Abelardo Riojas
ISC 4245 Homework 3
Dr. Meyer-Baese Fall 2020

Homework 3

2. Table 4.7
   a. Compute the overall Gini index for the training example

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2,$$

   Where c = number of classes and p(i|t) is the proportion of records that belong to class t
   c = 2 (C0 and C1), therefore
   Gini = 1 - (.5^2 +.5^2) = .5
   b. Compute the Gini index for the customer id attribute
   The proportion of unique customer IDs that belong to each class is always 1.
   Therefore,
   Gini = 1 - 1^2 = 0 for each ID. Overall, this means gini = 0.
   c. Compute the Gini index for gender
   Male:
   Gini = 1 - (.5^2 + .5^2) = .5
   Same for female:
   Gini = 1 - (.5^2 + .5^2) = .5
   d. Gini index for car type
   Family:
   Gini = 1 - ((1/4)^2 + (3/4)^2) = 0.375
   Luxury:
   Gini = 1 - ((1/8)^2 + (7/8)^2) = 0.21875
   Sports:
   Gini = 1 - 1 - ((8/8)^2 + (0/8)^2) = 0
   Overall:
   Gini = (4/20 * .375) + (8/20 * .21875) + (8/20 * 0) = .1625
   e. Gini index for shirt size
   Small:
   Gini = 1 - ((3/5)^2 + (2/5)^2) = .48
   Medium:
   Gini = 1 - ((3/7)^2 + (4/7)^2) = 0.489795918
   Large:
   Gini = 1 - ((2/4)^2 + (2/4)^2) = .5

Extra large:
Gini = Gini = 1 - ((2/4)^2 + (2/4)^2) = .5
Overall:
Gini = (5/20 * .48) + (7/20 * .489795918) + (4/20 * .5) + * (4/20*.5) = .4914

f. Best gini between car type, gender, and shirt size?
   Car type since it has the lowest of the three
g. Explain why customer ID shouldn't be used
   Because each customer is assigned a unique ID and the number is nominal (in name only). There is nothing you can predict with a customer ID.

3. Consider the training examples shown in Table 4.8 for a binary classification problem

a. What is the entropy of this collection of training examples with respect to the positive class?

$$\text{Entropy}(t) = -\sum_{i=1}^{k} p(i|t)\log_2 p(i|t)$$

Therefore
Entropy = - (5/9*log2(5/9) + 4/9*log2(4/9)) = 0.99107606

b. Information gains of a1 and a2 relative to positive and negative classes

$$\Delta = I_{\text{parent}} - \sum_{j=1}^{k} \frac{N_j}{N} I_j$$

Therefore,
Gain(a1) = .99107606 - (4/9 * - ((3/4)*log2(3/4) + 1/4*log2(1/4)) + 5/9 * + (5/9 * - ((1/5)*log2(1/5) + 4/5*log2(4/5)) = .2294
Gain(a2) = .99107606 - (-5/9 * (2/5)*log2(2/5) + 3/5*log2(3/5)) + -4/9 * + (5/9 * (2/4)*log2(2/4) + 2/4*log2(2/4)) = .007

c. For a3, calculate the entropy and information gain for each possible split
   Split = 2.0

a_3	+	-
Under 2.0	1	0
Over 2.0	3	5

Entropy(a_3) = 1/9*(-(1/1)*log2(1/1) - 0) + (8/9)*(-3/8)*log2(3/8) - (5/8)*log2(5/8)) = 0.848
Gain = .99107606 - .848 = 0.14307606

Split = 3.5

a_3	+	-
Under 3.5	1	1
Over 3.5	3	4

Entropy(a_3) = .988510772 =
-2/9*((1/2)*log2(1/2) + 1/2*log2(1/2)) + -7/9*(3/7*log2(3/7)+(4/7)*log2(4/7))
Gain = .99107606 - 0.988510772 = 0.002565288

Split = 4.5

a_3	+	-
Under 4.5	2	1
Over 4.5	2	4

Entropy(a_3) = .0.918295834 =
-3/9*((2/3)*log2(2/3) + 1/3*log2(1/3)) + -6/9*(2/6*log2(2/6)+(4/6)*log2(4/6))
Gain = . - .918295834 = .072780226

Split = 5.5

a_3	+	-
Under 5.5	2	3
Over 5.5	2	2

Entropy(a_3) = .983861441 =
-5/9*((2/5)*log2(2/5) + 3/5*log2(3/5)) + -4/9*(2/4*log2(2/4)+(2/4)*log2(2/4))
Gain = .99107606 - 0.983861441 = 0.007214619

Split = 6.5

a_3	+	-
Under 6.5	3	3
Over 6.5	2	1

Entropy(a_3) = .0.918295834 =

-6/9*((3/6)*log2(3/6) + 3/6*log2(3/6)) + -3/9*(2/3*log2(2/3)+(1/3)*log2(1/3))
Gain = .99107606 - .918295834 = .018310782
Split = 7.5

a_3	+	-
Under 7.5	4	4
Over 7.5	0	1

Entropy(a_3) = 0.888888889 =
-8/9*((4/8)*log2(4/8) + 4/8*log2(4/8)) + -1/9*(0/1*log2(0/1)+(1/1)*log2(1/1))
Gain = .99107606 - .8888888889 = .102187171

d. What is the best split (among α1 , α2 , and α3) according to the information gain?
a1 has the best split due to it having the highest gain.

e. What is the best split (among a1 and a2) in terms of the classification error?

$$\text{Classification error}(t) \;=\; 1 - \max_{i}^{i=0}[p(i|t)],$$

For a1, max(p(i|t)) = 7/9 therefore Classification error(a1) = 2/9
For a2, max(p(i|t)) = 5/9 therefore Classification error(a1) = 4/9

a1 is the better split.

f. Best split between a1 and a2 according to the gini index?
Gini(a1) = 4/9 * (1 - (3/4)^2 - (1/4)^2) + 5/9*(1 - (1/5)^2 - (4/5)^2) = .344
Gini(a2) = 5/9 * (1 - (2/5)^2 - (3/5)^2) + 4/9*(1 - (2/4)^2 - (2/4)^2) = .489

Gini is smaller for a1, therefore a1 is best split.

5. Binary class problem
Eparent = -(4/10*log2(4/10) + 6/10*log2(6/10)) = 0.970950594

Cross tabs:

a.

	T_A	F_A
+	4	0
-	3	3

Δ_A =.971 - (7/10)*((-4/7)*log2(4/7)+(-3/7)*log2(3/7)) + (-3/10)*(0+(-3/3)*log2(3/3))
= 0.281340305

	T_A	F_A
+	3	1
-	1	5

$\Delta_B =$
.971 - (4/10)*((-3/4)*log2(3/4)+(-1/4)*log2(1/4)) + (-6/10)*((-1/6)*log2(1/6)+(-5/6)*log2(5/6))
= 0.256475297

A > B, therefore A.

b. Gini index on A and B and choose split
Gini = 1 - ((4/10)^2 + (6/10)^2) = 0.48

$Gini_{AT}$ = 1 - ((4/7)^2 + (3/7)^2) = 0.489795918
$Gini_{\mathbf{AF}}$ = 1 - ((0/3)^2 + (3/3)^2) = 0

Δ_A =.48 -(7/10)*(.489795918)+(-3/10)(0) = 0.137142857

$Gini_{BT}$ = 1 - ((3/4)^2 + (1/4)^2) = 0.375
$Gini_{BF}$ = 1 - ((1/6)^2 + (5/6)^2) = 0.277777778

Δ_B = .48 - (4/10)*(.375) + (-6/10)*(.277777778) = 0.163333333

B > A, therefore B.

c. Yes it's possible for them to favor different attributes, in (b), it showed that different splits can be favored using different impurity measures.

7. 3 3 attributes, 2 class labels

a. First attribute:
$Error_{parent}$ = max(50/100,50/100) = .5

	T_A	F_A
+	25	25
-	0	50

E_{AT} = 1 - max(25/25,0/25) = 0
E_{AF} = 1 - max(25/75,50/75) = 25/75
E_A = .5 - ((25/100)*(0)) + (-75/100)*(25/75) = .5

	T_A	F_A
+	25	25
-	0	50

E_{AT} = 1 - max(25/25,0/25) = 0
E_{AF} = 1 - max(25/75,50/75) = 25/75
E_A = .5 - ((25/100)*(0)) + (-75/100)*(25/75) = .5

	T_B	F_B
+	30	20
-	20	30

E_{BT} = 1 - max(30/50,20/50) = 20/50
E_{BF} = 1 - max(20/50,30/50) = 20/50
E_B = .5 - ((50/100)*(20/50)) + (-50/100)*(20/50) = .1

	T_C	F_C
+	25	25
-	25	25

E_{CT} = 1 - max(25/50,25/50) = 25/50
E_{CF} = 1 - max(25/50,25/50) = 25/50
E_C = .5 - ((50/100)*(25/50)) + (-50/100)*(25/50) = .0

A > B and C, therefore A.

b. Repeat for two child nodes
A=F node:
E_{parent} = 1 - max(25/75,50/75) = 25/75

	T_B	F_B
+	25	0
-	20	30

E_{BT} = 1 - max(25/45,20/45) = 20/45

E_{BF} = 1 - max(0/30,30/30) = 0
$Gain_B$ = (25/75) - ((45/75)*(20/45)) + (-30/75)*(0) = 5/75

	T_C	F_C
+	0	25
-	25	25

E_{CT} = 1 - max(0/25,25/25) = 0
E_{CF} = 1 - max(25/50,25/50) = .5
$Gain_C$ = (25/75) - (25/75)*(0) + (-50/75)*(.5)= 0

B > C, therefore, B for A = F node,
A = T node is pure, so stop.

c. How many misclassifications
20/100 since 20 positive 'splits' were classified as Falses and zero negatives were counted as Trues.

d. Repeat parts a b and c using C as the first splitting attribute
$Error_{parent}$ = max(50/100,50/100) = .5

	T_A	F_A
+	25	25
-	0	50

E_{AT} = 1 - max(25/25,0/25) = 0
E_{AF} = 1 - max(25/75,50/75) = 25/75
E_A = .5 - ((25/100)*(0)) + (-75/100)*(25/75) = .5

	T_A	F_A
+	25	25
-	0	50

E_{AT} = 1 - max(25/25,0/25) = 0
E_{AF} = 1 - max(25/75,50/75) = 25/75
E_A = .5 - ((25/100)*(0)) + (-75/100)*(25/75) = .5

	T_B	F_B

+	30	20
-	20	30

E_{BT} = 1 - max(30/50,20/50) = 20/50
E_{BF} = 1 - max(20/50,30/50) = 20/50
E_{B} = .5 - ((50/100)*(20/50)) + (-50/100)*(20/50) = .1

	T_C	F_C
+	25	25
-	25	25

E_{CT} = 1 - max(25/50,25/50) = 25/50
E_{CF} = 1 - max(25/50,25/50) = 25/50
E_{C} = .5 - ((50/100)*(25/50)) + (-50/100)*(25/50) = .0

A > B and C, therefore A, but the problem tells us to use C, so we're using C.

C=T node:
E_{parent} = 1 - max(25/50,25/50) = 25/50

	T_A	F_A
+	25	0
-	0	25

E_{AT} = 1 - max(0/25,25/25) = 0
E_{AF} = 1 - max(25/50,25/50) = 0
$Gain_{A}$ = (25/50) - 0 = .5

	T_B	F_B
+	5	20
-	20	5

E_{BT} = 1 - max(5/25,20/25) = 5/25
E_{BF} = 1 - max(20/50,5/25) = 5/25
$Gain_{B}$ = (25/75) - (25/55)*(5/25) + (25/50)*(5/25) = .3

A > B, therefore, A for C = T node.

C=F node:
E_{parent} = 1 - max(25/50,25/50) = 25/50

	T_A	F_A
+	0	25
-	0	25

E_{AT} = 1 - max(0/0,0/0) = 0
E_{AF} = 1 - max(25/50,25/50) = .5
$Gain_A$ = (25/50) - (.5) = 0

	T_B	F_B
+	25	25
-	0	0

E_{BT} = 1 - max(25/25,0/25) = 0
E_{BF} = 1 - max(25/25,0/25) = 0
$Gain_B$ = 25/50 - 25/50*0 + 25/50*0 = .25
B > A, therefore, B for C = F node.

Error for both nodes is 0 as they both do not misidentify any of the records.

e. What does this tell you about the greedy nature of the algorithm?
Sometimes taking the dataset that splits the data the best first doesn't work, as illustrated here, as the overall error rate using C as the first splitting attribute yields a lower overall error rate.

9. Decision tree cost

- The total description length of a tree is given by:
  $$Cost(tree, data) = Cost(tree) + Cost(data|tree).$$
- Each internal node of the tree is encoded by the ID of the splitting attribute. If there are m attributes, the cost of encoding each attribute is $log_2(m)$ bits.
- Each leaf is encoded using the ID of the class it is associated with. If there are k classes, the cost of encoding a class is $log_2(k)$ bits.

- $Cost(tree)$ is the cost of encoding all the nodes in the tree. To simplify the computation, you can assume that the total cost of the tree is obtained by adding up the costs of encoding each internal node and each leaf node.
- $Cost(data|tree)$ is encoded using the classification errors the tree commits on the training set. Each error is encoded by $log_2(n)$ bits, where n is the total number of training instances.

Cost(tree1,data) = Cost(tree1) + Cost(data,tree1)

Cost(tree1) = cost(internal nodes) + cost(leaf nodes)
Cost of encoding each attribute in internal node is log2(16) = 4 bits

Cost of encoding each attribute in leaf node is log2(3) = 1.5849625 or 2 bits.

2 nodes = 8 bits
3 internal nodes = 6 bits.

14 + 7 * log(n) where n is the number of training instances

Cost(tree2) = cost(internal nodes) + cost(leaf nodes)
Cost of encoding each attribute in internal node is log2(16) = 4 bits

Cost of encoding each attribute in leaf node is log2(3) = 1.5849625 or 2 bits.

4 nodes = 16 bits
5 internal nodes = 10 bits.

26 + 4 * log(n) where n is the number of training instances

The graphs of both functions of n intersect at n = 16, therefore A is better under 16 training instances, and B is better over 16 training instances.