Mark Gameng

CS 422 – HW 4

Part 1 – Questions, Chapter 3

2. Questions on Table 3.5. Let **G** be the Gini index.
   - a. 10 C0 and 10 C1
     G = 1 – ((10/20)^2 + (10/20)^2)
     G = 1 – 0.5 = **0.5**
   - b. 20 Customer IDs – unique
     G1 = 1 – ((1/1)^2 + (0/1)^2) = 0 – for all 20
     Thus, weighted average:
     G = (1/20) * 0 + (1/20) * 0 + … = **0**
   - c. 10 Males and Females
     Gm = 1 – ((6/10)^2 + (4/10)^2) = 0.48
     Gf = 1 – ((4/10)^2 + (6/10)^2) = 0.48
     Thus, G = (10/20) * 0.48 + (10/20) * 0.48 = **0.48**
   - d. 4 Family, 8 Sports, 8 Luxury
     Gf = 1 – ((1/4)^2 + (3/4)^2) = 0.375
     Gs = 1 –((8/8)^2 + (0/8)^2) = 0
     Gl = 1 – ((1/8)^2 + (7/8)^2) = 0.21875
     Thus, G = (4/20) * 0.375 + (8/20) * 0 + (8/20) * 0.21875 = **0.1625**
   - e. 5 Small, 7 Medium, 4 Large, 4 Extra Large
     Gs = 1 – ((3/5)^2 + (2/5)^2) = 0.48
     Gm = 1 – ((3/7)^2 + (4/7)^2) = 0.49
     Gl = 1 – ((2/4)^2 + (2/4)^2) = 0.5
     Gel = 1 – ((2/4)^2 + (2/4)^2) = 0.5
     Thus, G = (5/20) * 0.48 + (7/20) * 0.49 + (4/20) * 0.5 + (4/20) * 0.5 = **0.4915**
   - f. **Car type** is better because the Gini index is the lowest out of the three, 0.1625.
   - g. Customer ID should not be used because it is unique for every customer, so it will always result in a Gini index of 0.
3. Questions on Table 3.6. Let **E** be the entropy
   - a. E = -((4/9)log(4/9) + (5/9)log(5/9)) = **0.991**
   - b. A1 -> 4 T (3 +, 1 -), 5 F (1 +, 4 -)
     Ea1 = -(4/9)((3/4)log(3/4)+(1/4)log(1/4))+(5/9)((1/5)log(1/5)+(4/5)log(4/5)) = **0.762**
     a1_gain = 0.991 – 0.762 = **0.229**
     A2 -> 5 T (2 +, 3 -), 4 F (2 +, 2 -)
     Ea2 = -(5/9)((2/5)log(2/5)+(3/5)log(3/5))+(4/9)((2/4)log(2/4)+(2/4)log(2/4)) = **0.983**
     a2_gain = 0.991 – 0.983 = **.008**
     The information gains of a1 and a2 are 0.229 and 0.008 respectively.
   - c.

A3	Class	Split	Entropy	Information Gain
1	+	2	0.848	0.143

3	-	3.5	0.988	0.003
4	+	4.5	0.918	0.073
5	-	5.5	0.983	0.008
5	-			
6	+	6.5	0.972	0.019
7	-	7.5	0.888	0.103
7	+			
8	-			

For first split, 2, E = -((1/9)((1/1)log(1/1))+(8/9)((3/8)log(3/8)+(5/8)log(5/8))) = **0.848**
Same formulas for the rest of the splits. The results are in the table above.

Looking at the information gain for each split, the most information gain is at 2, which makes the best split for a3 to be at the split point, 2, with an entropy of 0.848.

d. Looking at the information gains, the best split is a1 with information gain of 0.229 and entropy of 0.762.

e. The misclassification error rate for a1 and a2 are 2/9 and 4/9 respectively. Thus, a1 is the best split according to the misclassification rate.

f. Ga1 = (4/9)(1-(3/4)^2-(1/4)^2)+(5/9)(1-(1/5)^2-(4/5)^2) = 0.344
Ga2 = (5/9)(1-(2/5)^2+(3/5)^2)+(4/9)(1-(2/4)^2-(2/4)^2) = 0.888
Gini index for a1 is smaller, 0.344, thus a1 is the best split according to the Gini index.

5. Questions on the table

a. E = -((4/10)log(4/10)+(6/10)log(6/10)) = **0.971**
A -> 7 T, 3 F
Ea = (7/10)(-(4/7)log(4/7)-(3/7)log(3/7))+(3/10)(-(3/3)log(3/3)) = **0.689**
a_gain = 0.971 – 0.689 = **0.282**
B -> 4 T, 6 F
Eb = (4/10)(-(3/4)log(3/4)-(1/4)log(1/4))+(6/10)(-(1/6)log(1/6)-(5/6)log(5/6)) = **0.714**
b_gain = 0.971 – 0.714 = **0.257**
The information gain is more on A, so the decision tree induction algorithm would choose attribute A.

b. G = 1 – (4/10)^2 – (6/10)^2 = **0.48**
Ga = (7/10)(1-(4/7)^2-(3/7)^2+(3/10)(1-(3/3)^2) = **0.34**
a_gain = 0.48 – 0.34 = **0.14**
Gb = (4/10)(1-(3/4)^2-(1/4)^2)+(6/10)(1-(1/6)^2-(5/6)^2) = **0.32**
b_gain = 0.48 – 0.32 = **0.16**
Since the gain is more from B, the decision tree induction algorithm will choose attribute B.

c. Yes, it is possible that information gain and the gain in the Gini index favor different attributes. Even though entropy and Gini index are both monotonically increasing and decreasing in the same ranges, what matters is the gains. Though the entropy and Gini index have similar behaviors, the gain in Gini index and information gain don’t. For example, in this question, part a and b. Using information gain, splitting on A is better but using gain in the Gini index, splitting on B is better.