Nick Palacio

Raj Dasgupta

October 1, 2018

Introduction to AI

Homework 2

Mathematics Part

**Question 1.**

*You have a decision tree algorithm and you are trying to figure out which attribute is the best to test on first. You are using the “information gain” metric.*

* + *You are given a set of 128 examples, with 64 positively labeled and 64 negatively labeled.*
* *There are three attributes: Homeowner (H), In Debt (ID), and Rich (R).*
* *For 64 examples, Home Owner is true. The Homeowner=true examples are 1/4 negative and 3/4 positive.*
* *For 96 examples, In Debt is true. Of the In Debt=true examples, 1/2 are positive and half are negative.*
* *For 32 examples, Rich is true. 3/4 of the Rich=true examples are positive and 1/4 are negative*

*You must show all mathematical calculations/steps to get full points for each subpart (a) – (d) below. Just writing the final answer in each subpart (correct or not) will get zero points.*

1. *What is the entropy of the initial set of examples?*
2. *What is the information gain of splitting on the Home Owner attribute as the root node?*
3. *What is the information gain of splitting on the In Debt attribute as the root node?*
4. *What is the information gain of splitting on the Rich attribute as the root node?*
5. *Which attribute do you split on?*

1a. Since in the initial set half the examples are positive and half are negative our initial entropy is given as:

1b. For the 64 examples where Home Owner is true, 16 (1/4) are negative, and 48 (3/4) are positive. This means that for the other 64 examples where Home Owner is false, 48 (3/4) are negative and 16 (1/4) are positive. The information gain of splitting on the Home Owner as the root node is given as:

1c. For the 96 examples where In Debt is true, 48 (1/2) are negative and 48 (1/2) are positive. This means that for the other 32 examples where In Debt is false, 16 (1/2) are negative and 16 (1/2) are positive. The information gain of splitting on In Debt as the root node can be given as:

1d. For the 32 examples where Rich is true, 8 (1/4) are negative and 24 (3/4) are positive. This means that for the 96 examples where Rich is false, 56 are negative and 40 are positive. The information gain of splitting on Rich as the root node can be given as:

1e. Since:

I would split on Home Owner so I can have the largest information gain.