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?

Answer: The Gini value equals $1 - p_y^2 - p_n^2$ where p_y (p_n) is the percentage of the yes (no) records. Here, $p_y = 0.3$ and $p_n = 0.7$. Hence, the Gini value is $1 - 0.09 - 0.49 = 0.42$.

(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?

Answer: Consider the following table:

cheat	**refund**	
	yes	no
yes	0	3
no	3	4
total	3	7

The table should be read as follows. Splitting on *refund* creates a left (resp., right) child for *refund* = yes (resp., no). Then, the left child contains 3 records, among which 0 (resp., 3) satisfy *cheat* = yes (resp., no). Hence, GINI(left) $= 1 - (0/3)^2 - (3/3)^2 = 0$. The right child contains 7 records, among which 3 (resp., 4) satisfy *cheat* = yes (resp., no). Hence, GINI(right) $= 1 - (3/7)^2 - (4/7)^2 = 0.490$. The Gini of the split equals:

$$\frac{n_{left}}{n} \cdot \text{GINI(left)} + \frac{n_{right}}{n} \cdot \text{GINI(right)}$$

where n_{left} (n_{right}) is the number of records in the left (right) child, and $n = n_{left} + n_{right}$. Therefore, the Gini of the above split equals $(3/10) \cdot 0 + (7/10) \cdot 0.490 = 0.343$.

(iii) Calculate the Gini value for every split on the attribute *marital*.

Answer: There are (essentially) only 3 splits, each of which is illustrated by a table below:

cheat	**marital**	
	{single}	{married, divorced}
yes	2	1
no	2	5
total	4	6
	Gini of split = 0.367	

cheat	**marital**	
	{married}	{single, divorced}
yes	0	3
no	4	3
total	4	6
	Gini of split = 0.3	

cheat	**marital**	
	{divorced}	{single, married}
yes	1	2
no	1	6
total	2	8
	Gini of split = 0.4	

The best split is the second one.

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

Answer. There are 9 splits, as shown below:

cheat	**income**	
	≤ 60	> 60
yes	0	3
no	1	6
	Gini of split = 0.4	

cheat	**income**	
	≤ 70	> 70
yes	0	3
no	2	5
	Gini of split = 0.375	

cheat	**income**	
	≤ 75	> 75
yes	0	3
no	3	4
	Gini of split = 0.342	

cheat	**income**	
	≤ 85	> 85
yes	1	2
no	3	4
	Gini of split = 0.417	

cheat	**income**	
	≤ 90	> 90
yes	2	1
no	3	4
	Gini of split = 0.4	

cheat	**income**	
	≤ 95	> 95
yes	3	0
no	3	4
	Gini of split = 0.3	

cheat	**income**	
	≤ 100	> 100
yes	3	0
no	4	3
	Gini of split = 0.342	

cheat	**income**	
	≤ 120	> 120
yes	3	0
no	5	2
	Gini of split = 0.375	

cheat	**income**	
	≤ 125	> 125
yes	3	0
no	6	1
	Gini of split = 0.4	

The best one is to split at 95.

(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.

Answer. Actually 2 splits are equally the best, i.e., the best splits in (iii) and (iv), respectively. The Gini of each split is 0.3. Hence, its Gain is $0.42 - 0.3 = 0.12$. By the way, in this case, the decision tree construction algorithm will pick one of the two splits randomly to create the root.

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:

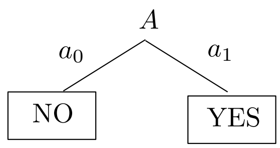
- 90% of the population have A value a_0.
- Among those people with $A = a_0$, 70% belong to the *yes* class.
- Among those people with $A = a_1$, 70% belong to the *no* class.

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

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

Answer. This tree makes a mistake in classification when a person belongs to the *no* class. The probability for a random person to belong to this class equals $0.9 \cdot 0.3 + 0.1 \cdot 0.7 = 0.34$.

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



Answer. This tree makes a mistake in classification when (i) a person has $A = a_0$ but belongs to the *yes* class, or (ii) a person has $A = a_1$ but belongs to the *no* class. The probability that (i) happens is $0.9 \cdot 0.7 = 0.63$, while the probability that (ii) happens is $0.1 \cdot 0.7 = 0.07$. Therefore, the mis-classification probability is $0.63 + 0.07 = 0.7$.

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.

Answer. Let S be the set of records in the table. Each record r is in the form (r_A, r_B), representing its values on A and B, respectively. For simplicity, we will assume that all records have distinct values on A. Removing this assumption requires only minor modification to our algorithm, which is left for you to figure out.

The value $a = r_A$ of each record $r \in S$ defines a possible split, which divides S into: (i) S_1, which includes all the records $r' \in S$ satisfying $r'_A \leq a$, and (ii) S_2 with records $r' \in S$ satisfying $r'_A > a$. To calculate the split's Gini value, we need 4 counts:

- The number of yes-records (i.e., yes on B) in S_1; denote this as $c_y^1(a)$;
- The number of no-records in S_1;
- The number of yes-records in S_2;
- The number of no-records in S_2.

Overall, we need to obtain $4(n-1)$ counts (there are only $n-1$ "meaningful" splits on A because if a is the maximum A value in S, the split has no effects). We will show that all these counts can be obtained in $O(n \log n)$ total time, after which it is easy to compute each split's Gini value in $O(n)$ time.

Due to symmetry, it suffices to explain how to obtain $c_y^1(a)$ for all possible a. Sort S in ascending order by A. Set a count $c = 0$. Scan the records of S in ascending order of A. For each record r, (i) increase c by 1 if $r.B$ is yes, or otherwise, do nothing to c, and then (ii) set $c_y^1(a)$ to c for $a = r.A$.