

## Identification Trees Solution (from Exam 2, Fall 2001)

You have been appointed as a research assistant at PacMan Institute of Technology, where you work for Pinky, Winky, Binky, and Hinky on research projects.

Binky asks you to build a Ghostbot that can collect and dispose of styrofoam foodtruck containers (a very useful task around P.I.T.). They should go into two different bins P (for Powerpellet Kitchen) or S (for Strawberry King).

You are given the training set (repeated on a tear off sheet for your convenience at the end of this examination). Note that "Maybe" is a possible result for the Unfinished? test.

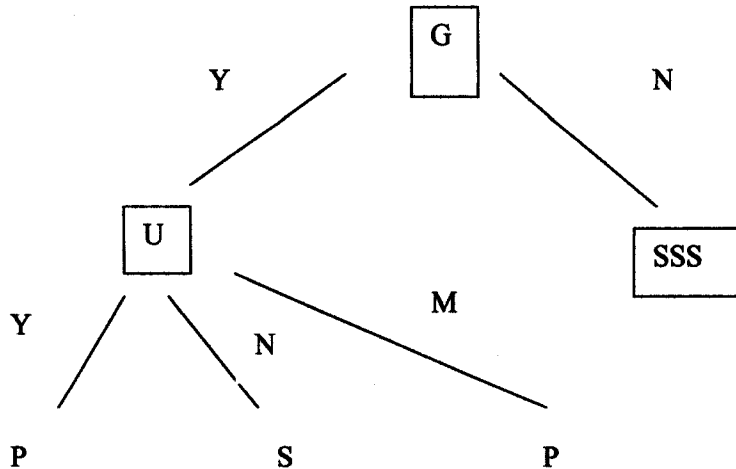
Training #	Colorful?	Unfinished?	Greasy?	Truck
1	Y	Y	N	S
2	N	Y	N	S
3	N	Y	N	S
4	Y	Y	Y	P
5	N	N	Y	S
6	N	Maybe	Y	P

And you are given the following table (also on the tear-off page), in case you forgot your calculator. If the number you need is not in the table, use the table to estimate the number you need.

	n	log <sub>2</sub> n	- n log <sub>2</sub> n - (1-n)log <sub>2</sub> (1-n)
	0.00	0.00	0.00
	0.05	-4.32	0.29
	0.10	-3.32	0.47
	0.15	-2.74	0.61
	0.20	-2.32	0.72
	0.25	-2.00	0.81
	0.30	-1.74	0.88
	0.35	-1.51	0.93
	0.40	-1.32	0.97
	0.45	-1.15	0.99
	0.50	-1.00	1.00
1/9	0.11	-3.17	0.50
1/8	0.13	-3.00	0.54
1/7	0.14	-2.81	0.59
1/6	0.17	-2.58	0.65
1/5	0.20	-2.32	0.72
1/4	0.25	-2.00	0.81
1/3	0.33	-1.58	0.92

**Part A (12 Points)**

Exhibit the identification tree produced by the standard disorder-based tree builder in the box below. You may abbreviate the tests as C, U, and G. (You may wish to show your work to ensure maximum partial credit).



**Part B (5 Points)**

Next, convert the tree from Part A into a set of equivalent IF-THEN rules. You may abbreviate the tests as C, U, and G.

**If G=Y and U=Y  
Then P**

**If G=Y and U=N  
Then S**

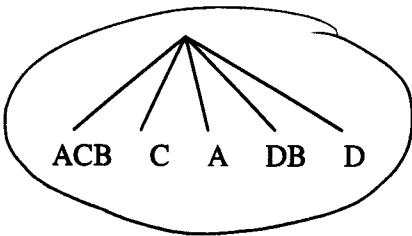
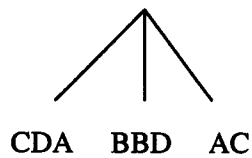
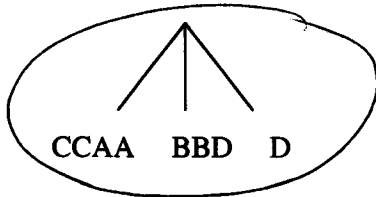
**If G=Y and U=M  
Then P**

**If G=N  
Then S**

*Each path to a leaf provides a rule. We allowed ors in the rules, but not elses, which don't make sense.*

**Part C (7 points)**

Next, you are commanded to decide which the following is the **best** candidate as the **next branch** of a decision tree under a node that receives samples CCAABBDD. **Circle One**, **calculate the average disorder for the one you circle**, and **write that average disorder in the box below**.



**Average disorder for best test:**

First tree:

$$\begin{aligned} & 1/2(-1/2\log_2 1/2 - 1/2\log_2 1/2) + 3/8(-1/3\log_2 1/3 - 2/3\log_2 2/3) \\ &= 1/2(2)(-1/2\log_2 1/2) + 3/8(-1/3\log_2 1/3 - 2/3\log_2 2/3) \\ &= 1/2*1 + 3/8*0.92 + 0 \\ &= 0.845 \end{aligned}$$

Second tree:

$$\begin{aligned} & 3/8(-1/3\log_2 1/3 - 1/3\log_2 1/3 - 1/3\log_2 1/3) + 3/8*0.92 + 1/4*1 \\ &= 3/8(3)(-1/3\log_2 1/3) + 3/8*0.92 + 1/4*1 \\ &= .592 + .345 + .25 = 1.187 \end{aligned}$$

Third tree:

$$\begin{aligned} & 3/8(-1/3\log_2 1/3 - 1/3\log_2 1/3 - 1/3\log_2 1/3) + 0 + 0 + 1/4*1 + 0 \\ &= .592 + .25 \\ &= 0.842 \end{aligned}$$

*First and third trees have nearly the same disorder; full credit for circling either the first or third tree and getting the disorder right.*