CS 5350/6350: Machine Learining Fall 2023

Homework 1

Handed out: 3 Sep, 2024
Due date: 11:59pm, 20 Sep, 2024

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] Use the ID3 algorithm with information gain to learn a decision tree from the training dataset in Table 1. Please list every step in your tree construction, including the data subsets, the attributes, and how you calculate the information gain of each attribute and how you split the dataset according to the selected attribute. Please also give a full structure of the tree. You can manually draw the tree structure, convert the picture into a PDF/EPS/PNG/JPG format and include it in your homework submission; or instead, you can represent the tree with a conjunction of prediction rules as we discussed in the lecture.

   Let the 7 test items be called $y(1), y(2), ..., y(7)$. Let entropy be denoted H, gain be denoted G, current data subset be denoted S, and current set of attributes be denoted A. We start by calculating the gain of each attribute at the root node of the tree.

$$S = \{y(0), y(1), ..., y(7)\}$$

$$A = \{x_1, x_2, x_3, x_4\}$$

$$H(S) = -\frac{2}{7}log\left(\frac{2}{7}\right) - \frac{5}{7}log\left(\frac{5}{7}\right) = 0.86$$

$$H(S|x_1 = 0) = -\frac{1}{5}log\left(\frac{1}{5}\right) - \frac{4}{5}log\left(\frac{4}{5}\right) = 0.72$$

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

$$H(S|x_2 = 0) = -\frac{2}{3}log\left(\frac{2}{3}\right) - \frac{1}{3}log\left(\frac{1}{3}\right) = 0.92$$

$$H(S|x_2 = 1) = -0log\left(0\right) - 1log\left(1\right) = 0$$

$$H(S|x_3 = 0) = -\frac{1}{4}log\left(\frac{1}{4}\right) - \frac{3}{4}log\left(\frac{3}{4}\right) = 0.81$$

$$H(S|x_3 = 1) = -\frac{1}{3}log\left(\frac{1}{3}\right) - \frac{2}{3}log\left(\frac{2}{3}\right) = 0.92$$

$$H(S|x_4 = 0) = -0log\left(0\right) - 1log\left(1\right) = 0$$

$$H(S|x_4 = 1) = -\frac{2}{3}log\left(\frac{2}{3}\right) - \frac{1}{3}log\left(\frac{1}{3}\right) = 0.92$$

$$G(S, x_1) = 0.86 - \left(\frac{5}{7} \cdot 0.81 + \frac{2}{7} \cdot 1\right) = 0.06$$

$$G(S, x_2) = 0.86 - \left(\frac{3}{7} \cdot 0.92 + \frac{4}{7} \cdot 0\right) = 0.47$$

$$G(S, x_3) = 0.86 - \left(\frac{4}{7} \cdot 0.81 + \frac{3}{7} \cdot 0.92\right) = 0.001$$

$$G(S, x_4) = 0.86 - \left(\frac{4}{7} \cdot 0 + \frac{3}{7} \cdot 0.92\right) = 0.47$$

The information gain is highest for x_2 so we will use this to split the data set.

$$\{S|x_2 = 0\} : \{y(1), y(3), y(4)\}$$

$$\{S|x_2 = 1\} : \{y(2), y(5), y(6), y(7)\}$$

The set S where $x_2 = 1$ is uniquely labeled with $y = 0$ so that is a leaf node. Now we will split the other chunk by first finding the information gain.

$$S = \{y(1), y(3), y(4)\}$$

$$A = \{x_1, x_3, x_4\}$$

$$H(S) = -\frac{2}{3}log\left(\frac{2}{3}\right) - \frac{1}{3}log\left(\frac{1}{3}\right) = 0.92$$

$$H(S|x_1 = 0) = -\frac{1}{2}log\left(\frac{1}{2}\right) - \frac{1}{2}log\left(\frac{1}{2}\right) = 1$$

$$H(S|x_1 = 1) = -1log\left(1\right) - 0log\left(0\right) = 0$$

$$H(S|x_3 = 0) = -1log\left(1\right) - 0log\left(0\right) = 0$$

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

$$H(S|x_4 = 0) = -0log\left(0\right) - 1log\left(1\right) = 0$$

$$H(S|x_4 = 1) = -1log\left(1\right) - 0log\left(0\right) = 0$$

$$G(S, x_1) = 0.92 - \left(\frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 0\right) = 0.25$$

$$G(S, x_3) = 0.86 - \left(\frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 1\right) = 0.25$$

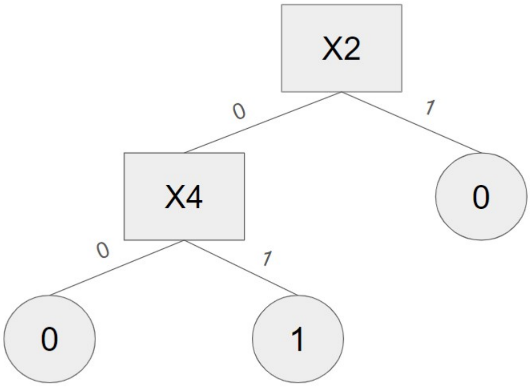
$$G(S, x_4) = 0.86 - \left(\frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 0\right) = 0.86$$

The information gain is highest for x_4 so we will use this to split the data set.

$$\{S|x_4 = 0\} : \{y(1)\}$$

$$\{S|x_4 = 1\} : \{y(3), y(4)\}$$

Both of these are uniquely labeled so we are done. Below is the final tree structure.



(b) [2 points] Write the boolean function which your decision tree represents. Please use a table to describe the function — the columns are the input variables and label, i.e., x_1, x_2, x_3, x_4 and y; the rows are different input and function values. **This boolean function is represented in Table 2.**

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	0	1	1	1
0	1	0	0.	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	1
1	0	1	0	0
1	0	1	1	1
1	1	0	0.	0
1	1	0	1	0
1	1	1	0	0
1	1	1	1	0

Table 2: Truth table for a Boolean classifier

2. [17 points] Let us use a training dataset to learn a decision tree about whether to play tennis (**Page 43, Lecture: Decision Tree Learning**, accessible by clicking the link http://www.cs.utah.edu/˜zhe/teach/pdf/3-decision-trees-learning.pdf). In the class, we have shown how to use information gain to construct the tree in ID3 framework.

   (a) [7 points] Now, please use majority error (ME) to calculate the gain, and select the best feature to split the data in ID3 framework. As in problem 1, please list every step in your tree construction, the attributes, how you calculate the gain of each attribute and how you split the dataset according to the selected attribute. Please also give a full structure of the tree.

   Let $y(1)$ to $y(14)$ be the 14 training data items. Let G be gain and ME be majority error. Let S be the data set and A be the current set of attributes. Note: majority error is easy to determine by simply looking at the data subset and counting items so I will not go into too much detail.

   $$S : \{y(1), y(2), ..., y(14)\}$$

   $$A : \{O, T, H, W\}$$

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

   $ME(S|O = s) = \frac{2}{5}$ $\quad ME(S|O = o) = 0$ $\quad ME(S|O = r) = \frac{2}{5}$

   $ME(S|T = h) = \frac{1}{2}$ $\quad ME(S|T = m) = \frac{1}{3}$ $\quad ME(S|T = c) = \frac{1}{4}$

$ME(S|H=h) = \frac{3}{7}$ $\qquad ME(S|H=n) = \frac{1}{7}$

$ME(S|W=s) = \frac{1}{2}$ $\qquad ME(S|W=w) = \frac{1}{4}$

$$G(S,O) = \frac{5}{14} - \left(\frac{5}{14}\cdot\frac{2}{5} + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot\frac{2}{5}\right) = 0.071$$

$$G(S,T) = \frac{5}{14} - \left(\frac{4}{14}\cdot\frac{1}{2} + \frac{6}{14}\cdot\frac{1}{3} + \frac{4}{14}\cdot\frac{1}{4}\right) = 0$$

$$G(S,H) = \frac{5}{14} - \left(\frac{7}{14}\cdot\frac{3}{7} + \frac{7}{14}\cdot\frac{1}{7}\right) = 0.071$$

$$G(S,W) = \frac{5}{14} - \left(\frac{6}{14}\cdot\frac{1}{2} + \frac{8}{14}\cdot\frac{1}{4}\right) = 0$$

We will split on O.

$$S_1 = \{S|O=s\} : \{y(1), y(2), y(8), y(9), y(11)\}$$

$$S_2 = \{S|O=o\} : \{y(3), y(7), y(12), y(13)\}$$

$$S_3 = \{S|O=r\} : \{y(4), y(5), y(6), y(10), y(14)\}$$

Notice that S_2 is uniquely labeled so that subset is already done. Let's split S_1 now.

$$S_1 : \{y(1), y(2), y(8), y(9), y(11)\}$$

$$A : \{T, H, W\}$$

$ME(S_1) = \frac{2}{5}$

$ME(S_1|T=h) = 0$ $\qquad ME(S_1|T=m) = \frac{1}{2}$ $\qquad ME(S_1|T=c) = 0$

$ME(S_1|H=h) = 0$ $\qquad ME(S_1|H=n) = 0$

$ME(S_1|W=s) = \frac{1}{2}$ $\qquad ME(S_1|W=w) = \frac{1}{3}$

$$G(S_1,T) = \frac{2}{5} - \left(\frac{2}{5}\cdot 0 + \frac{2}{5}\cdot\frac{1}{2} + \frac{1}{5}\cdot 0\right) = 0.2$$

$$G(S_1,H) = \frac{2}{5} - \left(\frac{3}{5}\cdot 0 + \frac{2}{5}\cdot 0\right) = 0.4$$

$$G(S_1,W) = \frac{2}{5} - \left(\frac{2}{5}\cdot\frac{1}{2} + \frac{3}{5}\cdot\frac{1}{3}\right) = 0$$

We will split S_1 on H.

$$\{S_1|H=h\}:\{y(1),y(2),y(8)\}$$

$$\{S_1|H=n\}:\{y(9),y(11)\}$$

$$\{S_1|H=l\}:\{\}$$

Each subset is uniquely labeled so this branch is done. Now let's split S_3.

$$S_3:\{y(4),y(5),y(6),y(10),y(14)\}$$

$$A:\{T,H,W\}$$

$ME(S_3)=\frac{2}{5}$

$ME(S_3|T=m)=\frac{1}{3}$ $\quad$ $ME(S_3|T=c)=\frac{1}{2}$

$ME(S_3|H=h)=\frac{1}{3}$ $\quad$ $ME(S_3|H=n)=\frac{1}{3}$

$ME(S_3|W=s)=0$ $\quad$ $ME(S_3|W=w)=0$

$$G(S_3,T)=\frac{2}{5}-\left(\frac{3}{5}\cdot\frac{1}{3}+\frac{2}{5}\cdot\frac{1}{2}\right)=0$$

$$G(S_3,H)=\frac{2}{5}-\left(\frac{2}{5}\cdot\frac{1}{3}\right)=0.067$$

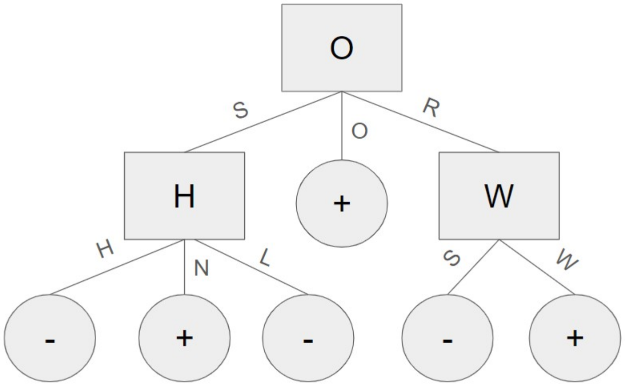
$$G(S_3,W)=\frac{2}{5}-\left(\frac{3}{5}\cdot 0+\frac{2}{5}\cdot 0\right)=0.4$$

We will split S_3 on W.

$$\{S_3|W=s\}:\{y(6),y(14)\}$$

$$\{S_3|W=w\}:\{y(4),y(5),y(10)\}$$

Each subset is uniquely labeled so this branch is done. The completed tree is shown below.



(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.

Let GI be the Gini Index.

$$S : \{y(1), y(2), ..., y(14)\}$$

$$A : \{O, T, H, W\}$$

$$GI(S) = 1 - \left(\frac{5}{14}^{2} + \frac{9}{14}^{2}\right) = 0.46$$

$$GI(S|O = s) = 1 - \left(\frac{2}{5}^{2} + \frac{3}{5}^{2}\right) = 0.48$$

$$GI(S|O = o) = 1 - \left(1^2 + 0^2\right) = 0$$

$$GI(S|O = r) = 1 - \left(\frac{3}{5}^{2} + \frac{2}{5}^{2}\right) = 0.48$$

$$GI(S|T = h) = 1 - \left(\frac{1}{2}^{2} + \frac{1}{2}^{2}\right) = 0.5$$

$$GI(S|T = m) = 1 - \left(\frac{2}{3}^{2} + \frac{1}{3}^{2}\right) = 0.44$$

$$GI(S|T = c) = 1 - \left(\frac{3}{4}^{2} + \frac{1}{4}^{2}\right) = 0.38$$

$$GI(S|H = h) = 1 - \left(\frac{3}{7}^{2} + \frac{4}{7}^{2}\right) = 0.49$$

$$GI(S|H = n) = 1 - \left(\frac{6}{7}^{2} + \frac{1}{7}^{2}\right) = 0.24$$

$$GI(S|W = s) = 1 - \left(\frac{1}{2}^{2} + \frac{1}{2}^{2}\right) = 0.5$$

$$GI(S|W = w) = 1 - \left(\frac{3}{4}^{2} + \frac{1}{4}^{2}\right) = 0.38$$

$$G(S, O) = 0.46 - \left(\frac{5}{14} \cdot 0.48 + \frac{4}{14} \cdot 0 + \frac{5}{14} \cdot 0.48\right) = 0.12$$

$$G(S, T) = 0.46 - \left(\frac{4}{14} \cdot 0.5 + \frac{6}{14} \cdot 0.44 + \frac{4}{14} \cdot 0.38\right) = 0.02$$

$$G(S, H) = 0.46 - \left(\frac{7}{14} \cdot 0.49 + \frac{7}{14} \cdot 0.24\right) = 0.095$$

$$G(S, W) = 0.46 - \left(\frac{6}{14} \cdot 0.5 + \frac{8}{14} \cdot 0.38\right) = 0.03$$

We will split on O.

$$S_1 = \{S|O = s\} : \{y(1), y(2), y(8), y(9), y(11)\}$$

$$S_2 = \{S|O = o\} : \{y(3), y(7), y(12), y(13)\}$$

$$S_3 = \{S|O = r\} : \{y(4), y(5), y(6), y(10), y(14)\}$$

Notice that S_2 is uniquely labeled so that subset is already done. Let's split S_1 now.

$$S_1 : \{y(1), y(2), y(8), y(9), y(11)\}$$

$$A : \{T, H, W\}$$

$$GI(S_1) = 1 - \left(\frac{2}{5}^2 + \frac{3}{5}^2\right) = 0.48$$

$$GI(S_1|T = h) = 1 - \left(0^2 + 1^2\right) = 0$$

$$GI(S_1|T = m) = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = 0.5$$

$$GI(S_1|T = c) = 1 - \left(1^2 + 0^2\right) = 0$$

$$GI(S_1|H = h) = 1 - \left(0^2 + 1^2\right) = 0$$

$$GI(S_1|H = n) = 1 - \left(1^2 + 0^2\right) = 0$$

$$GI(S_1|W = s) = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = 0.5$$

$$GI(S_1|W = w) = 1 - \left(\frac{1}{3}^2 + \frac{2}{3}^2\right) = 0.44$$

$$G(S_1, T) = 0.48 - \left(\frac{2}{5} \cdot 0 + \frac{2}{5} \cdot 0.5 + \frac{1}{5} \cdot 0\right) = 0.28$$

$$G(S_1, H) = 0.48 - \left(\frac{3}{5} \cdot 0 + \frac{2}{5} \cdot 0\right) = 0.48$$

$$G(S_1, W) = 0.48 - \left(\frac{2}{5} \cdot 0.5 + \frac{3}{5} \cdot 0.44\right) = 0.016$$

We will split S_1 on H.

$$\{S_1|H = h\} : \{y(1), y(2), y(8)\}$$

$$\{S_1|H = n\} : \{y(9), y(11)\}$$

$$\{S_1|H = l\} : \{\}$$

Each subset is uniquely labeled so this branch is done. Now let's split S_3.

$$S_3 : \{y(4), y(5), y(6), y(10), y(14)\}$$

$$A : \{T, H, W\}$$

$$GI(S_3) = 1 - \left(\frac{3}{5}^2 + \frac{2}{5}^2\right) = 0.48$$

$$GI(S_3|T = h) = 1 - \left(\frac{2}{3}^2 + \frac{1}{3}^2\right) = 0.44$$

$$GI(S_3|T = m) = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = 0.5$$

$$GI(S_3|H = h) = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = 0.5$$

$$GI(S_3|H = n) = 1 - \left(\frac{2}{3}^2 + \frac{1}{3}^2\right) = 0.44$$

$$GI(S_3|W = s) = 1 - \left(0^2 + 1^2\right) = 0$$

$$GI(S_3|W = w) = 1 - \left(1^2 + 0^2\right) = 0$$

$$G(S_3, T) = 0.48 - \left(\frac{3}{5} \cdot 0.44 + \frac{2}{5} \cdot 0.5\right) = 0.016$$

$$G(S_3, H) = 0.48 - \left(\frac{2}{5} \cdot 0.5 + \frac{3}{5} \cdot 0.44\right) = 0.016$$

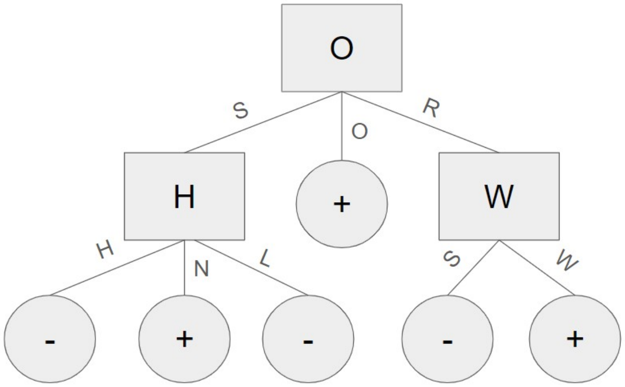
$$G(S_3, W) = 0.48 - \left(\frac{2}{5} \cdot 0 + \frac{3}{5} \cdot 0\right) = 0.48$$

We will split S_3 on W.

$$\{S_3|W = s\} : \{y(6), y(14)\}$$

$$\{S_3|W = w\} : \{y(4), y(5), y(10)\}$$

Each subset is uniquely labeled so this branch is done. The completed tree is shown below.



(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?

All 3 threes are identical. This is because there is a clear choice at each step so the differences in the gain calculation don't matter.

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.

We will choose sunny as the missing value because it is the most common.

$$H(S) = -\frac{10}{15} log\left(\frac{10}{15}\right) - \frac{5}{15} log\left(\frac{5}{15}\right) = 0.918$$

$$H(S|O = s) = -\frac{1}{2} log\left(\frac{1}{2}\right) - \frac{1}{2} log\left(\frac{1}{2}\right) = 1$$
$$H(S|O = o) = -1 log\left(1\right) - 0 log\left(0\right) = 0$$
$$H(S|O = r) = -\frac{3}{5} log\left(\frac{3}{5}\right) - \frac{2}{5} log\left(\frac{2}{5}\right) = 0.97$$
$$H(S|T = h) = -\frac{1}{2} log\left(\frac{1}{2}\right) - \frac{1}{2} log\left(\frac{1}{2}\right) = 1$$
$$H(S|T = m) = -\frac{5}{7} log\left(\frac{5}{7}\right) - \frac{2}{7} log\left(\frac{2}{7}\right) = 0.86$$
$$H(S|T = c) = -\frac{3}{4} log\left(\frac{3}{4}\right) - \frac{1}{4} log\left(\frac{1}{4}\right) = 0.81$$
$$H(S|H = h) = -\frac{3}{7} log\left(\frac{3}{7}\right) - \frac{4}{7} log\left(\frac{4}{7}\right) = 0.99$$
$$H(S|H = n) = -\frac{7}{8} log\left(\frac{7}{8}\right) - \frac{1}{8} log\left(\frac{1}{8}\right) = 0.54$$
$$H(S|W = s) = -\frac{1}{2} log\left(\frac{1}{2}\right) - \frac{1}{2} log\left(\frac{1}{2}\right) = 1$$
$$H(S|W = w) = -\frac{7}{9} log\left(\frac{7}{9}\right) - \frac{2}{9} log\left(\frac{2}{9}\right) = 0.76$$

$$G(S, O) = 0.918 - \left(\frac{6}{15} \cdot 1 + \frac{4}{15} \cdot 0 + \frac{5}{15} \cdot 0.97\right) = 0.195$$

$$G(S,H) = 0.918 - \left(\frac{4}{15}\cdot 1 + \frac{7}{15}\cdot 0.86 + \frac{4}{15}\cdot 0.81\right) = 0.034$$

$$G(S,T) = 0.918 - \left(\frac{7}{15}\cdot 0.99 + \frac{8}{15}\cdot 0.54\right) = 0.168$$

$$G(S,W) = 0.918 - \left(\frac{6}{15}\cdot 1 + \frac{9}{15}\cdot 0.76\right) = 0.062$$

Outlook is still the best attribute.

(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.

This time, Overcast will be the missing value because it is the most common among positive examples. This change only affects the Entropy and Gain for Outlook. The rest are the same as Part A.

$$H(S|O=s) = -\frac{2}{5}log\left(\frac{2}{5}\right) - \frac{3}{5}log\left(\frac{3}{5}\right) = 0.97$$

$$H(S|O=o) = -1log\,(1) - 0log\,(0) = 0$$

$$G(S,O) = 0.918 - \left(\frac{5}{15}\cdot 0.97 + \frac{5}{15}\cdot 0 + \frac{5}{15}\cdot 0.97\right) = 0.27$$

Outlook is still the best attribute.

(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.

Using fractional counts is still very similar to Part A with the only changes being the Entropy and Gain for Outlook.

$$H(S|O=s) = -\frac{2.357}{5.357}log\left(\frac{2.357}{5.357}\right) - \frac{3}{5.357}log\left(\frac{3}{5.357}\right) = 0.99$$

$$H(S|O=o) = -1log\,(1) - 0log\,(0) = 0$$

$$H(S|O=r) = -\frac{3.357}{5.357}log\left(\frac{3.357}{5.357}\right) - \frac{2.357}{5.357}log\left(\frac{2.357}{5.357}\right) = 0.95$$

$$G(S,O) = 0.918 - \left(\frac{5.357}{15}\cdot 0.99 + \frac{4.357}{15}\cdot 0 + \frac{5.357}{15}\cdot 0.95\right) = 0.26$$

Outlook is still the best attribute.

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

We will split on O. Let F be a fractional example containing $\{T = m, H = n, W = w, Label = +\}$.

$$S_1 = \{S|O = s\} : \left\{y(1), y(2), y(8), y(9), y(11), \frac{5}{14} \cdot F\right\}$$

$$S_2 = \{S|O = o\} : \left\{y(3), y(7), y(12), y(13), \frac{4}{14} \cdot F\right\}$$

$$S_3 = \{S|O = r\} : \left\{y(4), y(5), y(6), y(10), y(14), \frac{5}{14} \cdot F\right\}$$

We will now split S_1.

$$A : \{T, H, W\}$$

$$H(S_1) = -\frac{2.357}{5.357} log\left(\frac{2.357}{5.357}\right) - \frac{3}{5.357} log\left(\frac{3}{5.357}\right) = 0.99$$

$$H(S_1|T = h) = -0log\left(0\right) - 1log\left(1\right) = 0$$

$$H(S_1|T = m) = -\frac{1.357}{2.357} log\left(\frac{1.357}{2.357}\right) - \frac{1}{2.357} log\left(\frac{1}{2.357}\right) = 0.98$$

$$H(S_1|T = c) = -1log\left(1\right) - 0log\left(0\right) = 0$$

$$H(S_1|H = h) = -0log\left(0\right) - 1log\left(1\right) = 0$$

$$H(S_1|H = n) = -1log\left(1\right) - 0log\left(0\right) = 0$$

$$H(S_1|W = s) = -\frac{1}{2} log\left(\frac{1}{2}\right) - \frac{1}{2} log\left(\frac{1}{2}\right) = 1$$

$$H(S_1|W = w) = -\frac{1.357}{3.357} log\left(\frac{1.357}{3.357}\right) - \frac{2}{3.357} log\left(\frac{2}{3.357}\right) = 0.97$$

$$G(S_1, T) = 0.99 - \left(\frac{2}{5.357} \cdot 0 + \frac{2.357}{5.357} \cdot 0.98 + \frac{1}{5.357} \cdot 0\right) = 0.55$$

$$G(S_1, H) = 0.99 - \left(\frac{3}{5.357} \cdot 0 + \frac{2}{5.357} \cdot 0\right) = 0.99$$

$$G(S_1, W) = 0.99 - \left(\frac{2}{5.357} \cdot 1 + \frac{3.357}{5.357} \cdot 0.97\right) = 0.01$$

We will split S_1 on H.

$$\{S_1|H = h\} : \{y(1), y(2), y(8)\}$$

$$\{S_1|H = n\} : \{y(9), y(11), \frac{5}{14} \cdot F\}$$

$$\{S_1 | H = l\} : \{\}$$

Each subset is uniquely labeled so this branch is done. S_2 is also uniquely labeled so that branch is done too. Now let's split S_3.

$$A : \{T, H, W\}$$

$$H(S_3) = -\frac{3.357}{5.357}log\left(\frac{3.357}{5.357}\right) - \frac{2}{5.357}log\left(\frac{2}{5.357}\right) = 0.95$$

$$H(S_3|T = m) = -\frac{2.357}{3.357}log\left(\frac{2.357}{3.357}\right) - \frac{1}{3.357}log\left(\frac{1}{3.357}\right) = 0.88$$

$$H(S_3|T = c) = -\frac{1}{2}log\left(\frac{1}{2}\right) - \frac{1}{2}log\left(\frac{1}{2}\right) = 1$$

$$H(S_3|H = h) = -\frac{1}{2}log\left(\frac{1}{2}\right) - \frac{1}{2}log\left(\frac{1}{2}\right) = 1$$

$$H(S_3|H = n) = -\frac{2.357}{3.357}log\left(\frac{2.357}{3.357}\right) - \frac{1}{3.357}log\left(\frac{1}{3.357}\right) = 0.88$$

$$H(S_3|W = s) = -0log\left(0\right) - 1log\left(1\right) = 0$$

$$H(S_3|W = w) = -1log\left(1\right) - 0log\left(0\right) = 0$$

$$G(S_3, T) = 0.95 - \left(\frac{3.357}{5.357} \cdot 0.88 + \frac{2}{5.357} \cdot 1\right) = 0.028$$

$$G(S_3, H) = 0.95 - \left(\frac{2}{5.357} \cdot 1 + \frac{3.357}{5.357} \cdot 0.88\right) = 0.028$$

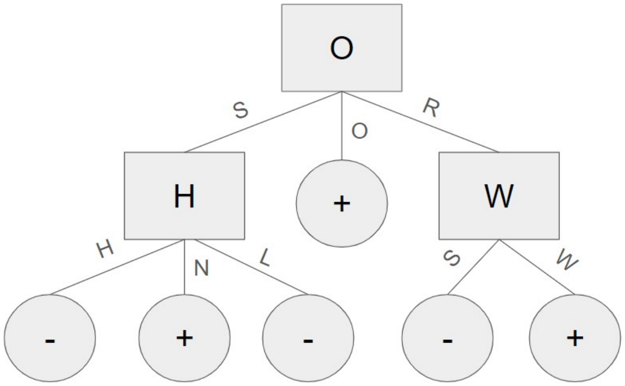
$$G(S_3, W) = 0.95 - \left(\frac{2}{5.357} \cdot 0 + \frac{3.357}{5.357} \cdot 0\right) = 0.95$$

We will split S_3 on W.

$$\{S_3 | W = s\} : \{y(6), y(14)\}$$

$$\{S_3 | W = w\} : \{y(4), y(5), y(10), \frac{5}{14} \cdot F\}$$

These subsets are uniquely labeled which means the tree is done, It is identical to the other 3 times that we have calculated this tree.



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)

   Skip

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?

   You want to select an attribute to split on that decreases the standard deviation of the labels in the subset of data. The gain calculation should find the standard deviation of the full subset and subtract the average of the standard deviations of the divided subsets, similarly to what we do for Entropy, Majority Error, and Gini Index.

2 Decision Tree Practice [60 points]

1. [5 Points] Starting from this assignment, we will build a light-weighted machine learning library. To this end, you will first need to create a code repository in Github.com. Please refer to the short introduction in the appendix and the official tutorial to create an account and repository. Please commit a README.md file in your repository, and write one sentence: "This is a machine learning library developed by **Your Name** for CS5350/6350 in University of Utah". You can now create a first folder, "DecisionTree". Please leave the link to your repository in the homework submission. We will check if you have successfully created it.

   Link to my repo: https://github.com/EpicMark98/CS6350MachineLearning

2. [30 points] We will implement a decision tree learning algorithm for car evaluation task. The dataset is from UCI repository(`https://archive.ics.uci.edu/ml/datasets/car+evaluation`). Please download the processed dataset (car.zip) from Canvas. In this task, we have 6 car attributes, and the label is the evaluation of the car. The attribute and label values are listed in the file "data-desc.txt". All the attributes are categorical. The training data are stored in the file "train.csv", consisting of $1,000$ examples. The test data are stored in "test.csv", and comprise 728 examples. In both training and test datasets, attribute values are separated by commas; the file "data-desc.txt" lists the attribute names in each column.

   (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.

   See implementation on GitHub.

(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, 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.

Tree Depth	Entropy	Majority Error	Gini Index
1	30%	30%	30%
2	22%	30%	22%
3	18%	24%	18%
4	8%	13%	9%
5	3%	4%	3%
6	0%	0%	0%

Table 3: Average training error of each heuristic for various tree depths

Tree Depth	Entropy	Majority Error	Gini Index
1	30%	30%	30%
2	22%	32%	22%
3	20%	26%	18%
4	15%	24%	13%
5	9%	17%	9%
6	9%	17%	9%

Table 4: Average test error of each heuristic for various tree depths

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

As you increase the tree depth, both the training error and the test error decrease until depth 5 when the test error stops decreasing. In many trees, increasing the depth too much will overfit the data and cause the test error to start increasing.

3. [25 points] Next, modify your implementation a little bit to support numerical attributes. We will use a simple approach to convert a numerical feature to a binary one. We choose the media (NOT the average) of the attribute values (in the training set) as the threshold, and examine if the feature is bigger (or less) than the threshold. We will use another real dataset from UCI repository(`https://archive.ics.uci.edu/ml/datasets/Bank+Marketing`). This dataset contains 16 attributes, including both numerical and categorical ones. Please download the processed dataset from Canvas (bank.zip). The attribute and label values are listed in the file "data-desc.txt". The training set is the file "train.csv", consisting of 5,000 examples, and the test "test.csv"

with 5,000 examples as well. In both training and test datasets, attribute values are separated by commas; the file “data-desc.txt” lists the attribute names in each column.

(a) [10 points] Let us consider “unknown” as a particular attribute value, and hence we do not have any missing attributes for both training and test. Vary the maximum tree depth from 1 to 16 — for each setting, run your algorithm to learn a decision tree, and use the tree to predict both the training and test examples. Again, if your tree cannot grow up to 16 levels, 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.

See Tables 5 and 6 for results

Tree Depth	Entropy	Majority Error	Gini Index
1	12%	11%	11%
2	11%	10%	10%
3	10%	10%	9%
4	8%	8%	7%
5	6%	7%	6%
6	5%	7%	5%
7	3%	6%	3%
8	3%	6%	3%
9	2%	5%	2%
10	2%	4%	2%
11	1%	4%	1%
12	1%	3%	1%
13	1%	3%	1%
14	1%	2%	1%
15	1%	2%	1%
16	1%	1%	1%

Table 5: Average training error of each heuristic for various tree depths

(b) [10 points] Let us consider ”unknown” as attribute value missing. Here we simply complete it with the majority of other values of the same attribute in the training set. Vary the maximum tree depth from 1 to 16 — for each setting, run your algorithm to learn a decision tree, and use the tree to predict both the training and test examples. Report in a table the average prediction errors on each dataset when you use information gain, majority error and gini index heuristics, respectively.

See Tables 7 and 8 for results

(c) [5 points] What can you conclude by comparing the training errors and the test errors, with different tree depths, as well as different ways to deal with ”unknown” attribute values?

At first the training and test errors decrease together as depth increases. Soon, the test error starts to increase while the training error

Tree Depth	Entropy	Majority Error	Gini Index
1	12%	12%	12%
2	11%	11%	11%
3	11%	11%	11%
4	11%	11%	12%
5	12%	11%	13%
6	13%	12%	15%
7	14%	12%	15%
8	15%	12%	16%
9	15%	13%	17%
10	15%	13%	17%
11	16%	14%	17%
12	16%	14%	17%
13	16%	15%	17%
14	16%	15%	17%
15	16%	16%	17%
16	16%	16%	17%

Table 6: Average test error of each heuristic for various tree depths

continues to decrease. I'm not sure why my training error did not reach 0 at a depth of 16 (maybe there is a bug). Replacing the "unknown" values with the most common value seems to make the error very slightly worse across the board. A depth of 3 seems to have the best test error in both cases.

Tree Depth	Entropy	Majority Error	Gini Index
1	12%	11%	11%
2	11%	10%	10%
3	10%	10%	9%
4	8%	8%	8%
5	6%	7%	6%
6	5%	7%	5%
7	4%	6%	3%
8	3%	6%	3%
9	3%	5%	2%
10	2%	5%	2%
11	2%	4%	2%
12	2%	4%	2%
13	2%	3%	2%
14	2%	3%	2%
15	2%	2%	2%
16	2%	2%	2%

Table 7: Average training error of each heuristic for various tree depths replacing unknown values

Tree Depth	Entropy	Majority Error	Gini Index
1	12%	11%	12%
2	11%	11%	11%
3	11%	11%	11%
4	12%	11%	12%
5	12%	12%	13%
6	14%	12%	15%
7	15%	12%	15%
8	15%	13%	16%
9	16%	13%	17%
10	16%	14%	17%
11	16%	15%	17%
12	17%	15%	17%
13	17%	15%	17%
14	17%	16%	17%
15	17%	16%	17%
16	17%	16%	17%

Table 8: Average test error of each heuristic for various tree depths replacing unknown values