Machine Learning

Author: Alec Peterson (ap3842@drexel.edu)

Assignment 5 - Decision Trees

Fall 2023

1 Theory

1. Consider the following set of training examples for an unknown target function: $(x_1, x_2) \rightarrow y$:

Y	x_1	x_2	Count
+	T	T	3
+	T	F	4
+	F	T	4
+	F	F	1
-	T	T	0
-	T	F	1
-	F	T	3
-	F	F	5

(a) What is the sample entropy for the class label overall, $H(Y)$ from this training data (using log base 2) (3pts)?

Reported values are rounded to nearest thousandth.

i. $P_{Y=+} = \frac{3+4+4+1}{3+4+4+1+0+1+3+5} = \frac{12}{21}$

ii. $P_{Y=-} = 1 - P_{Y=+} = 1 - \frac{12}{21} = \frac{9}{21}$

iii. $H(Y) = -P_{Y=+} log_2(P_{Y=+}) - -P_{Y=-} log_2(P_{Y=-})$

$$\implies H(Y) = -(\tfrac{12}{21}) log_2(\tfrac{12}{21}) - (\tfrac{9}{21}) log_2(\tfrac{9}{21}) = 0.985$$

(b) What are the weighed average entropies for branching on variables x_1 and x_2 (6pts)?

Reported values are rounded to nearest thousandth.
Calculations performed / propagated with un-rounded values.

i. $x_1 = T$:

- $P(y=+) = \frac{3+4}{3+4+0+1} = \frac{7}{8}$
- $P(y=-) = 1 - P(y=+) = 1 - \frac{7}{8} = \frac{1}{8}$

$$\implies H_{x1=T} = -\frac{7}{8}\log_2(\frac{7}{8}) - \frac{1}{8}\log_2(\frac{1}{8}) = 0.544$$

ii. $x_1 = F$:

- $P(y=+) = \frac{4+1}{4+1+3+5} = \frac{5}{13}$
- $P(y=-) = 1 - P(y=+) = 1 - \frac{5}{13} = \frac{8}{13}$

$$\implies H_{x1=F} = -\frac{5}{13}\log_2(\frac{5}{13}) - \frac{8}{13}\log_2(\frac{8}{13}) = 0.961$$

iii. $x_2 = T$:

- $P(y=+) = \frac{3+4}{3+4+0+3} = \frac{7}{10}$
- $P(y=-) = 1 - P(y=+) = 1 - \frac{7}{10} = \frac{3}{10}$

$$\implies H_{x2=T} = -\frac{7}{10}\log_2(\frac{7}{10}) - \frac{3}{10}\log_2(\frac{3}{10}) = 0.881$$

iv. $x_2 = F$:

- $P(y=+) = \frac{4+1}{4+1+1+5} = \frac{5}{11}$
- $P(y=-) = 1 - P(y=+) = 1 - \frac{5}{11} = \frac{6}{11}$

$$\implies H_{x2=F} = -\frac{5}{11}\log_2(\frac{5}{11}) - \frac{6}{11}\log_2(\frac{6}{11}) = 0.994$$

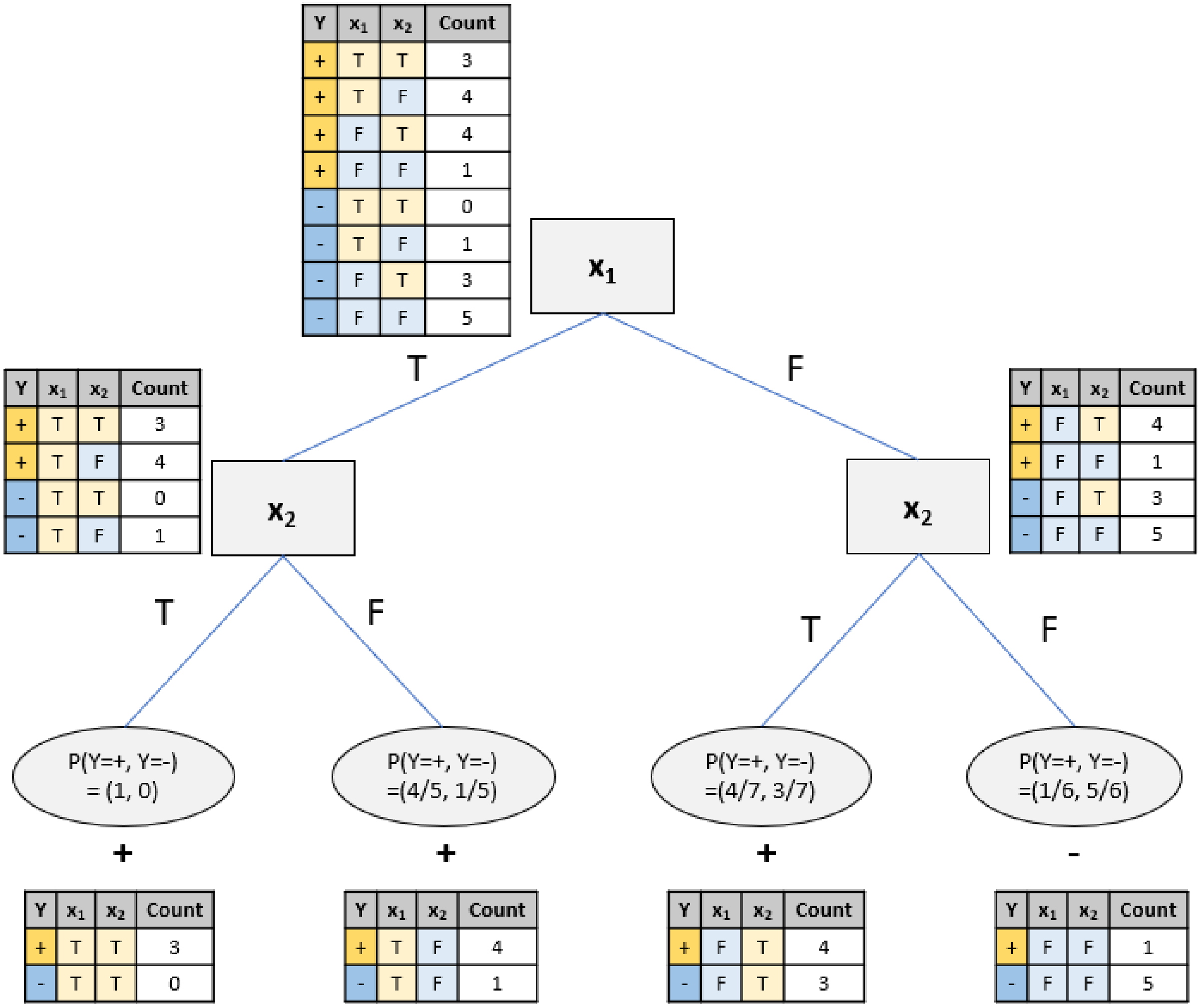
v. Weight-averaged entropy Calculations:

$$TotalCount = 3+4+4+1+0+1+3+5 = 21$$

$$E_{x1} = \frac{3+4+0+1}{21}H_{x1=T} + \frac{4+1+3+5}{21}H_{x1=F} = \frac{8}{21}(0.544) + \frac{13}{21}(0.961) = \mathbf{0.802}$$

$$E_{x2} = \frac{3+4+0+3}{21}H_{x2=T} + \frac{4+1+1+5}{21}H_{x2=F} = \frac{10}{21}(0.881) + \frac{11}{21}(0.994) = \mathbf{0.951}$$

(c) Draw the decision tree that would be learned by the ID3 algorithm without pruning from this training data. You may use software to draw this or draw it by hand. But either way the figure should be embedded in your PDF submission. (6pts)



Y	x_1	x_2	Count
+	T	T	3
+	T	F	4
+	F	T	4
+	F	F	1
-	T	T	0
-	T	F	1
-	F	T	3
-	F	F	5

Y	x_1	x_2	Count
+	T	T	3
+	T	F	4
-	T	T	0
-	T	F	1

Y	x_1	x_2	Count
+	F	T	4
+	F	F	1
-	F	T	3
-	F	F	5

Y	x_1	x_2	Count
+	T	T	3
-	T	T	0

Y	x_1	x_2	Count
+	T	F	4
-	T	F	1

Y	x_1	x_2	Count
+	F	T	4
-	F	T	3

Y	x_1	x_2	Count
+	F	F	1
-	F	F	5

2 Decision Tree

In your report you will need:

1. Description of any additional pre-processing of the dataset you did.

   In training set, continuous columns in dataset were all made into binary categorical columns by comparing against respective mean of each column. If value in column was less than mean, made to 0, else made to 1.

   In validation set, continous columns were all similarly made into binary categorical columns by comparing against respective mean of each column from training set. If value in column was less than training mean, made to 0, else made to 1.

2. The validation accuracy of your system.

   Accuracy of validation set = 0.86

3. Your confusion matrix.

   $$\textbf{Confusion Matrix} = \begin{bmatrix} 510 & 26 & 19 \\ 31 & 56 & 4 \\ 6 & 12 & 44 \end{bmatrix}$$

3 Additional Dataset

(a) Description of any additional pre-processing of the dataset you did.

In training set, continuous columns in dataset were all made into binary categorical columns by comparing against respective mean of each column. If value in column was less than mean, made to 0, else made to 1.

In validation set, continous columns were all similarly made into binary categorical columns by comparing against respective mean of each column from training set. If value in column was less than training mean, made to 0, else made to 1.

(b) Validation accuracies and confusion matrices

Accuracy of validation set = 0.80

Confusion Matrix for training set =

$$\begin{bmatrix}
7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7
\end{bmatrix}$$

(see next page)

Confusion Matrix for validation set =

$$\begin{bmatrix}
4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 2
\end{bmatrix}$$