CS 5350/6350: Machine Learining Spring 2021

Homework 1

Handed out: 2 Feb, 2021
Due date: 11:59pm, 19 Feb, 2021

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. [7 points] Decision tree construction.

   (a) [5 points]

   Step 1: $ID3(S, Attributes, Label)$ Where:

   S = The set of examples (i.e., the training data)

   Attributes = x_1, x_2, x_3, x_4

   Label = Target attribute (prediction) = 1

   Step 2: If all examples aren't the same, create a root node for the tree.

   Step 3: Find the attribute that best Splits S.

   Find current entropy:

$$Entropy(S) = H(S) = -p_+ \log(p_+) - p_- \log(p_-)$$

   $S = \{0, 0, 1, 1, 0, 0, 0\}$

   $Entropy(S) = -\frac{2}{7}\log_2(\frac{2}{7}) - \frac{5}{7}\log_2(\frac{5}{7}) = .8631$

   Next, find which attributes has the largest information gain:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Current attributes: x_1, x_2, x_3, x_4

Let the first row of the training data be t1, the second row be t2..., and the last row be t7

$S_1 = \{t4, t6\}$ $S_0 = \{t1, t2, t3, t5, t7\}$

$Gain(S, x_1) = .8631 - [\frac{2}{7}(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})) - \frac{5}{7}(-\frac{1}{5}\log_2(\frac{1}{5}) - \frac{4}{5}\log_2(\frac{4}{5}))] = .0617$

$S_1 = \{t2, t5, t6, t7\}$ $S_0 = \{t1, t3, t4\}$

$Gain(S, x_2) = .8631 - [\frac{4}{7}(-0\log_2(0) - \frac{4}{4}\log_2(\frac{4}{4})) - \frac{3}{7}(-\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}))] = .469$

$S_1 = \{t1, t3, t5\}$ $S_0 = \{t2, t4, t6, t7\}$

$Gain(S, x_3) = .8631 - [\frac{3}{7}(-\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3})) - \frac{4}{7}(-\frac{1}{4}\log_2(\frac{1}{4}) - \frac{3}{4}\log_2(\frac{3}{4}))] = .029$

$S_1 = \{t3, t4, t7\}$ $S_0 = \{t1, t2, t5, t6\}$

$Gain(S, x_4) = .8631 - [\frac{3}{7}(-\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3})) - \frac{4}{7}(-0\log_2(0) - \frac{4}{4}\log_2(\frac{4}{4}))] = .469$

x_2 and x_4 both have the largest information gain therefore we will split S with attribute x_2

Step 4: Next, add a branch for every value of A{i.e., 1,0}

Step 5: Then, let S_v be the subset of examples of S where $v = A$:

$S_1 = \{2, 5, 6, 7\}$ and $S_2 = \{1, 3, 4\}$

Step 6: $ID3(S_1, Attributes - \{x_2\}, 0)$

Because all of the values in S_1 have the same label, a leaf node with label 0 will be returned.

Step 7: $ID3(S_0, Attributes - \{x_2\}, 1)$ Step 2-6 will be repeated here:

Step 8: If all examples aren't the same, create a root node for the tree.

Step 9: Find the attribute that best Splits S.

Find current entropy:

$S = \{0, 1, 1\}$

$Entropy(y) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = .918$

Next, find which attributes has the largest information gain:

Current attributes: x_1, x_3, x_4

Let the first row of the training data be t1, the second row be t2..., and the last row be t7

$S_1 = \{t4\} \qquad S_0 = \{t1, t3\}$

$Gain(S, x_1) = .918 - [\frac{1}{3}(-\frac{1}{1}\log_2(\frac{1}{1}) - 0\log_2(0) - \frac{2}{3}(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}))] = .215$

$S_1 = \{t1, t3\} \qquad S_0 = \{t4\}$

$Gain(S, x_3) = .918 - [\frac{2}{3}(-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})) - \frac{1}{3}(\frac{1}{1}\log_2(\frac{1}{1}) - 0\log_2(0))] = .251$

$S_1 = \{t3, t4\} \qquad S_0 = \{t1\}$

$Gain(S, x_4) = .918 - [\frac{2}{3}(-\frac{2}{2}\log_2(\frac{2}{2}) - 0\log_2(0)) - \frac{1}{3}(-0\log_2(0) - \frac{1}{1}\log_2(\frac{1}{1}))] = .918$

x_4 has the largest information gain therefore we will split S with attribute x_4

Step 10: Next, add a branch for every value of A{i.e., 1,0}

Step 11: Then, let S_v be the subset of examples of S where $v = A$:

$S_1 = \{t3, t4\}$ and $S_0 = \{t1\}$

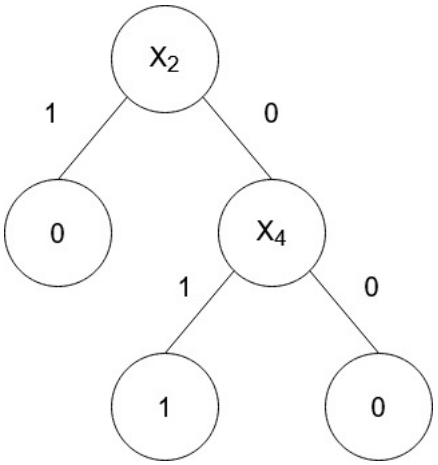
Step 12: $ID3(S_1, Attributes - \{x_4\}, 1)$

Because all of the values in S_1 have the same label, a leaf node with label 1 will be returned.

Step 13: $ID3(S_0, Attributes - \{x_4\}, 0)$

Because all of the values in S_0 have the same label, a leaf node with label 0 will be returned.

The final decision tree looks like:



(b) [2 points]

x_2	x_4	y
1	1	0
1	0	0
0	1	1
0	0	0

Table 2: Boolean function for decision tree

2. [17 points]

(a) [7 points]

Step 1: Find the attribute that best Splits S.

Find current Majority Error (ME):

ME(S) = "Suppose the tree was not grown below this node and the majority label were chosen, what would be the error?"

$S = \{-, -, +, +, +, -, +, -, +, +, +, +, +, -\}$

$ME(S) = \frac{5}{14} = .357$

Next, find which attributes has the largest information gain:

$$Gain(S, A) = ME(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} ME(S_v)$$

Current attributes A = {Outlook, Temperature, Humidity, Wind}

$S_S = \{t1, t2, t8, t9, t11\}$ $S_O = \{t3, t7, t12, t1\}$ $S_R = \{t4, t5, t6, t10, t14\}$

$Gain(S, Outlook) = .357 - [(\frac{5}{14})(\frac{2}{5}) + (\frac{4}{14})(\frac{0}{4}) + (\frac{5}{14})(\frac{2}{5})] = .0712$

$S_H = \{t1, t2, t3, t13\}$ $S_M = \{t4, t8, t10, t11, t12, t14\}$ $S_C = \{t5, t6, t7, t9\}$

$Gain(S, Temperature) = .357 - [(\frac{4}{14})(\frac{2}{4}) + (\frac{6}{14})(\frac{2}{6}) + (\frac{4}{14})(\frac{1}{4})] = 0$

$S_H = \{t1, t2, t3, t4, t8, t12, t14\}$ $S_N = \{t5, t6, t7, t9, t10, t11, t13\}$ $S_L = \{\}$

$Gain(S, Humidity) = .357 - [(\frac{7}{14})(\frac{3}{7}) + (\frac{7}{14})(\frac{1}{7}) + 0)] = .0712$

$S_S = \{t2, t6, t7, t11, t12, t14\}$ $S_W = \{t1, t3, t4, t5, t8, t9, t10, t13\}$

$Gain(S, Wind) = .357 - [(\frac{6}{14})(\frac{3}{6}) + (\frac{8}{14})(\frac{2}{8}))] = 0$

Outlook and Humidity both have the largest information gain therefore we will split S with attribute Outlook.

Because all of the values in Overcast have the same label, a leaf node with label Yes will be returned.

Step 2: Find the attribute that best Splits S again for Sunny.

Find current Majority Error(ME):

$S = \{t1, t2, t8, t9, t11\}$ ME(S) $= \frac{2}{5} = .4$

Next, find which attributes has the largest information gain:

Current attributes: A = {Temperature, Humidity, Wind}

$S_H = \{t1, t2\}$ $S_M = \{t8, t11\}$ $S_C = \{t9\}$

$Gain(S, Temperature) = .4 - [0 + (\frac{2}{5})(\frac{1}{2}) + 0)] = .2$

$S_H = \{t1, t2, t8\}$ $S_N = \{t9, t11\}$ $S_L = \{\}$

$Gain(S, Humidity) = .4 - [(0 + 0 + 0)] = .4$

$S_S = \{t2, t11\}$ $S_W = \{t1, t8, t9\}$

$Gain(S, Wind) = .4 - [(\frac{2}{5})(\frac{1}{2}) + (\frac{3}{5})(\frac{1}{3}))] = 0$

Humidity has the largest information gain therefore we will split S with attribute of Humidity

Because all of the values in S_H, S_N, and S_L (i.e., High, Normal, or Low Humidity) have the same label, a leaf node with label No, Yes, No will be returned respectively.

Step 3: Find the attribute that best Splits S again for Rain.

Find current Majority Error(ME):

$S = \{t4, t5, t6, t10, t14\}$ ME(S) $= \frac{2}{5} = .4$

Next, find which attributes has the largest information gain:

Current attributes: A = {Temperature, Humidity, Wind}

$S_H = \{\}$ $S_M = \{t4, t10, t14\}$ $S_C = \{t5, t6\}$

$Gain(S, Temperature) = .4 - [0 + (\frac{3}{5})(\frac{1}{3}) + 0)] = .2$

$S_H = \{t4, t14\}$ $S_N = \{t5, t6, t10\}$ $S_L = \{\}$

$Gain(S, Humidity) = .4 - [(\frac{2}{5})(\frac{1}{2}) + (\frac{3}{5})(\frac{1}{3}) + 0)] = 0$

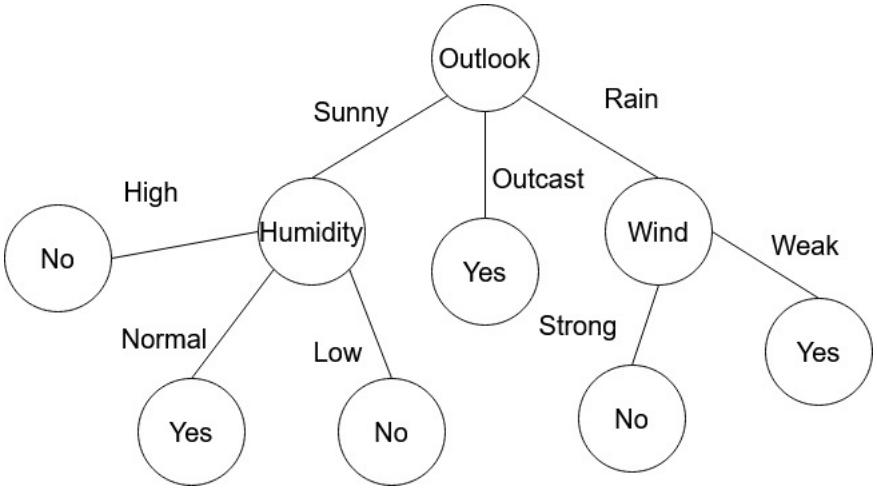
$S_S = \{t4, t5, t10\}$ $S_W = \{t6, t14\}$

$Gain(S, Wind) = .4 - [0 + 0] = .4$

Wind has the largest information gain therefore we will split S with attribute of Wind

Because all of the values in S_S and S_W (i.e., Strong and Weak Wind) have the same label, a leaf node with label No and Yes will be returned respectively.

The final decision tree looks like:



(b) [7 points] Please use gini index (GI) to calculate the gain, and conduct tree learning with ID3 framework. List every step and the tree structure.

Step 1: Find the attribute that best Splits S.

Find current Gini Index (GI):

GI(S) =

$$1 - \sum_{t=1}^{k} p^2 k$$

$S = \{-, -, +, +, +, -, +, -, +, +, +, +, +, -\}$

$GI(S) = 1 - (\frac{9}{14})^2 - (\frac{5}{14})^2 = .714$

Next, find which attributes has the largest information gain:

$$Gain(S, A) = GI(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} GI(S_v)$$

Current attributes A = {Outlook, Temperature, Humidity, Wind}

$S_S = \{t1, t2, t8, t9, t11\}$ $S_O = \{t3, t7, t12, t1\}$ $S_R = \{t4, t5, t6, t10, t14\}$

$Gain(S, Outlook) = .714 - [(\frac{5}{14})(1 - (\frac{3}{5})^2 - (\frac{2}{5})^2)] + [(\frac{4}{14})(1 - (\frac{4}{4})^2 - (\frac{0}{5})^2)] + [(\frac{5}{14})(1 - (\frac{2}{5})^2 - (\frac{3}{5})^2)] = .371$

$S_H = \{t1, t2, t3, t13\}$ $S_M = \{t4, t8, t10, t11, t12, t14\}$ $S_C = \{t5, t6, t7, t9\}$

$Gain(S, Temperature) = .714 - [(\frac{4}{14})(1 - (\frac{2}{4})^2 - (\frac{2}{4})^2)] + [(\frac{6}{14})(1 - (\frac{2}{6})^2 - (\frac{4}{6})^2)] + [(\frac{4}{14})(1 - (\frac{1}{4})^2 - (\frac{3}{4})^2)] = .225$

$S_H = \{t1, t2, t3, t4, t8, t12, t14\}$ $S_N = \{t5, t6, t7, t9, t10, t11, t13\}$ $S_L = \{\}$

$Gain(S, Humidity) = .714 - [(\frac{7}{14})(1 - (\frac{3}{7})^2 - (\frac{4}{7})^2)] + [(\frac{7}{14})(1 - (\frac{1}{7})^2 - (\frac{6}{7})^2)] + [0] = .346$

$S_S = \{t2, t6, t7, t11, t12, t14\}$ $S_W = \{t1, t3, t4, t5, t8, t9, t10, t13\}$

$Gain(S, Wind) = .714 - [(\frac{6}{14})(1 - (\frac{3}{6})^2 - (\frac{3}{6})^2)]S + [(\frac{8}{14})(1 - (\frac{2}{8})^2 - (\frac{6}{8})^2)] = 0$

Outlook has the largest information gain therefore we will split S with attribute Outlook.

Because all of the values in Overcast have the same label, a leaf node with label Yes will be returned.

Step 2: Find the attribute that best Splits S again for Sunny.

Find current Gini Index (GI):

$S = \{t1, t2, t8, t9, t11\}$

$GI(S) = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = .48$

Next, find which attributes has the largest information gain:

Current attributes: A = {Temperature, Humidity, Wind}

$S_H = \{t1, t2\}$ $S_M = \{t8, t11\}$ $S_C = \{t9\}$

$Gain(S, Temperature) = .48 - [0] + [(\frac{2}{5})(1 - (\frac{1}{2})^2 - (\frac{1}{2})^2)] + [0] = .28$

$S_H = \{t1, t2, t8\} \qquad S_N = \{t9, t11\} \qquad S_L = \{\}$
$Gain(S, Humidity) = .48 - [0] + [0] + [0] = .48$

$S_S = \{t2, t11\} \qquad S_W = \{t1, t8, t9\}$
$Gain(S, Wind) = .48 - [(\frac{2}{5})(1 - (\frac{1}{2})^2 - (\frac{1}{2})^2)] + [(\frac{3}{5})(1 - (\frac{1}{3})^2 - (\frac{2}{3})^2)] = .0133$

Humidity has the largest information gain therefore we will split S with attribute of Humidity

Because all of the values in S_H, S_N, and S_L (i.e., High, Normal, or Low Humidity) have the same label, a leaf node with label No, Yes, No will be returned respectively.

Step 3: Find the attribute that best Splits S again for Rain.

Find current Gini Index (GI):

$S = \{t4, t5, t6, t10, t14\}$

$GI(S) = 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = .48$

Next, find which attributes has the largest information gain:

Current attributes: A = {Temperature, Humidity, Wind}

$S_H = \{\} \qquad S_M = \{t4, t10, t14\} \qquad S_C = \{t5, t6\}$
$Gain(S, Temperature) = .48 - [0] + [(\frac{3}{5})(1 - (\frac{1}{3})^2 - (\frac{2}{3})^2)] + [0] = .213$

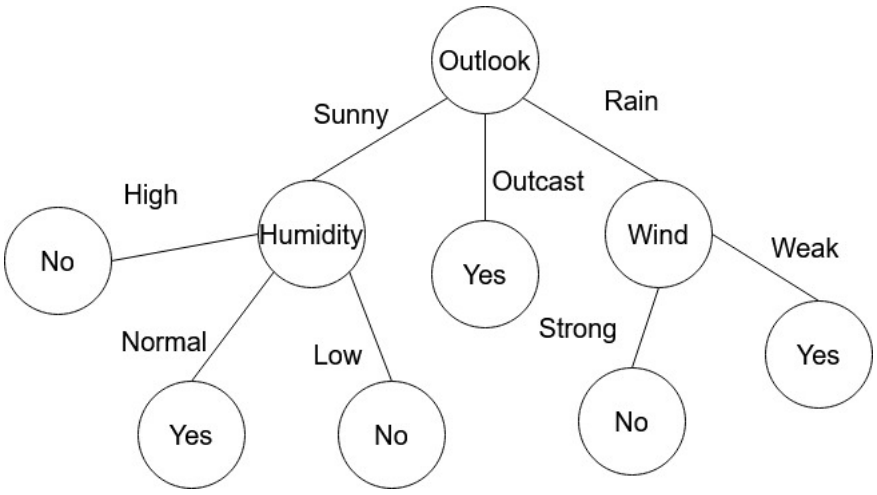
$S_H = \{t4, t14\} \qquad S_N = \{t5, t6, t10\} \qquad S_L = \{\}$
$Gain(S, Humidity) = .48 - [(\frac{2}{5})(1 - (\frac{1}{2})^2 - (\frac{1}{2})^2)] + [(\frac{3}{5})(1 - (\frac{1}{3})^2 - (\frac{2}{3})^2)] + [0]$
$= .0133$

$S_S = \{t4, t5, t10\} \qquad S_W = \{t6, t14\}$
$Gain(S, Wind) = .48 - [0] + [0] = .48$

Wind has the largest information gain therefore we will split S with attribute of Wind

Because all of the values in S_S and S_W (i.e., Strong and Weak Wind) have the same label, a leaf node with label No and Yes will be returned respectively.

The final decision tree is the same as the descision tree that was produced when using Marginal Error(ME) instead of Gini Index(GI):



(c) [3 points] Compare the two trees you just created with the one we built in the class (see Page 62 of the lecture slides). Are there any differences? Why?

There aren't any differences in the two trees because we are using an algorithm (ID3) to produce trees and the only way to get different trees is to have different data or to use different variants of determining information gain. In our case, even using different variants of information gain didn't produce a different tree.

3. [16 points] Continue with the same training data in Problem 2. Suppose before the tree construction, we receive one more training instance where Outlook's value is missing: {Outlook: Missing, Temperature: Mild, Humidity: Normal, Wind: Weak, Play: Yes}.

(a) [3 points] Use the most common value in the training data as the missing value, and calculate the information gains of the four features. Note that if there is a tie for the most common value, you can choose any value in the tie. Indicate the best feature.

$ME(S) = \frac{5}{15} = .333$

$Gain(S, Outlook) = .333 - [(\frac{6}{15})(\frac{3}{6}) + (\frac{4}{15})(\frac{0}{4}) + (\frac{5}{15})(\frac{2}{5})] = 0$

$Gain(S, Temperature) = .333 - [(\frac{4}{15})(\frac{2}{4}) + (\frac{7}{15})(\frac{2}{7}) + (\frac{4}{15})(\frac{1}{4})] = 0$

$Gain(S, Humidity) = .333 - [(\frac{7}{15})(\frac{3}{7}) + (\frac{8}{15})(\frac{1}{8}) + 0)] = .066$

$Gain(S, Wind) = .357 - [(\frac{6}{15})(\frac{3}{6}) + (\frac{9}{15})(\frac{2}{9}))] = 0$

(b) [3 points] Use the most common value among the training instances with the same label, namely, their attribute "Play" is "Yes", and calculate the information gains of the four features. Again if there is a tie, you can choose any value in the tie. Indicate the best feature.

The best feature is overcast as it is the most common value among the "yes" label. The gain for outlook only needs to be recalculated as all other attributes do not change from changing outlook from sunny to overcast.

$ME(S) = \frac{5}{15} = .333$
$Gain(S, Outlook) = .333 - [(\frac{5}{15})(\frac{2}{5}) + (\frac{5}{15})(\frac{0}{5}) + (\frac{5}{15})(\frac{2}{5})] = 0.066$

(c) [3 points] Use the fractional counts to infer the feature values, and then calculate the information gains of the four features. Indicate the best feature.-

Similarly, only the gain for Outlook needs to be recalculated.

Best feature: Sunny $\frac{5}{14} = .357$, $Overcast\frac{4}{14} = .285$, $Rain\frac{5}{14} = .357$

$ME(S) = \frac{5}{15} = .333$
$Gain(S, Outlook) = .333 - [(\frac{5.357}{15})(\frac{2.357}{5.357}) + (\frac{4.285}{15})(\frac{0}{4.285}) + (\frac{5.357}{15})(\frac{2}{5.357})] = .0428$

(d) [7 points] Continue with the fractional examples, and build the whole free with information gain. List every step and the final tree structure.

Step 1: Find the attribute that best Splits S.

From the the above problems we can see that Humidity has the largest information gain therefore we will split s with attribute Humidity.

Step 2: Find the attribute that best Splits S again for High.

Find current Majority Error(ME):

$S_H = \{t1, t2, t3, t4, t8, t12, t14\}$ ME(S) $= \frac{3}{7} = .428$

Next, find which attributes has the largest information gain:

Current attributes: A = {Outlook, Temperature, Wind}

$S_S = \{+0, -3\}$ $S_O = \{+2, -0\}$ $S_R = \{+1, -1\}$
$Gain(S, Outlook) = .428 - [(\frac{3}{7})(\frac{0}{3}) + (\frac{2}{7})(\frac{0}{2}) + (\frac{2}{7})(\frac{1}{2}))] = .285$

$S_H = \{+1, -2\}$ $S_M = \{+2, -2\}$ $S_C = \{+0, -0\}$
$Gain(S, Temperature) = .428 - [(\frac{3}{7})(\frac{1}{3}) + (\frac{4}{7})(\frac{2}{4}) + 0] = 0$

$S_S = \{+1, -2\}$ $S_W = \{+2, -2\}$
$Gain(S, Wind) = .428 - [(\frac{3}{7})(\frac{1}{3}) + (\frac{4}{7})(\frac{2}{4})] = 0$

Outlook has the largest information gain therefore we will split S_H with attribute of Outlook

Because all of the values in S_H and S_O (i.e., Sunny and Overcast) have the same label, a leaf node with label No, Yes, No will be returned respectively.

Step 3: Find the attribute that best Splits S_R again for Rain.

Find current Majority Error(ME):

$S = \{t4, t14\}$ ME(S) $= \frac{1}{2} = .5$

Next, find which attributes has the largest information gain:

Current attributes: A = {Temperature, Wind}

$S_H = \{+0, -0\}$ $S_M = \{+1, -1\}$ $S_C = \{+0, -0\}$

$Gain(S, Temperature) = .5 - [0 + (\frac{2}{2})(\frac{1}{2}) + 0)] = 0$

$S_S = \{+0, -1\}$ $S_W = \{+1, -0\}$
$Gain(S, Wind) = .5 - [(\frac{1}{2})(\frac{0}{2}) + (\frac{1}{2})(\frac{0}{2}))] = .5$

Wind has the largest information gain therefore we will split S_R with attribute of Wind

Because all of the values in S_S and S_W (i.e., Strong and Weak Wind) have the same label, a leaf node with label No and Yes will be returned respectively.

Step 4: Find the attribute that best Splits S again for Normal.

Find current Majority Error(ME):

$S_H = \{t5, t6, t7, t9, t10, t11, t13, t15\}$ ME(S) $= \frac{1}{8} = .125$

Next, find which attributes has the largest information gain:

Current attributes: A = {Outlook, Temperature, Wind}

$S_S = \{+2.357, -0\}$ $S_O = \{+2.285, -0\}$ $S_R = \{+2.357, -1\}$
$Gain(S, Outlook) = .125 - [0 + 0(\frac{3.357}{8})(\frac{1}{3.357}))] = 0$

$S_H = \{+1, -0\}$ $S_M = \{+3, -0\}$ $S_C = \{+3, -1\}$
$Gain(S, Temperature) = .125 - [0 + 0 + (\frac{4}{8})(\frac{1}{4})] = 0$

$S_S = \{+2, -1\}$ $S_W = \{+5, -0\}$
$Gain(S, Wind) = .125 - [(\frac{3}{8})(\frac{1}{3}) + 0] = .125$

None of the attributes produce any information gain therefore we will arbitrarily split S_N with attribute of Outlook

Because all of the values in S_S and S_O (i.e., Sunny and Overcast) have the same label, a leaf node with label Yes and Yes will be returned respectively.

Step 5: Find the attribute that best Splits S again for Rain.

Find current Majority Error(ME):

$S_R = \{t5, t6, t10, t15\}$ ME(S) $= \frac{1}{3.357} = .297$

Next, find which attributes has the largest information gain:

Current attributes: A = {Temperature, Wind}

$S_H = \{+0, -0\}$ $S_M = \{+1.357, -0\}$ $S_C = \{+1, -1\}$
$Gain(S, Temperature) = .297 - [0 + 0 + (\frac{2}{3.357})(\frac{1}{2})] = 0$

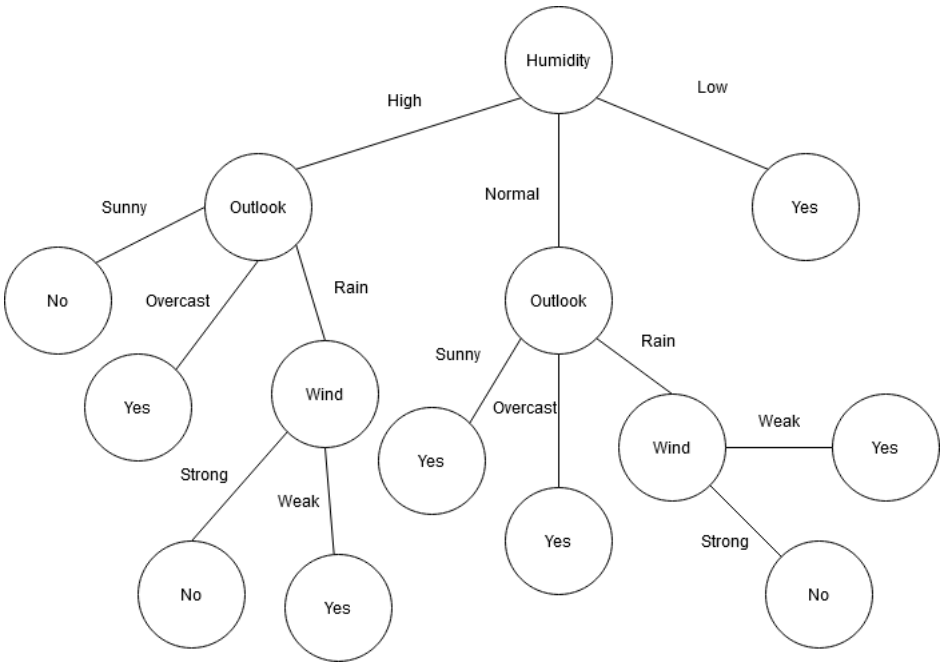
$S_S = \{+0, -1\}$ $S_W = \{+2.357, -0\}$
$Gain(S, Wind) = .297 - [0 + 0] = .297$

Wind has the largest information gain therefore we will split S_R with attribute of Wind

Because all of the values in S_S and S_W (i.e., Strong and Weak) have the same label, a leaf node with label No and Yes will be returned respectively.

Step 6: S_L (i.e., Low Humidity) is empty, therefore we will return a leaf node with the most common label in s which is Yes.

The final decision tree looks like:



4. [**Bonus question 1**] [5 points]. Prove that the information gain is always non-negative. That means, as long as we split the data, the purity will never get worse! (Hint: use convexity)

5. [**Bonus question 2**] [5 points]. We have discussed how to use decision tree for regression (i.e., predict numerical values) — on the leaf node, we simply use the average of the (numerical) labels as the prediction. Now, to construct a regression tree, can you invent a gain to select the best attribute to split data in ID3 framework?

2 Decision Tree Practice [60 points]

1. [5 Points] `https://github.com/SergioRemigio/DecisionTree`

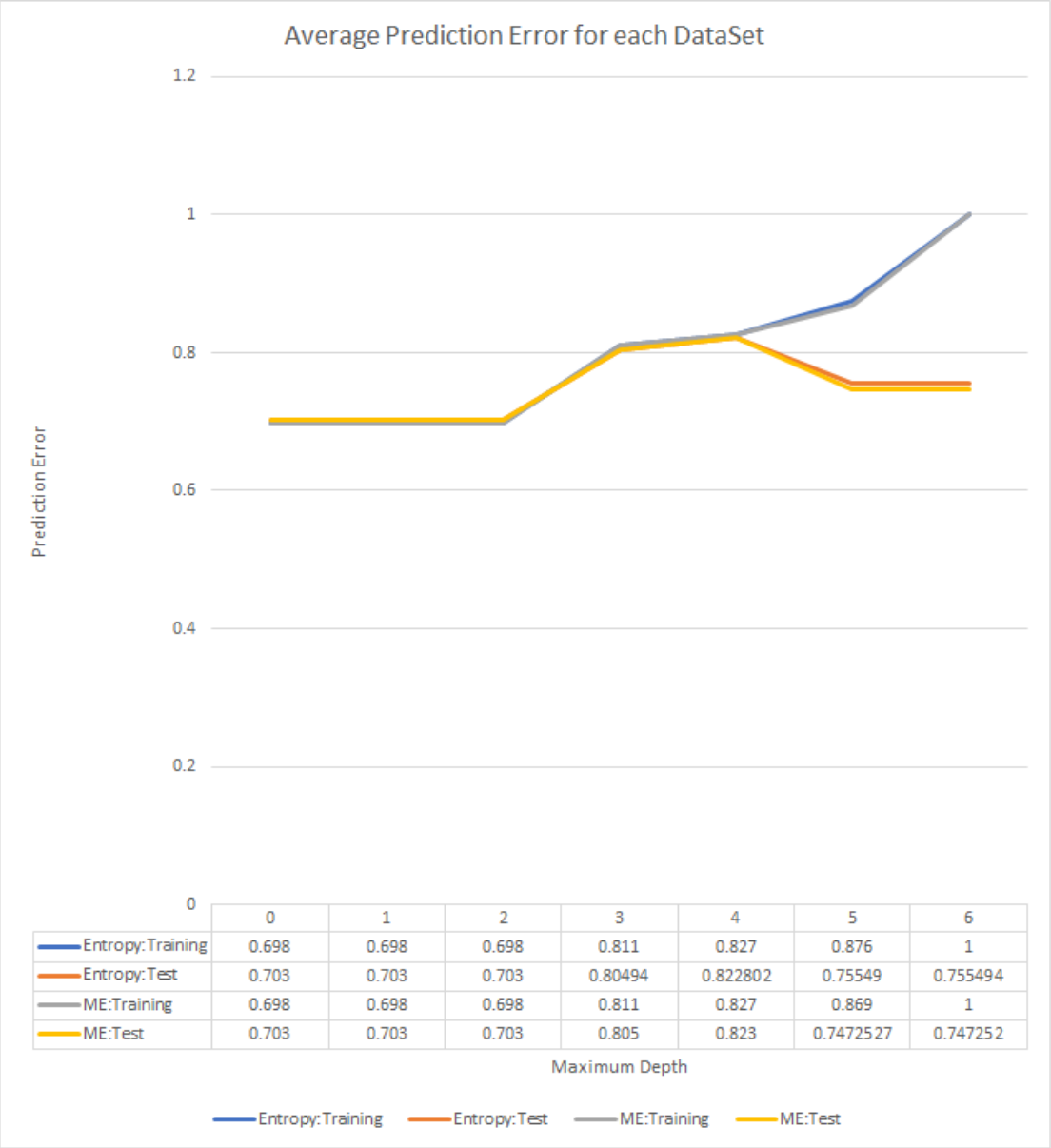
2. [30 points]

Note: we highly recommend you to use Python for implementation, because it is very convenient to load the data and handle strings. For example, the following snippet reads the CSV file line by line and split the values of the attributes and the label into a list, "terms". You can also use "dictionary" to store the categorical attribute values. In the web are numerous tutorials and examples for Python. if you have issues, just google it!

```
with open(CSVfile, 'r') as f:
    for line in f:
        terms = line.strip().split(',')
        process one training example
```

(a) [15 points] Implement the ID3 algorithm that supports, information gain, majority error and gini index to select attributes for data splits. Besides, your ID3 should allow users to set the maximum tree depth. Note: you do not need to convert categorical attributes into binary ones and your tree can be wide here.

(b) [10 points] Use your implemented algorithm to learn decision trees from the training data. Vary the maximum tree depth from 1 to 6 — for each setting, run your algorithm to learn a decision tree, and use the tree to predict both the training and test examples. Note that if your tree cannot grow up to 6 levels, then you can stop at the maximum level. Report in a table the average prediction errors on each dataset when you use information gain, majority error and gini index heuristics, respectively.



	0	1	2	3	4	5	6
Entropy:Training	0.698	0.698	0.698	0.811	0.827	0.876	1
Entropy:Test	0.703	0.703	0.703	0.80494	0.822802	0.75549	0.755494
ME:Training	0.698	0.698	0.698	0.811	0.827	0.869	1
ME:Test	0.703	0.703	0.703	0.805	0.823	0.7472527	0.747252

(c) [5 points] What can you conclude by comparing the training errors and the test errors?

We can conclude that the training data will always predict the correct label when it is allowed to reach maximum depth.

We can also conclude that the decision tree gives better predictions when it isn't allowed to reach maximum depth. However if it isn't allowed to have more than a certain depth, it will predict much worse than it potentially could. Therefore there is a optimal depth that a decision tree must have and testing all its depths is needed to find out the optimal depth.

.