CS 5350/6350: Machine Learining Fall 2017

Yucheng Yang
Homework 1

Handed out: 29 August, 2017
Due date: 12 September, 2017

1 Decision Tree

1. [6 points] Write the following Boolean functions as decision trees. (You can write your decision trees as a series of if-then-else statements, or use your favorite drawing program to draw a tree. You can use 1 to represent True and 0 to represent False.)

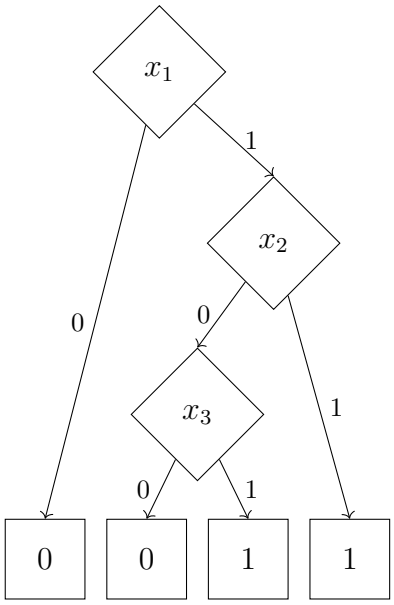
   (a) $(x_1 \wedge x_2) \vee (x_1 \wedge x_3)$

   (b) $(x_1 \wedge x_2)$ xor x_3

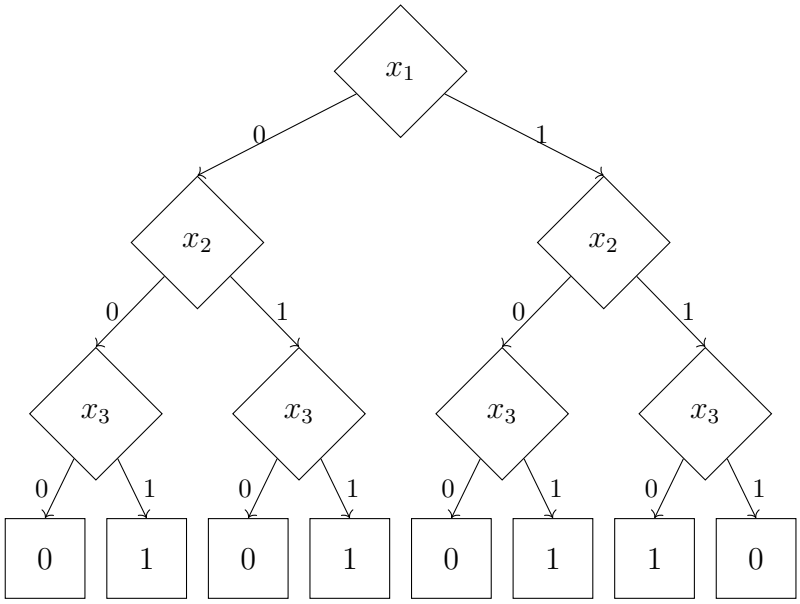
   (c) $\neg A \vee \neg B \vee \neg C \vee \neg D$

   Solution:

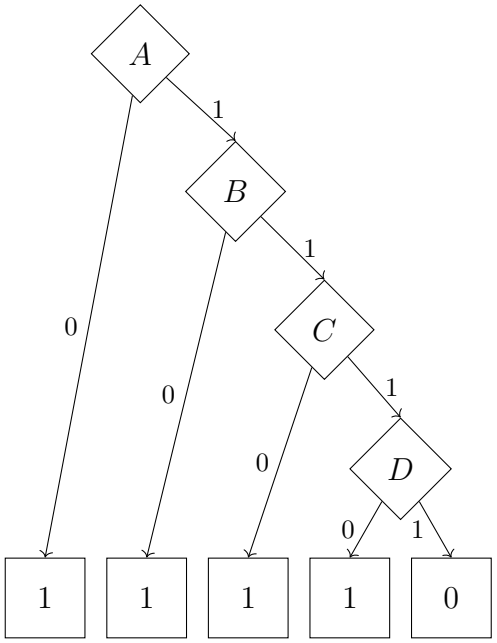
   (a)



(b)



(c)



2. [24 points] Solution:

   (a) $2 * 2 * 3 * 4 = 64$
   There is only 9 functions consistent with the given training dataset.

   (b) $H(Invade?) = -\frac{5}{9} * \log_2(\frac{5}{9}) - \frac{4}{9} * \log_2(\frac{4}{9}) = 0.47111 + 0.52 = 0.99111$

      i. Technology = Yes : 3 of 9 examples

$$p = 1/3, n = 2/3, H_y = 0.9183$$

Technology = No : 6 of 9 examples

$$p = 4/6, n = 2/6, H_n = 0.9183$$

Expected entropy: $H(tech) = (3/9) * 0.9183 + (6/9) * 0.9183 = 0.9183$
Information gain: $Gain(tech) = 0.99111 - 0.9183 = 0.07281$

ii. Environment = Yes : 5 of 9 examples

$$p = 4/5, n = 1/5, H_y = 0.72193$$

Environment = No : 4 of 9 examples

$$p = 1/4, n = 3/4, H_n = 0.81128$$

Expected entropy: $H(en) = (5/9) * 0.72193 + (4/9) * 0.81128 = 0.76164$
Information gain: $Gain(en) = 0.99111 - 0.76164 = 0.22947$

iii. Human = Like : 4 of 9 examples

$$p = 1/4, n = 3/4, H_l = 0.81128$$

Human = Not Care : 4 of 9 examples

$$p = 4/4, n = 0, H_n = 0$$

Human = Hate : 1 of 9 examples

$$p = 0, n = 1/1, H_h = 0$$

Expected entropy: $H(human) = (4/9)*0.81128+(4/9)*0+(1/9)*0 = 0.36057$
Information gain: $Gain(human) = 0.99111 - 0.36057 = 0.63054$

iv. Distance = 1 : 2 of 9 examples

$$p = 1/2, n = 1/2, H_1 = 1$$

Distance = 2 : 1 of 9 examples

$$p = 1/1, n = 0, H_2 = 0$$

Distance = 3 : 3 of 9 examples

$$p = 2/3, n = 1/3, H_3 = 0.9183$$

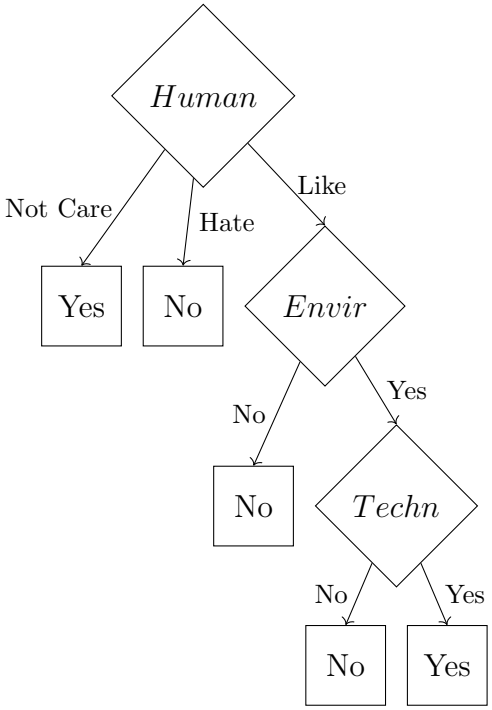
Distance = 4 : 3 of 9 examples

$$p = 1/3, n = 2/3, H_4 = 0.9183$$

Expected entropy: $H(Dist) = (2/9) * 1 + (1/9) * 0 + (1/3) * 0.9183 + (1/3) * 0.9183 = 0.83442$
Information gain: $Gain(Dist) = 0.99111 - 0.83442 = 0.15669$

(c) Human. From the entropy in last problem, choose Human as the root of the tree, the reason is that Human provides the maximum information gain.

(d)



(e)

Here is mine prediction: The accuracy of the classifier is 66.7%.

Technology	Environment	Human	Distance	Invade?	prediction	Correct?
Yes	Yes	Like	2	No	Yes	False
No	No	Hate	3	No	No	True
Yes	Yes	Like	4	Yes	Yes	True

Table 1: Prediction of the test data

3.

(a) [6 points] Using the $MajorityError$ measure, calculate the information gain for the four features respectively. Use 3 significant digits.

(b) [4 points] According to your results in the last question, which attribute should be the root for the decision tree? Do these two measures (entropy and majority error) lead to the same tree?

Solution:

(a) $ME(Invade?) = 1 - \frac{5}{9} = \frac{4}{9}$

i. Technology = Yes : 3 of 9 examples

$$p = \frac{1}{3}, n = \frac{2}{3}, ME_y = 1 - \frac{2}{3} = \frac{1}{3}$$

Technology = No : 6 of 9 examples

$$p = \frac{4}{6}, n = \frac{2}{6}, ME_n = 1 - \frac{4}{6} = \frac{1}{3}$$

Expected entropy: $ME(tech) = (3/9) * \frac{1}{3} + (6/9) * \frac{1}{3} = \frac{1}{3}$
Information gain:

$$Gain(tech) = \frac{4}{9} - \frac{1}{3} = \frac{1}{9} = 0.111$$

ii. Environment = Yes : 5 of 9 examples

$$p = \frac{4}{5}, n = \frac{1}{5}, ME_y = 1 - \frac{4}{5} = \frac{1}{5}$$

Environment = No : 4 of 9 examples

$$p = \frac{1}{4}, n = \frac{3}{4}, ME_n = 1 - \frac{3}{4} = \frac{1}{4}$$

Expected entropy: $ME(en) = \frac{5}{9} * \frac{1}{5} + \frac{4}{9} * \frac{1}{4} = \frac{2}{9}$
Information gain:

$$Gain(en) = \frac{4}{9} - \frac{2}{9} = \frac{2}{9} = 0.222$$

iii. Human = Like : 4 of 9 examples

$$p = \frac{1}{4}, n = \frac{3}{4}, ME_l = \frac{3}{4} = \frac{1}{4}$$

Human = Not Care : 4 of 9 examples

$$p = 4/4, n = 0, ME_n = 0$$

Human = Hate : 1 of 9 examples

$$p = 0, n = 1/1, ME_h = 0$$

Expected entropy: $ME(human) = \frac{4}{9} * \frac{1}{4} + \frac{4}{9} * 0 + \frac{1}{9} * 0 = \frac{1}{9}$
Information gain:

$$Gain(human) = \frac{4}{9} - \frac{1}{9} = \frac{1}{3} = 0.333$$

iv. Distance = 1 : 2 of 9 examples

$$p = 0.5, n = 0.5, ME_1 = 0.5$$

Distance = 2 : 1 of 9 examples

$$p = 1/1, n = 0, ME_2 = 0$$

Distance = 3 : 3 of 9 examples

$$p = 2/3, n = 1/3, ME_3 = \frac{1}{3}$$

Distance = 4 : 3 of 9 examples

$$p = 1/3, n = 2/3, ME_4 = \frac{1}{3}$$

Expected entropy: $ME(Dist) = \frac{2}{9} * 0.5 + \frac{1}{9} * 0 + \frac{3}{9} * \frac{1}{3} + \frac{3}{9} * \frac{1}{3} = \frac{1}{3}$
Information gain:

$$Gain(Dist) = \frac{4}{9} - \frac{1}{3} = \frac{1}{9} = 0.111$$

(b) Human, From the top information gain, the root note will be Human which is the largest information.
These two measures lead to the same root, in this case, these two measures will lead to the same tree.

2 Linear Classifier

1. Solution:

$$x_1 + x_2 + x_3 + x_4 > 1$$

So

$$w = [1, 1, 1, 1], b = -1.5$$

2. Solution:
The accuracy is about 0.857.

x1	x2	x3	x4	o	prediction
0	0	0	1	1	-1
0	0	1	1	1	1
0	0	0	0	-1	-1
1	0	1	0	1	1
1	1	0	0	1	1
1	1	1	1	1	1
1	1	1	0	1	1

3. Solution: As x_1 and x_4 is positive weight. x_2 and x_3 is negative weight. And the absolute value of positive weight are bigger than the absolute value of negative weight. When x1,x2,x3,x4 are 0, o is -1, set $bias = -0.25$, then the linear classifier will be

$$2 * x_1 - 0.5 * x_2 - 0.5 * x_3 + 2 * x_4 - 0.25 > 0$$

So

$$w = [2, -0.5, -0.5, 2], b = -0.25$$

3 Experiments

1. [25 points] **Implementation**

   (a) First, I want to find out how many examples I have. So there is 445 training examples. I decide to choose all binary attributes. Therefore, $2^6 = 64, 2^7 = 128, 2^8 = 256$. In order to make each possible have at least one examples, I think 6 binary attributes would be a good choice. And we only have names, so we need to make the attributes about names. The first name length and last name length, and so on may have some inner relations to help me. Reform the + and - to 1 and 0.

   (b) I choose 6 features and they are all binary attributes.

   i. The first name longer than the last name(Yes:1,No:0)
   ii. The name has a middle name(Yes:1,No:0)
   iii. the first name start and end with the same letter(Yes:1,No:0)
   iv. the first name comes alphabetically before the last name(Yes:1,No:0)
   v. the second letter of their first name a vowel (a,e,i,o,u)(Yes:1,No:0)
   vi. the number of letters in the last name even(Yes:1,No:0)

   (c) After training, there are 21 errors of my decision tree. the row number of the data is [8 17 24 29 33 65 86 101 144 171 218 225 234 256 276 280 282 356 376 434 440]. The accuracy is 0.9527.

   (d) Running the test data, there are 4 errors of my decision tree. the row number of the data is [26 35 51 74]. The accuracy is 0.9640.

   (e) The max depth of my decision tree is 6.

2. [20 points] **Limiting Depth**

   (a) [10 points] Solution:

depth	1	2	3	4	5	6
data1,2,3	0.8649	0.8559	0.8559	0.8649	0.9279	0.9369
data2,3,0	0.8108	0.8829	0.9009	0.9459	0.9459	0.9550
data3,0,1	0.8108	0.8919	0.9189	0.9550	0.9459	0.9459
data0,1,2	0.8198	0.8919	0.8919	0.9459	0.9550	0.9550

depth	1	2	3	4	5	6
maxAccuracy	0.8649	0.8919	0.9189	0.9550	0.9550	0.9550
varAccuracy	0.0007	0.0003	0.0007	0.0018	0.0001	0.0001
meanAccuracy	0.8266	0.8806	0.8919	0.9279	0.9437	0.9482

Set the ID3 function to build a tree only [1,2,3,4,5,6] depth,reform the training data based on three data sets from data1 to data4 in 4 times. And then checking the decision tree in the rest piece of data set, in order to find out how many error of the prediction. Above tables is the result of running. First is the accuracy of prediction as the input is different data set and different depth. The second table is the max-accuracy, variance of accuracy, and mean of accuracy of each depth.

When the depth is 4 ,5 and 6 , it has the same max-accuracy (0.9550). Even the var-accuracy of depth 5 and 6 are smaller than depth as 4. So I will choose data0, data1, and data3 as training data also setting the depth as 4.

(b) [5 points] The most accuracy is 4. After training as depth 4, the decision tree will only have 3 error predictions running for test data. The accuracy is 0.9730.

(c) [5 points] From the performance of depth 4 and depth 6. I think the limiting depth is a good idea which increases the accuracy from 0.9640 to 0.9730 as spending less time and space.

Experiment Submission Guidelines

1. The report should detail your experiments. For each step, explain in no more than a paragraph or so how your implementation works. You may provide the results for the final step as a table or a graph.

2. *Your code should run on the CADE machines.* You should include a shell script, `run.sh`, that will execute your code in the CADE environment. Your code should produce similar output to what you include in your report.

   You are responsible for ensuring that the grader can execute the code using only the included script. If you are using an esoteric programming language, you should make sure that its runtime is available on CADE.

3. Please do not hand in binary files! We will *not* grade binary submissions.

4 Decision Lists (For CS 6350 students)

I am undergraduate.