	12.48
2	11.02	11.04	11.14
3	11.56	10.84	10.78
4	11.58	11.64	11.78
5	11.60	12.26	12.74
6	11.96	13.26	13.64
7	11.96	14.30	14.32
8	12.04	15.48	15.24
9	12.08	16.00	16.14
10	12.08	16.24	16.88
11	12.08	16.56	16.88
12	12.08	16.52	17.08
13	12.08	16.54	17.06
14	12.08	16.54	17.06
15	12.08	16.54	17.06
16	12.08	16.54	17.06

(c)

The training dataset errors become smaller as the depth of the tree is increased. But for the testing dataset the error goes down as the depth of the tree grows. The training error is always usually smaller then the test error. The error also dropped slightly when unknown was added as a new attribute.