

Homework 9

A Little Decision Trees

Due: 5:00pm on 4/21/08

Problem 9.1

The informational entropy of a probability distribution is defined as the - expected-value $\log(p(X))$. (Use a base 2 log)

- a) Calculate the entropy of a fair coin toss (one that is heads half the time and tails the other half).
- b) Calculate the entropy of a biased coin that is heads 75% of the time.
- c) Explain in your own words what you think entropy measures and explain why the answer to (1) is greater than the answer to (2).

Problem 9.2

For each of the following problems, say if you think decision trees would work well, decision trees with bagging would work, or if decision trees are the wrong thing to use entirely. Please give a reason for your answer. (You might note that any time decision trees are appropriate, DT with bagging is also appropriate. Choose the one that you feel is more applicable for the situation. You dont want to use bagging unless you think it will be a significant improvement over DT without bagging.)

In CIT 219, everyone has gathered for cs141 lecture. Today is a special day though because everyone is wearing single color shirts. Can you use decision trees to see if the class falls into the following categories:

- a. Diverse: Everyone in the class is wearing a different color shirt.
- b. Well Taught: As you know, the color sense of the Professor and the TAs often determines the quality of the class. If the Meinolf, Andy and Lucia wear Blue, Green, or Purple shirts, the class is Well Taught.
- c. Dominant: A majority of people in the class are wearing the same color shirt.

Problem 9.3

As usual, we want to know if mushrooms are edible.

Let A_{tall} and $A_{spotted}$ be two binary attributes.

Here is what we know about mushrooms:

$$P(A_{tall} = true) = .5 \quad P(A_{spotted} = true) = .5$$

$$P(edible = true|A_{tall} = true) = .1 \quad P(edible = true|A_{tall} = false) = .3$$

$$P(edible = true|A_{spotted} = true) = .125 \quad P(edible = true|A_{spotted} = false) = .25$$

- a. Compute the Gini index value for A_{tall} and $A_{spotted}$
- b. Based on the Gini values, which attribute should you branch on?
- c. Compute the entropy for A_{tall} and $A_{spotted}$
- d. Based on the entropy values, which attribute should you branch on?