CS 5350/6350: Machine Learining Spring 2019

Homework 1

RK Yoon | U1136736

1 Decision Tree [40 points + 10 bonus]

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

Table 1: Training data for a Boolean classifier

1. **Solution.** (a) Let S, A denote dataset and space of all attributes. Initially, we have 7 numbers of data with labels. Denote each data as number n such as $S = \{1, 2, 3, 4, 5, 6, 7\}$ for initial dataset. For each step, we need to choose the best attribute using the information gain based on entropy.

   1) $S = \{1, 2, 3, 4, 5, 6, 7\}$ and $A = \{x_1, x_2, x_3, x_4\}$. Here, $p_0 = \frac{5}{7}$ and $p_1 = \frac{2}{7}$. So the total entropy is calculated as

$$H(S) = -\frac{5}{7}\log\left(\frac{5}{7}\right) - \frac{2}{7}\log\left(\frac{2}{7}\right) = 0.2598$$

   i) When attribute is x_1,

$$H(x_1 = 0) = -\frac{4}{5}\log\left(\frac{4}{5}\right) - \frac{1}{5}\log\left(\frac{1}{5}\right) = 0.2173$$

$$H(x_1 = 1) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 0.3010$$

   Hence the information gain is computed as

$$Gain(S, x_1) = H(S) - \sum_{v \in Value} \frac{|S_v|}{|S|} H(x_1 = v) = 0.2598 - \left(\frac{5}{7}\right) 0.2173 - \left(\frac{2}{7}\right) 0.3010$$
$$= 0.0186.$$

ii) When attribute is x_2,

$$H(x_2 = 0) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) = 0.2764, \qquad H(x_2 = 1) = 0$$

$$Gain(S, x_2) = 0.2598 - \left(\frac{3}{7}\right)0.2764 = 0.1413$$

iii) When attribute is x_3,

$$H(x_3 = 0) = -\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) = 0.2442$$

$$H(x_3 = 1) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) = 0.2764$$

$$Gain(S, x_3) = 0.2598 - \left(\frac{4}{7}\right)0.8112 - \left(\frac{3}{7}\right)0.9183 = 0.018$$

iv) When attribute is x_4,

$$H(x_4 = 0) = 0, \qquad H(x_4 = 1) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) = 0.2764$$

$$Gain(S, x_2) = 0.2598 - \left(\frac{3}{7}\right)0.2764 = 0.1413$$

Since larger information gain implies the higher purity and lower uncertainty, I will choose x_2 as the best attribute. Hence if $x_2 = 1$, then the label $y = 0$ and if $x_2 = 0$, we apply ID3 algorithm again with subdata $S = \{1, 3, 4\}$ and reduced attributes set $A = \{x_1, x_3, x_4\}$.

2) $S = \{1, 3, 4\}$ and $A = \{x_1, x_3, x_4\}$. So the total entropy is calculated as

$$H(S) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) = 0.2764$$

i) When attribute is x_1,

$$H(x_1 = 0) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 0.3010, \qquad H(x_1 = 1) = 0$$

$$Gain(S, x_1) = 0.2764 - \left(\frac{2}{3}\right)0.3010 = 0.076$$

i) When attribute is x_3,

$$H(x_3 = 0) = 0, \qquad H(x_3 = 1) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 0.3010,$$

$$Gain(S, x_1) = 0.2764 - \left(\frac{2}{3}\right)0.3010 = 0.076$$

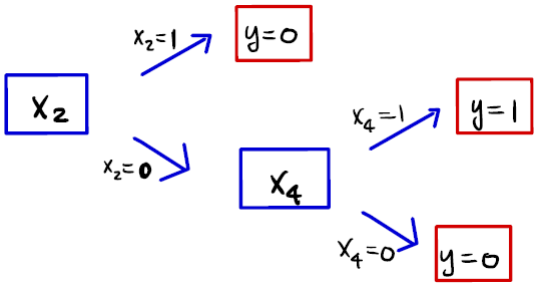
iii) When attribute is x_4,

$$H(x_4 = 0) = 0, \qquad H(x_4 = 1) = 0 \qquad \Longrightarrow \qquad Gain(S, x_4) = 0.2764$$

In this case, I will choose x_4 as the best attribute at this stage. Hence if $x_4 = 0$, then the label $y = 0$ and if $x_4 = 1$, then the label $y = 1$.

Since all data has the same label at every node, this algorithm stops. Here, Figure 1. shows the picture of obtained decision tree.

(b) Then the corresponding boolean function is given as below



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

Figure 1: (left) Figure of decision tree. (right) Boolean function for obtained decision tree

2. **Solution.** **(a)** Let $S = \{n \mid n = 1, \ldots, 14\}$ be the initial dataset and let $A = \{O, T, H, W\}$ denote initial space of attributes. For each step, we need to choose the best attribute using the majority error(ME).

1) $S = \{1, \ldots 14\}$ and $A = \{O, T, H, W\}$. Here, $p_+ = \frac{9}{14}$ and $p_- = \frac{5}{14}$. So the total ME is calculated as $ME(S) = \frac{5}{14}$.

i) When attribute is $Outlook(O)$,

$$ME(O = Sunny) = \frac{2}{5}, \quad ME(O = Overcast) = 0, \quad ME(O = Rainy) = \frac{2}{5}$$

Hence the Gain based on ME is computed as

$$Gain(S, Outlook) = ME(S) - \sum_{v \in Value} \frac{|S_v|}{|S|} ME(O = v) = \frac{5}{14} - \frac{5}{14}\frac{2}{5} - \frac{5}{14}\frac{2}{5} = \frac{1}{14}$$

ii) When attribute is $Temperature(T)$,

$$ME(T = Hot) = \frac{2}{4}, \quad ME(T = Medium) = \frac{2}{6}, \quad ME(T = Cool) = \frac{1}{4}.$$

$$Gain(S, Temperature) = \frac{5}{14} - \frac{6}{14}\frac{2}{6} - \frac{4}{14}\frac{1}{4} - \frac{4}{14}\frac{2}{4} = 0$$

iii) When attribute is $Humidity(H)$,

$$ME(H = High) = \frac{3}{7}, \quad ME(H = Normal) = \frac{1}{7}, \quad ME(H = Low) = 0.$$

$$Gain(S, Humidity) = \frac{5}{14} - \frac{7}{14}\frac{3}{7} - \frac{7}{14}\frac{1}{7} = \frac{1}{14}$$

iv) When attribute is $Wind(W)$,

$$ME(W = Strong) = \frac{3}{6}, \quad ME(W = Weak) = \frac{2}{8}.$$

$$Gain(S, Wind) = \frac{5}{14} - \frac{6}{14}\frac{3}{6} - \frac{8}{14}\frac{2}{8} = 0$$

Here the largest gain is obtained when the attribute is O and H. However, our training dataset miss the one of the feature, Low, in the Humidity attribute. So I choose the "Outlook" as the best attribute in this stage. Hence if $O = Overcast$, then it returns the label $y =$ " $+$ " and if $O = Sunny$ or $O = Rainy$, then we repeat this recursive ID3 algorithm respectively.

2)-1. Below the branch $O = Sunny$ with subdata $S = \{1, 2, 8, 9, 11\}$ and $A = \{T, H, W\}$. Here So the total ME is calculated as $ME(S) = \frac{2}{5}$. Here, we can easily check that H is the best attribute because

$$ME(H = High) = 0, \quad ME(H = Normal) = 0, \quad ME(H = Low) = 0,$$

which implies that $Gain(S, Humidity) = \frac{2}{5}$ where it is the maximum gain based on the ME at this depth. Hence if $H = High$, then it returns a label $y =$ " $-$ " and if $H = Normal$, then it returns a label $y =$ " $+$ ". Here I will assign the common label $y =$ " $+$ " for missing feature $H = Low$.

2)-2. Below the branch $O = Rainy$ with subdata $S = \{4, 5, 6, 10, 14\}$ and $A = \{T, H, W\}$. Here So the total ME is calculated as $ME(S) = \frac{2}{5}$. Similarly, it is obvious that W is the best attribute because

$$ME(W = Strong) = 0, \quad ME(W = Weak) = 0, \qquad \Longrightarrow \qquad Gain(S, Humidity) = \frac{2}{5},$$

where it is the maximum gain based on the ME at this level. Hence if $W = Strong$, then it returns a label $y =$ "$-$" and if $W = Weak$, then it returns a label $y =$ "$+$".

Then Figure2. shows the diagram for constructed decision tree using the ME.

(b) We will find the best splitting attribute using the GI(Gini Index). Withe the same process above, let $S = \{n \mid n = 1, \ldots, 14\}$ and $A = \{O, T, H, W\}$ be the initial dataset and space of attributes respectively.

1) $S = \{1, \ldots 14\}$ and $A = \{O, T, H, W\}$. Here, $p_+ = \frac{9}{14}$ and $p_- = \frac{5}{14}$. So the total GI is calculated as $GI(S) = 1 - \left(\frac{5}{14}\right)^2 - \left(\frac{9}{14}\right)^2 = 0.4591$.

i) When attribute is $Outlook(O)$,

$$GI(O = Sunny) = GI(O = Rainy) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48,$$

and $GI(O = Overcast) = 0$. Hence the gain based on the GI is computed as

$$Gain(S, Outlook) = GI(S) - \sum_{v \in Value} \frac{|S_v|}{|S|} GI(O = v) = 0.4591 - \frac{5}{14}0.48 - \frac{5}{14}0.48 = 0.1162$$

ii) When attribute is $Temperature(T)$,

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

$$GI(T = Medium) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.44,$$

$$GI(T = Cool) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0375.$$

$$Gain(S, Temperature) = 0.4591 - \frac{6}{14}0.44 - \frac{4}{14}0.5 - \frac{4}{14}0.375 = 0.02$$

iii) When attribute is $Humidity(H)$,

$$GI(H = High) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.4878,$$

$$GI(H = Normal) = 1 - \left(\frac{1}{7}\right)^2 - \left(\frac{6}{7}\right)^2 = 0375, \quad GI(H = Low) = 0.$$

$$Gain(S, Humidity) = 0.4591 - \frac{7}{14}0.4898 - \frac{7}{14}0.2449 = 0.09175.$$

iv) When attribute is $Wind(W)$,

$$GI(W = Strong) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$GI(W = Weak) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375.$$

$$Gain(S, Wind) = 0.4591 - \frac{6}{14}0.5 - \frac{8}{14}0.375 = 0.0305$$

Hence I choose the "Outlook" as the best attribute in this stage because the largest gain is obtained when the attribute is O. So Hence if $O = Overcast$, then it returns the label $y = " + "$ and if $O = Sunny$ or $O = Rainy$, then we repeat this recursive ID3 algorithm respectively.

2)-1. Below the branch $O = Sunny$ with subdata $S = \{1, 2, 8, 9, 11\}$ and $A = \{T, H, W\}$. Here So the total GI is calculated as $GI(S) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$. With the same reason in part(a), H is the best attribute because

$$GI(H = High) = GI(H = Normal) = GI(H = Low) = 0,$$

which implies that $Gain(S, Humidity) = 0.48$ where it is the maximum gain based on the GI at this depth. Again, I will assign the common label $y = " + "$ for missing feature $H = Low$.

2)-2. Below the branch $O = Rainy$ with subdata $S = \{4, 5, 6, 10, 14\}$ and $A = \{T, H, W\}$. Here So the total GI is equal to $GI(S) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$. Similarly, it is obvious that W is the best attribute because

$$GI(W = Strong) = GI(W = Weak) = 0, \qquad \Longrightarrow \qquad Gain(S, Humidity) = 0.48.$$

Then Figure2. shows the diagram for constructed decision tree using the GI.

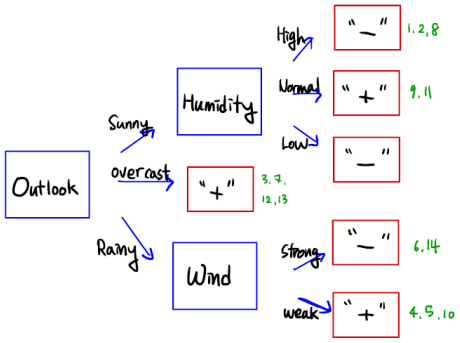


Figure 2: Figure of decision tree using ME(Majority error) and GI(Gini Index). I got the same decision tree.

(c) By comparing those two diagram with the decision tree what we got in the class, we can conclude that we got the same decision tree. The reason is that the entropy, majority error and Gini index are the methods to compute the purity of the dataset, more precisely, bigger number has higher purity and lower uncertainty. In this tennis example case, three methods measure its purity similarly. But there can be special case of dataset which returns different decision tree. For example, when I compute the majority error at the first depth, $ME(H)$ and $ME(O)$ provides same results even if their GI are different. Depending on the choice to same scores, resulting decision tree can be different but the tendency of three measurements would be similar.

3. **Solution.** Now we add one more train data which has missing feature value.
(a) In the training data, there are five S, four O and five R. Let the missing value is the common value S. As before, let $S = \{1, \ldots, 15\}$ and A={O,T,H,W} be the dataset and space of attributes. Here, $p_+ = \frac{10}{15}$ and $p_- = \frac{5}{15}$. So the total entropy is calculated as

$$H(S) = -\frac{10}{15}\log\left(\frac{10}{15}\right) - \frac{5}{15}\log\left(\frac{5}{15}\right) = 0.2764$$

O	T	H	W	Play
S	M	N	W	+

i) When attribute is $Outlook(O)$,

$$H(O = Sunny) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 0.3010,$$

$$H(O = Rainy) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.2923,$$

and $H(O = Overcast) = 0$. Hence the information gain is computed as

$$Gain(S, Outlook) = 0.2764 - \left(\frac{6}{15}\right)0.3010 - \left(\frac{5}{15}\right)0.2923 = 0.0586.$$

ii) When attribute is $Temperature(T)$,

$$H(T = High) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 0.3010$$

$$H(T = Medium) = -\frac{5}{7}\log\left(\frac{5}{7}\right) - \frac{2}{7}\log\left(\frac{2}{7}\right) = 0.2598$$

$$H(T = Cool) = -\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right) = 0.2442$$

$$Gain(S, Temperature) = 0.2764 - \left(\frac{4}{15}\right)0.3010 - \left(\frac{7}{15}\right)0.2598 - \left(\frac{4}{15}\right)0.2442 = 0.0098.$$

iii) When attribute is $Humidity(H)$,

$$H(H = High) = -\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right) = 0.2966$$

$$H(H = Normal) = -\frac{7}{8}\log\left(\frac{7}{8}\right) - \frac{1}{8}\log\left(\frac{1}{8}\right) = 0.1636$$

and $H(H = Low) = 0$. Then the information gain is computed as

$$Gain(S, Humidity) = 0.2764 - \left(\frac{7}{15}\right)0.2966 - \left(\frac{8}{15}\right)0.1636 = 0.0507.$$

iv) When attribute is $Wind(W)$,

$$H(W = Strong) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 0.3010$$

$$H(W = weak) = -\frac{2}{9}\log\left(\frac{2}{9}\right) - \frac{7}{9}\log\left(\frac{7}{9}\right) = 0.2300$$

$$Gain(S, Temperature) = 0.2764 - \left(\frac{6}{15}\right)0.3010 - \left(\frac{9}{15}\right)0.2300 = 0.0018.$$

Even if we add one more data, the best attribute at the first depth is still the *Outlook*.

Actually, R also can be considered as the common feature value. I manually computed the gain when I replace missing value by R and the resulting best attribute is still *Outlook*.

(b) Among the same label " + ", we have two S, four O and three R. Hence, let's choose O as the value for missing data. And then do a same process with part (a). Now,

O	T	H	W	Play
O	M	N	W	+

$S = \{1, \ldots, 15\}$ and A={O,T,H,W} . We can keep everything except the information gain of "Outlook" because we only change its value. Again, we still have total entropy $H(S) = 0.2764$. Actually, we will get the same information gain except when attribute is "Outlook". But by adding the " + " label having $O = overcast$, we will get lower uncertainty, i.e. higher purity. More precisely, the entropy of each values are computed as

$$H(O = Sunny) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = 0.2922,$$

$$H(O = Rainy) = 0.2923, \qquad H(O = Overcast) = 0.$$

Thus its information gain will increase to

$$Gain(S, Outlook) = 0.2764 - \left(\frac{5}{15}\right)0.2922 - \left(\frac{5}{15}\right)0.2923 = 0.0816.$$

Hence this attribute *Outlook* would be more powerful attribute splitting my dataset. Hence the best feature is still *Outlook*.

(c) Now choose the missing value using the fractional counts as below. Consider the

name	counts	O	T	H	W	Play
15	$\frac{5}{14}$	S	M	N	W	+
16	$\frac{4}{14}$	0	M	N	W	+
17	$\frac{5}{14}$	R	M	N	W	+

information gain for *Outlook*. Here the total entropy is preserved as $H(S) = 0.2764$ because the total number of label " + " is same.

i) When $O = Sunny$,

$$p_+ = \frac{2 + \frac{5}{14}}{5 + \frac{5}{14}} = 0.44, \; p_- = \frac{3}{5 + \frac{5}{14}} = 0.56 \;\Rightarrow\; H(O = S) = 0.2979$$

ii) When $O = Rainy$,

$$p_+ = \frac{3 + \frac{5}{14}}{5 + \frac{5}{14}} = 0.6267, \; p_- = \frac{2}{5 + \frac{5}{14}} = 0.3733 \;\Rightarrow\; H(O = R) = 0.2869$$

Then the information gain for "$Outlook$" is computed as

$$Gain(S, Outlook) = 0.2764 - \frac{5 + \frac{5}{14}}{15} * 0.2979 - \frac{5 + \frac{5}{14}}{15} * 0.3733 = 0.03668,$$

which is less than $Gain(S, Humidity) = 0.0507$. Hence, now the best attribute is $Humidity$ instead of $Outlook$.

(d). Now, we consider the ID3 algorithm at depth 2.

2)-1. Below the branch $H = High$ with subdata $S = \{1, 2, 3, 4, 8, 12, 14\}$ and $A = \{O, T, W\}$. Here the total entropy is

$$H(S) = -\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right) = 0.2966$$

i) When next attribute is $Outlook(O)$,

$$H(O = Sunny) = H(O = Overcast) = 0$$

$$H(O = Rainy) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 0.3010.$$

$$Gain(S, Outlook) = 0.2966 - \frac{2}{7}0.3010 = 0.2106$$

ii) When next attribute is $Temperature(T)$,

$$H(T = High) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) = 0.2764$$

$$H(T = Medium) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 0.3010.$$

$$Gain(S, Temperature) = 0.2966 - \frac{3}{7} * 0.2764 - \frac{4}{7}0.3010 = 0.0061$$

iii) When next attribute is $Wind(W)$,

$$H(W = Strong) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) = 0.2764$$

$$H(W = Weak) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 0.3010.$$

$$Gain(S, Wind) = 0.2966 - \frac{3}{7} * 0.2764 - \frac{4}{7}0.3010 = 0.0061$$

Hence the best attribute is "Outlook". If $O = Sunny$, it returns label $y = " - "$, and if $O = Overcast$, then it provides label $y = " + "$. When $O = Rainy$, we apply ID3 one more time with subdata $S = \{4, 14\}$ and attributes set $A = \{T, W\}$. Actually, the subdata has same value "Medium" for attribute $Temperature(T)$, the next best attribute splitting subdata is $Wind(W)$. If $W = Weak$, then it returns label $y = " + "$ and if $W = Strong$, then it provides label $y = " - "$

2)-2. Below the branch $H = Normal$ with subdata $S = \{5, 6, 7, 9, 10, 11, 13, 15, 16, 17\}$ and $A = \{O, T, W\}$. Here the total entropy is

$$H(S) = -\frac{1}{8}\log\left(\frac{1}{8}\right) - \frac{7}{8}\log\left(\frac{7}{8}\right) = 0.1636$$

i) When next attribute is $Outlook(O)$,

$$H(O = Sunny) = H(O = Overcast) = 0$$

$$H(O = Rainy) = -\frac{1}{3+\frac{5}{14}}\log\left(\frac{1}{3+\frac{5}{14}}\right) - \frac{2+\frac{5}{14}}{3+\frac{5}{14}}\log\left(\frac{2+\frac{5}{14}}{3+\frac{5}{14}}\right) = 0.2645.$$

$$Gain(S, Outlook) = 0.1636 - \frac{3+\frac{5}{14}}{8}0.2645 = 0.0526$$

ii) When next attribute is $Temperature(T)$,

$$H(T = High) = H(T = Medium) = 0$$

$$H(T = Cool) = -\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right) = 0.2442.$$

$$Gain(S, Temperature) = 0.1636 - \frac{4}{8} * 0.2442 = 0.0415$$

iii) When next attribute is $Wind(W)$,

$$H(W = Weak) = 0$$

$$H(W = Strong) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) = 0.2442.$$

$$Gain(S, Wind) = 0.1636 - \frac{3}{8} * 0.2764 = 0.05995$$

Hence the best attribute is "Wind". If $W = Weak$, it returns label $y = " + "$. When $W = Strong$, we apply ID3 one more time with subdata $S = \{6, 7, 11\}$ and attributes set $A = \{O, T\}$.

If we choose $Temperature$ as next node, we need one more branch because there are no common node when $T = Cool$. On the other hand, if we choose $Outlook$ as the next node, we can stop. Since we would like to get less deeper decision tree, the best attribute at that depth is the $Outlook$. If $O = Rainy$, it returns label $y = " - "$ and if $O = Overcast$ or $O = Sunny$, it provides the label $y = " + "$.

Actually, the subdata has same value "Medium" for attribute $Temperature(T)$, the next best attribute splitting subdata is $Wind(W)$. If $W = Weak$, then it returns label $y = " + "$ and if $W = Strong$, then it provides label $y = " - "$

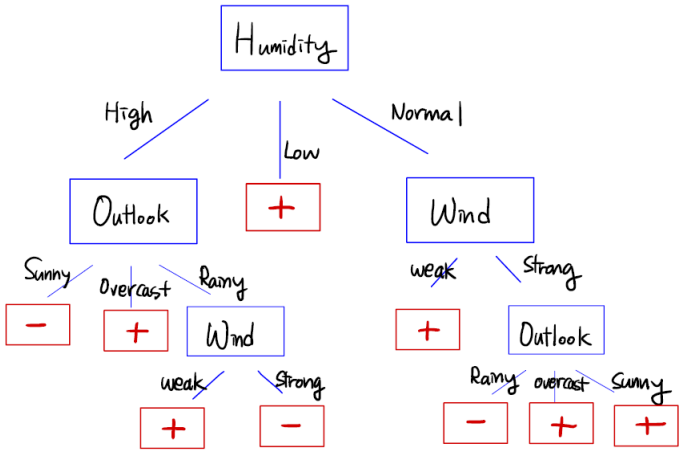


Figure 3: Decision tree after adding new data with the fractional counts method.

4. [**Bonus question 1.**] **Solution.** The information gain is defined as

$$Gain = H(s) - \sum_{v \in \text{Value}} \frac{|S_v|}{|S|} H(S_v) = \sum_{i \in \text{Label}} -p(i)\log p(i) - \sum_v p(v) \sum_i -p(i|v)\log p(i|v).$$

By the definition of conditional property, $p(i|v)p(v) = p(i,v)$,

$$Gain = \sum_i \sum_v -p(i,v)\log p(i) + \sum_i \sum_v p(v)p(i|v)\log p(i|v)$$

$$= -\sum_i \sum_v p(v)p(i|v)[\log p(i) - \log p(i|v)] = -\sum_i \sum_v p(v)p(i|v)\log\left(\frac{p(i)}{p(i|v)}\right).$$

By changing the order of summation,

$$Gain = -\sum_v p(v) \sum_i p(i|v)\log\left(\frac{p(i)}{p(i|v)}\right).$$

Since $y = \log x$ is convex function and $p(i,v) \geq 0, \ \ \forall i$ and $\sum_i p(i|v) = 1$, we can apply the Jensen's inequality $f(\sum_i \alpha_i x_i) \leq \sum_i \alpha_i f(x_i)$, and thus

$$Gain \geq -\sum_v p(v)\log\left(\sum_i p(i|v)\frac{p(i)}{p(i|v)}\right) = -\sum_v p(v)\log\sum_i p(i) = -\sum_v p(v)\log 1 = 0.$$

Hence the information gain is always non-negative.

5. [**Bonus question 2**]. **Solution.** The information gain measures the purity of the labels. So we need to find the measurement determining the purity of the numerical labels. Actually, the standard deviation measures the amount of variation a set of data values. That is, large standard deviation indicates that the dataset tend to be far to the average of the set. Hence, instead of finding the information gain of each attribution,

we can compute the standard deviation of numerical labels of each attribute. Then the attribute having the smallest standard deviation implies that the corresponding labels are around the its averages and its purity is high. So that attribute becomes the best attribute splitting dataset.

2 Decision Tree Practice [60 points]

1. Here is the link to my repository : `https://github.com/rkyoon12/CS6350-spring-2019`

2. **Solution.**

   (a) **Solution.** I implement the ID3 algorithm supporting information gain, majority error and gini index. And I allow user to vary the depth form 1 to 6.

   (b) **Solution.** I implement code to do prediction by varying the depth form 1 to 6. Here the accuracy is computed by

$$\text{Accuracy} = \frac{\text{Prediction} = \text{Label}}{\text{Total number of data}} * 100(\%)$$

   Table 2. and **Table 3.** show the accuracy of prediction about training and testing dataset at each depth by using three different splitting methods as below.

Depth	InfoGain(%)	ME(%)	GI(%)
1	69.80	69.80	69.80
2	77.80	70.80	77.80
3	81.90	79.40	82.40
4	91.80	84.00	91.10
5	97.30	88.40	97.3
6	100	89.60	100

Table 2: The accuracy of predicting training dataset.

Depth	InfoGain(%)	ME(%)	GI(%)
1	70.33	70.33	70.33
2	77.75	68.68	77.75
3	80.36	77.88	81.59
4	84.89	79.12	86.26
5	90.11	82.01	90.11
6	87.50	80.63	87.50

Table 3: The accuracy of predicting test dataset.

   (c) **Solution.** As shown above table, the accuracy of predicting training data increases as a depth of decision tree is deeper. On the other hand, the accuracy of predicting testing data increase upto upto depth=5, but after that, it decreases

again. And three different methods provides similar tendency. As we learned in class. this phenomena is caused by **overfitting** of obtained decision tree. When we make the decision tree fit with the training data, it would be fail in predicting general data(i.e. fail in generalization). Among three ways finding the best attributes, the performance of majority error is worse than the others when the label is not binary (i.e. with the multilabels).

3. **Solution**. **(a)** In this part, we consider the value "unkown" as a particular attribute value. **Table 4.** presents the accuracy of training and test data with information gain, ME, and GI by varying the tree depth from 1 to 16.

   (b) In this part, we consider the "unknown" as the missing value. And this missing value is replaced by the common value in the same attribute. For example, there are "unknown"s in the "Contact". Since the common value of this feature is "cellular" in the training set, so every "unknown" values in "Contact" is substituted by "cellular". **Table 5.** shows the accuracy of training and test data with information gain, ME, and GI by varying the tree depth from 1 to 16.

   (c) According to the result, tendency of the test accuracy for every methods is increasing as the depth is deeper, but after some depth, test accuracy decreases. Again, if we construct tree fitting to the training data, we lose the generalization and thus is not fitting on the test dataset. Depending on the method I used for gain, I got the different trees, but their performances are good, get almost 89% of accuracy on the testdata. After a specific depth, the accuracy stop increasing or decreasing. I think it happens because the left attributes are highly correlated with the previous used attribute.

   By comparing with the result of (a) and (b), I got more nice performance on the (a). Even if I complete the "unknown" as common value in same feature, it is still inaccurate in explaining the dataset and finding the relation with label and given information. Instead of completing missing data with the common value, I think the fractional counting is better to approximate the data.

IG	Train(%)	Test(%)
1	88.08	87.44
2	89.40	88.78
3	89.94	89.10
4	92.00	87.34
5	93.76	86.34
6	95.20	85.18
7	96.28	84.28
8	97.08	83.65
9	97.78	83.18
10	98.18	82.72
11	98.50	82.68
12	98.64	82.40
13	98.68	82.34
14	-	-
15	-	-
16	-	-

ME	Train(%)	Test(%)
1	89.12	88.34
2	89.58	89.10
3	90.34	88.58
4	91.52	87.28
5	93.12	86.28
6	93.64	85.38
7	93.94	85.32
8	94.02	85.24
9	94.06	85.24
10	94.08	85.22
11	94.10	85.22
12	-	-
13	-	-
14	-	-
15	-	-
16	-	-

GI	Train(%)	Test(%)
1	89.12	88.34
2	89.58	89.10
3	90.64	88.38
4	92.46	86.88
5	93.96	85.40
6	95.16	83.80
7	96.34	82.80
8	97.34	82.22
9	97.86	82.00
10	98.28	81.66
11	98.60	81.32
12	98.64	81.30
13	98.68	81.24
14	-	-
15	-	-
16	-	-

Table 4 : The accuracy of predicting train and test data by replacing missing by changing numerical features to binary one.

IG	Train(%)	Test(%)
1	88.08	87.44
2	89.40	88.78
3	89.78	88.86
4	91.24	87.20
5	93.02	86.46
6	94.40	85.40
7	95.56	84.76
8	96.34	84.56
9	96.96	83.90
10	97.56	83.62
11	97.94	83.50
12	98.18	83.34
13	98.20	83.34
14	-	-
15	-	-
16	-	-

ME	Train(%)	Test(%)
1	89.12	88.34
2	89.50	88.96
3	90.24	88.38
4	91.34	87.70
5	92.18	87.38
6	92.52	87.40
7	92.66	87.20
8	92.72	87.16
9	92.76	87.08
10	92.76	87.08
11	-	-
12	-	-
13	-	-
14	-	-
15	-	-
16	-	-

GI	Train(%)	Test(%)
1	89.12	88.34
2	89.48	88.88
3	89.90	88.96
4	91.18	87.38
5	92.80	86.16
6	94.36	84.94
7	95.56	84.12
8	96.50	83.70
9	97.04	83.14
10	97.64	82.84
11	98.00	82.84
12	98.16	82.70
13	98.20	82.66
14	-	-
15	-	-
16	-	-

Table5 : The accuracy of predicting train and test data by replacing missing by the majority attribute from the training dataset.