

Homework 9

A Little Decision Trees

Due: 5:00pm on 4/23/08

Problem 9.1

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 don't 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:

- Diverse: Everyone in the class is wearing a different color shirt.
- 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.
- Dominant: A majority of people in the class are wearing the same color shirt.

Solution: taken from jamo and yechin's homework

- Diverse:** To use a decision tree to determine Diverse will need to have the tree run through all people in the class to see if there are any two wearing the same color shirt. This would yield an exponentially large decision tree representation. Therefore, decision trees are the wrong thing to use entirely.
- Well Taught:** Using a decision tree will work well to determine Well Taught. The tree simply needs to branch on Meinolf, Andy and Lucia's shirt color to reach a decision.
- Dominant:** Using decision trees with bagging seems an applicable approach to determine Dominance. A forest can be built with small trees that only test a limited number (n) of people's shirt color to see if there is a dominant color. To determine the Dominant shirt color of the whole class, a randomly selected subset of (n) people will be applied on each decision tree. And if the majority of decision trees in the forest agree on the same dominant color, the whole class is classified to be Dominant

Problem 9.2

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:

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

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

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

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

Solution:

a. gini: $R(A_{tall}) = .5 * (1 - .1^2 - .9^2) + .5 * (1 - .3^2 - .7^2) = .3$ $R(A_{spotted}) = .5 * (1 - .125^2 - .875^2) + .5 * (1 - .25^2 - .75^2) = .2969$

b. you branch on the smaller of the remainder (largest info gain), so gini would branch on $A_{spotted}$

c. entropy: $R(A_{tall}) = .5 * (-.1 * \log_2(.1) - .9 * \log_2(.9)) + .5 * (-.3 * \log_2(.3) - .7 * \log_2(.7)) = 0.6751$ $R(A_{spotted}) = .5 * (-.125 * \log_2(.125) - .875 * \log_2(.875)) + .5 * (-.25 * \log_2(.25) - .75 * \log_2(.75)) = 0.6774$

d. A_{tall}

Problem 9.3

Take a look at Russel and Norvig page 659 "Choosing attribute tests" and notice that their definition of the information gain only works for two-class classification (where their two classes are p and n... positive and negative). How could you extend the definition to work with multi-class classification? This way we aren't restricted to binary classifications. For example, we could classify different text documents into k different categories. (Also you will be using this formula for your upcoming decision tree project)

Note: The book's method of choosing the best attribute is a bit inefficient. $\arg \max_A \text{Gain}(A) = \arg \max_A [I(\frac{p}{p+n}, \frac{n}{p+n}) - \text{Remainder}(A)] = \arg \min \text{Remainder}(A)$ as $I(\frac{p}{p+n}, \frac{n}{p+n})$ is constant over all attributes (remember n is the number of negative examples and p is the number of positive examples in binary classification). we want you to define $\text{Remainder}(A)$ for multi-class classification.

Hint: Remember conceptually, $\text{Remainder}(A) = \sum_{\text{choice} \in A} q_{\text{choice}} I(q_{\text{choice}_p}, q_{\text{choice}_n})$, where q_{choice} is the probability of being in choice A and q_{choice_p} is the probability that given we are in choice of an example is classified as positive and q_{choice_n} is the same for negative instead.

Solution: $\text{Remainder}(A) = \sum_{c \in A} q_c I(\vec{q}_c) = \sum_{c \in A} q_c \sum_{i=1}^k -q_{c_i} \log q_{c_i}$ **where q_{c_i} is the probability given we are in choice of an example is classified as category i .**