CMSC5724: Exercise List 1

By Yufei Tao

Problem 1. Assume that we have the following training set:

refund	**marital**	**income**	**cheat**
yes	single	125	no
no	married	100	no
no	single	70	no
yes	married	120	no
no	divorced	95	yes
no	married	60	no
yes	divorced	220	no
no	single	85	yes
no	married	75	no
no	single	90	yes

In the above table, *cheat* is the class label, whereas the other columns are the attributes.

(i) What is the Gini value of the original table?

(ii) Now let us create the first internal node in our decision tree. Recall that our algorithm does so by looking for the best split, for which purpose the algorithm examines each dimension in turn. Let us consider first attribute *refund*. Since this is a binary attribute, there is only one possible split. What is the Gini of this split?

(iii) Let us now focus on attribute *marital*. How many splits are possible on this attribute? Which one is the best one, and what is its Gini?

(iv) Repeat the above for attribute *income*.

(v) Considering all dimensions, which one is the best split? What is its Gain? Recall that the *Gain* of a split equals the difference between (a) Gini value before the split and (b) the Gini of the split.

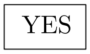
Problem 2. Consider a classification problem where there is only one attribute A and one class label B. Both A and B have binary domains. Specifically, A has two values a_0 and a_1 while B has two values *yes* and *no*. We want to build a decision tree such that given a person, by looking at her/his A value, we predict her/his class label.

Suppose that we already know the following statistics:

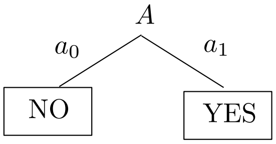
- 90% of the population have A value a_0.
- Among those people with $A = a_0$, 70% belong to the *yes* class.
- Among those people with $A = a_1$, 70% belong to the *no* class.

We can assume that each person to be classified is randomly picked from the population. Answer the following questions.

(i) What is the error probability of the following decision tree (which contains only one leaf)?



(ii) What is the error probability of the following decision tree?



Problem 3 (Finding the Best Split on an Ordered Dimension). Consider a table with two attributes (A, B) where A is an ordered attribute, and B the class label. Let n be the number of records in the table. Describe an algorithm that computes the best split along dimension A in $O(n \log n)$ time.