COEN 240 MACHINE LEARNING

HOMEWROK FIVE

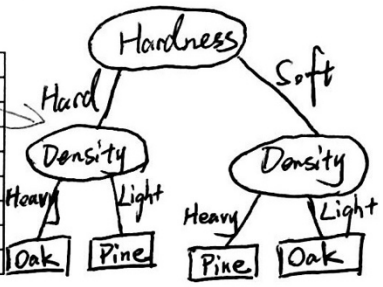
NAME: BOSEN YANG **STUDENT ID: 1589880**

Guideline: Please complete the following problems and generate a PDF file. Please refer to HomeworkFormat.pdf for the format of the submitted PDF file.

Problem 1

You are a robot in a lumber yard, and must learn to discriminate Oak wood from Pine wood. You choose to learn a Decision Tree classifier. You are given the following examples:

Example	Density	Grain	Hardness	Class
Example #1	Heavy	Small	Hard	Oak
Example #2	Heavy	Large	Hard	Oak
Example #3	Heavy	Small	Hard	Oak
Example #4	Light	Large	Soft	Oak
Example #5	Light	Large	Hard	Pine
Example #6	Heavy	Small	Soft	Pine
Example #7	Heavy	Large	Soft	Pine
Example #8	Heavy	Small	Soft	Pine



1.1 Which attribute will be chosen as the root of the tree (show derivations)?

1.2 Derive the complete decision tree by recursively applying the smallest entropy criterion to select root nodes of sub-trees (show derivations). Then draw the complete decision tree.

1.1 $H(C|Density) = \frac{6}{8} \cdot H(C|Density=Heavy) + \frac{2}{8} \cdot H(C|Density=Light)$ ~~$+\frac{3}{8} \cdot H(C|Density$~~

$= \frac{6}{8} \cdot H(\frac{3}{6}, \frac{3}{6}) + \frac{2}{8} \cdot H(\frac{1}{2}, \frac{1}{2}) = \frac{6}{8} \cdot 1 + \frac{2}{8} \cdot 1 = 1$ bit

$H(C|Grain) = \frac{4}{8} \cdot H(C|Grain=Small) + \frac{4}{8} \cdot H(C|Grain=Large)$

$= \frac{4}{8} \cdot H(\frac{2}{4}, \frac{2}{4}) + \frac{4}{8} \cdot H(\frac{2}{4}, \frac{2}{4}) = \frac{4}{8} \cdot 1 + \frac{4}{8} \cdot 1 = 1$ bit

$H(C|Hardness) = \frac{4}{8} \cdot H(C|Hardness=Hard) + \frac{4}{8} \cdot H(C|Hardness=Soft)$

$= \frac{4}{8} \cdot H(\frac{3}{4}, \frac{1}{4}) + \frac{4}{8} \cdot H(\frac{1}{4}, \frac{3}{4}) = \frac{1}{4} \cdot \log_2 4 + \frac{3}{4} \cdot \log_2 \frac{4}{3} = 0.811$ bit

Hardness will be chosen as the root because it has the smallest entropy (uncertainty)

1.2 $H(C|Density, Hardness=Hard) = \frac{3}{4} \cdot H(1,0) + \frac{1}{4} \cdot H(0,1) = 0$ bit

$H(C|Grain, Hardness=Hard) = \frac{3}{4} \cdot H(\frac{2}{3}, \frac{1}{3}) + \frac{1}{4} \cdot H(0,1) > H(C|Density, Hardness=Hard)$

We choose Density as node under condition Hardness = Hard.

$H(C|Density, Hardness=Soft) = \frac{1}{4} \cdot H(1,0) + \frac{3}{4} \cdot H(0,1) = 0$ bit

$H(C|Grain, Hardness=Soft) = \frac{2}{4} \cdot H(\frac{1}{2}, \frac{1}{2}) + \frac{2}{4} \cdot H(0,1) > H(C|Density, Hardness=Soft)$

Therefore, Density will be chosen as node under condition Hardness = Soft. Till now, we've reached entropy = 0 on all leaf nodes. We are able to draw the complete decision tree as follows.

Problem 2

!!!

NASA wants to discriminate Martians (M) from Humans (H) based on these features (attributes): Green ∈ {N, Y}, Legs∈ {2,3}, Height∈ {S, T}, Smelly ∈ {N, Y}. Your available training data is as follows (N=No, Y=Yes, S=Short, T=Tall):

Example Number		Height	Green	Legs	Smelly		Target: Species
1		S	Y	3	Y		M
2		T	Y	3	N		M
3		S	Y	3	N		M
4		T	Y	3	N		M
5		T	N	2	Y		M
??? 6		T	Y	2	Y		H
7		S	N	2	N		H
8		T	N	3	N		H
9		S	N	3	N		H
10		T	N	3	N		H

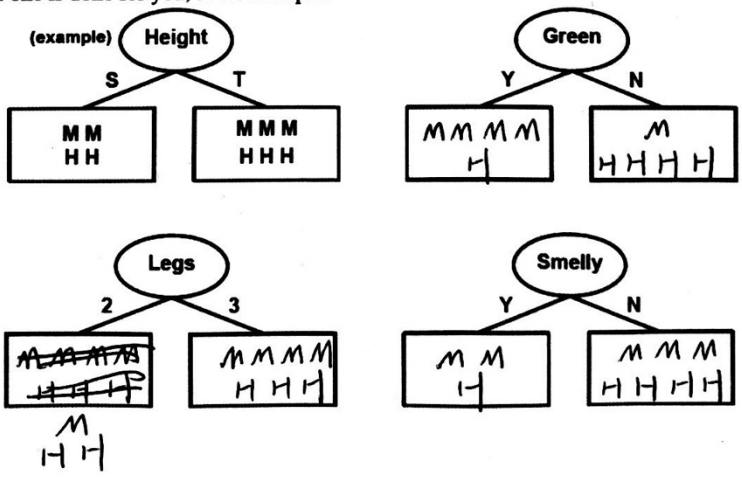
Please note:
A human might be green or have three legs for many possible reasons, e.g., if they were an actor playing a Martian as a role in a film or play. Anyway, it's a made-up problem

Great Point!

(a) What is the entropy of the target species before testing any attribute?

$H(\frac{5}{10}, \frac{5}{10}) = H(\frac{1}{2}, \frac{1}{2}) = 1$ bit

(b) For each possible choice of root attribute below, show the resulting species distribution. Give your answer as M over H. The first one is done for you, as an example.



(c) What is the conditional entropy of species under attribute Height?

$H(C|\text{Height}) = \frac{4}{10} \cdot H(\frac{1}{2}, \frac{1}{2}) + \frac{6}{10} \cdot H(\frac{1}{2}, \frac{1}{2}) = H(\frac{1}{2}, \frac{1}{2}) = 1$ bit

(d) What is the conditional entropy of species under attribute Green?

$H(C|\text{Green}) = \frac{5}{10} \cdot H(\frac{4}{5}, \frac{1}{5}) + \frac{5}{10} \cdot H(\frac{1}{5}, \frac{4}{5}) = H(\frac{1}{5}, \frac{4}{5}) = 0.722$ bit

(e) What is the conditional entropy of species under attribute Legs?

$H(C|\text{Legs}) = \frac{3}{10} \cdot H(\frac{1}{3}, \frac{2}{3}) + \frac{7}{10} \cdot H(\frac{4}{7}, \frac{3}{7}) = \frac{3}{10} \cdot 0.9183 + \frac{7}{10} \cdot 0.9852$

$= 0.965$ bit

(f) What is the conditional entropy of species under attribute Smelly?

$H(C|\text{Smelly}) = \frac{3}{10} \cdot H(\frac{2}{3}, \frac{1}{3}) + \frac{7}{10} \cdot H(\frac{3}{7}, \frac{4}{7}) = H(C|\text{Legs}) = 0.965$ bit

(g) Which attribute would you select as the root attribute (i.e., the attribute to test first)? Why?

Green will be selected because it has the smallest entropy/uncertainty.

(h) Derive the complete decision tree by recursively applying the smallest entropy criterion to select root nodes of sub-trees (show derivations). Then draw the complete decision tree.

From previous questions we know Green will be chosen as root.

$H(C|\text{Height}, \text{Green}=Y) = \frac{2}{5} \cdot H(1,0) + \frac{3}{5} \cdot H(\frac{2}{3}, \frac{1}{3})$

$H(C|\text{Legs}, \text{Green}=Y) = \frac{4}{5} \cdot H(1,0) + \frac{1}{5} \cdot H(0,1)$

} Legs will be selected as node under condition Green = Y.

$H(C|\text{Height}, \text{Green}=N) = \frac{2}{5} \cdot H(0,1) + \frac{3}{5} \cdot H(\frac{1}{3}, \frac{2}{3})$

$H(C|\text{Legs}, \text{Green}=N) = \frac{3}{5} \cdot H(0,1) + \frac{2}{5} \cdot H(\frac{1}{2}, \frac{1}{2})$

$H(C|\text{Smelly}, \text{Green}=N) = \frac{1}{5} \cdot H(1,0) + \frac{4}{5} \cdot H(0,1)$

} Smelly will be selected as node under condition Green = N.

Now on all leaf nodes we have entropy = 0, which means we are able to draw the complete decision tree.