Tiffany Wong A20442087
CS422 - HW 5 Exercises 10/14/2021

1. Questions
1.1 (1 point) Tan, Chapter 3 Exercise 2, 3, 5.

2. Consider the training examples shown in Table 3.5 classification problem.

Table 3.5. Data set for Exercise 2.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

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

a. *Compute the Gini index for the overall collection of training examples.*

$Total\ number\ of\ records\ =\ 20$

$P(C0)\ =\ 10/20$

$P(C1)\ =\ 10/20$

$Gini\ index\ =\ 1\ -\ [(10/20)^2\ +\ (10/20)^2]$

$=\ 1\ -\ [1/4\ +\ 1/4]\ =\ 1 - 1/2 = 1/2$

$Gini\ index\ =\ 0.5$

b. *Compute the Gini index for the Customer ID attribute.*

$Total\ number\ of\ records\ =\ 20$

All Customer IDs are either C0 or C1, no overlaps because each customer ID is unique.

$Gini\ index\ for\ Customer\ ID\ =\ 1\ -\ [1^2 + 0^2]\ =\ 1 - 1 = 0$

$Gini\ index\ =\ 0$

c. *Compute the Gini index for the Gender attribute.*

$Total\ number\ of\ records\ =\ 20$

$P(Female)\ =\ 10/20$

$P(Male)\ =\ 10/20$

$P(Male|C0)\ =\ 6/10$

$P(Male|C1)\ =\ 4/10$

$P(Female|C0)\ =\ 4/10$

$P(Female|C1)\ =\ 6/10$

$Gini\ index\ for\ Male\ =\ 1\ -\ [(6/10)^2\ +\ (4/10)^2] = 48/100$ (same for female)

$Gini\ index\ =\ (1/2\ *\ 48/100)\ +\ (1/2\ *\ 48/100) = 24/100\ +\ 24/100 = 48/100$

$Gini\ index\ =\ 0.48$

d. *Compute the Gini index for the Car Type attribute using multiway split.*

$Total\ number\ of\ records\ =\ 20$

$P(Family)\ =\ 4/20$

$P(Sport)\ =\ 8/20$

$P(Luxury)\ =\ 8/20$

$P(Family|C0)\ =\ 1/4$

$P(Family|C1)\ =\ 3/4$

$P(Sport|C0)\ =\ 8/8$

$P(Sport|C1)\ =\ 0/8$

$P(Luxury|C0)\ =\ 1/8$

$P(Luxury|C1)\ =\ 7/8$

$Gini\ index\ for\ Family\ =\ 1\ -\ [(1/4)^2\ +\ (3/4)^2] = 6/16$

$Gini\ index\ for\ Sport\ =\ 1\ -\ [(8/8)^2\ +\ (0/8)^2] = 0$

$Gini\ index\ for\ Luxury\ =\ 1\ -\ [(1/8)^2 + (7/8)^2] = 14/64$
$Gini\ index\ =\ (4/20\ * 6/16) + (8/20\ * 14/64) = 6/80 + 14/160 = 26/160$
$Gini\ index\ =\ 0.1625$

e. *Compute the Gini index for the Shirt Size attribute using multiway split.*

$Total\ number\ of\ records\ =\ 20$
$P(Small)\ =\ 5/20$
$P(Medium)\ =\ 7/20$
$P(Large)\ =\ 4/20$
$P(Extra\ Large)\ =\ 4/20$
$P(Small|C0)\ =\ 3/5$
$P(Small|C1)\ =\ 2/5$
$P(Medium|C0)\ =\ 3/7$
$P(Medium|C1)\ =\ 4/7$
$P(Large|C0)\ =\ 2/4$
$P(Large|C1)\ =\ 2/4$
$P(Extra\ Large|C0)\ =\ 2/4$
$P(Extra\ Large|C1)\ =\ 2/4$
$Gini\ index\ for\ Small\ =\ 1\ -\ [(3/5)^2 + (2/5)^2] = 12/25$
$Gini\ index\ for\ Medium\ =\ 1\ -\ [(3/7)^2 + (4/7)^2] = 24/49$
$Gini\ index\ for\ Large\ =\ 1\ -\ [(2/4)^2 + (2/4)^2] = 8/16$ (same for Extra Large)
$Gini\ index\ =\ (5/20\ * 12/25) + (7/20\ * 24/49) + 2 * (4/20 * 1/2) = 12/100 + 24/140 + 1/5 = 0.4914$
$Gini\ index\ =\ 0.49142$

f. *Which attribute is better: Gender, Car Type, or Shirt Size?*

Car Type attribute is better because it gives me the lowest Gini Index, meaning the best split.

g. *Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.*

Customer ID has the lowest Gini index, but this doesn’t mean it should be used for splitting because every customer has a different ID to identify them. The purpose of this attribute to help identify the customer or keep track of how many customers there are, but it doesn’t help to predict any insightful information from the overall customer.

3. Consider the training examples shown in Table 3.6 for a binary classification problem.

Table 3.6. Data set for Exercise 3.

Instance	a1	a2	a3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+

3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

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

$$\text{Entropy: } H(Y) = -\sum_{i=0}^{c-1} P(Y = y_i)\ log_2 P(Y = y_i)$$

P(+) = 4/9

P(-) = 5/9

$Entropy\ =\ -\ [(4/9\ *log_2(4/9))\ +\ (5/9\ *log_2(5/9))]\ =\ -\ [-\ 0.1565255636\ -\ 0.2552725051]\ =\ 0.41$

$Entropy\ =\ 0.991076$

b. *What are the information gains of a1 and a2 relative to these training examples?*

P(+) = 4/9

P(-) = 5/9

a1’s info gain:

$P(a1\ =\ T|\ +)\ =\ 3/4$

$P(a1\ =\ F|\ +)\ =\ 1/4$

$P(a1\ =\ T|\ -)\ =\ 1/5$

$P(a1\ =\ F|\ -)\ =\ 4/5$

$Entropy(a1|\ +)\ =\ -\ [(3/4\ *log_2(3/4))\ +\ (1/4\ *log_2(1/4))]\ =\ 0.8112781245$

$Entropy(a1|\ -)\ =\ -\ [(1/5\ *log_2(1/5))\ +\ (4/5\ *log_2(4/5))]\ =\ 0.7219280949$

$Entropy(a1)\ =\ (4/9\ *\ 0.8112781245)\ +\ (5/9\ *\ 0.7219280949)\ =\ 0.7616392191$

$Information\ Gain\ (a1)\ =\ 0.991076\ -\ 0.7616392191\ =\ 0.2294$

a2’s info gain:

$P(a2\ =\ T|\ +)\ =\ 2/4$

$P(a2\ =\ F|\ +)\ =\ 2/4$

$P(a2\ =\ T|\ -)\ =\ 3/5$

$P(a2\ =\ F|\ -)\ =\ 2/5$

$Entropy(a2|\ +)\ =\ -\ [(2/4\ *log_2(2/4))\ +\ (2/4\ *log_2(2/4))]\ =\ 1$

$Entropy(a2|\ -)\ =\ -\ [(3/5\ *log_2(3/5))\ +\ (2/5\ *log_2(2/5))]\ =\ 0.9709505945$

$Entropy(a2)\ =\ (4/9\ *\ 1)\ +\ (5/9\ *\ 0.9709505945)\ =\ 0.9838614414$

$Information\ Gain\ (a2)\ =\ 0.991076\ -\ 0.9838614414\ =\ 0.0072145586$

c. For a3, which is a continuous attribute, compute the information gain for every possible split.

a3	Target Class label	Point of Split	Entropy	Information Gain
1.0	+	2.0	0.8484	0.1427
3.0	-	3.5	0.9885	0.0026
4.0	+	4.5	0.9183	0.0728
5.0 5.0	- -	5.5	0.9829	0.0072
6.0	+	6.5	0.9728	0.0183
7.0 7.0	+ -	7.5	0.8889	0.1022

The most optimal split for a3 happens when the point of split is at 2.0.

d. What is the best split (among a1, a2, and a3) according to the information gain?

Looking at the information gain of each split among a1, a2, and a3, a1 is the best split because it results in the biggest information gain.

e. What is the best split (between a1 and a2) according to the misclassification error rate?

The best split is a1 because its error rate is 2/9, while a2's error rate is 5/9.

f. What is the best split (between a1 and a2) according to the Gini index?

Looking at a1's and a2's Gini index, a1 is better because its Gini index is smaller.

- a1 Gini index

$\frac{4}{9} * (1 - \frac{3^2}{4} - \frac{1^2}{4}) + \frac{5}{9} * (1 - \frac{1^2}{5} - \frac{4^2}{5}) = 0.3444$

- a2 Gini index

$\frac{4}{9} * (1 - \frac{2^2}{4} - \frac{2^2}{4}) + \frac{5}{9} * (1 - \frac{3^2}{5} - \frac{2^2}{5}) = 0.4899$

5. Consider the following data set for a binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

a. Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

- overall entropy

$entropy = -[(4/10 * log_2(4/10)) + (6/10 * log_2(6/10))] = 0.9710$

- A's entropy

$entropy = -7/10[(4/7 * log_2(4/7)) + (3/7 * log_2(3/7))] + (3/10 * 0) = 0.6897$

- A's info gain

$Information\ Gain\ (A) = 0.9710 - 0.6897 = 0.2813$

- B's entropy

$entropy = -6/10[(1/6 * log_2(1/6)) + (5/6 * log_2(5/6))] - 4/10[(3/4 * log_2(3/4)) + (1/4 * log_2(1/4$

- B's info gain

$Information\ Gain\ (A) = 0.9710 - 0.7145 = 0.2564$

Based on the information gains calculated above, the decision tree induction algorithm would choose to split on A since it has a higher gain value.

b. Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

- overall Gini index

$Gini\ index\ =\ 1-\frac{4^2}{10}-\frac{6^2}{10})\ =\ 0.48$

- A's Gini index

$Gini\ index\ =\ \frac{7}{10}*(1-\frac{4^2}{7}-\frac{3^2}{7})\ +\ \frac{3}{10}*(1-\frac{0^2}{3}-\frac{3^2}{3})\ =\ 0.3429$

- A's Gini index gain

$Gini\ index\ Gain\ (A)\ =\ 0.48\ -\ 0.3429\ =\ 0.1371$

- B's Gini index

$Gini\ index\ =\ \frac{6}{10}*(1-\frac{1^2}{6}-\frac{5^2}{6})\ +\ \frac{4}{10}*(1-\frac{3^2}{4}-\frac{1^2}{4})\ =\ 0.3167$

- B's Gini index gain

$Gini\ index\ Gain\ (A)\ =\ 0.48\ -\ 0.3167\ =\ 0.1633$

Based on the Gini index gains calculated above, the decision tree induction algorithm would choose to split on B since it has a higher gain value.

c. Figure 3.11 shows that entropy and the Gini index are both monotonically increasing on the range [0, 0.5] and they are both monotonically decreasing on the range [0.5, 1]. Is it possible that information gain and the gain in the Gini index favor different attributes? Explain.
It is possible that information gain and the gain in the Gini index favor different attributes because as you can see from part (a) and (b) that they are based on the same dataset, but the Gini Index favored B while entropy favored A. This is due to how each method scales the frequencies, like the Gini index uses log.

1.2 (1 point) Tan, Chapter 4 Exercise 18. (Show your work, don't just provide the answer without showing how you derived it.)

18. Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains instances from two classes, "+" and "−." Half of the data set is used for training while the remaining half is used for testing.

a. Suppose there are an equal number of positive and negative instances in the data and the decision tree classifier predicts every test instance to be positive. What is the expected error rate of the classifier on the test data?

The expected error rate of the classifier on the test data is 0.5. This is because the value is binary, so if the classifier predicts every test instance to be positive, then the expected error rate is ½ = 0.5.

b. Repeat the previous analysis assuming that the classifier predicts each test instance to be positive class with probability 0.8 and negative class with probability 0.2.

- true positives: 8/10 * ½
- true negatives: 2/10 * ½
- false negatives: 2/10 * ½
- false positives: 8/10 * ½

error rate is $\frac{8+2}{10} = \frac{10}{20} = 0.5$

c. *Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test instance to be positive?*

The expected error rate of the classifier on the test data is ⅓ . This is because if it predicts every test instance as positive, it will misclassify every negative instance, so the error rate is the frequency of the negative instances, or ⅓.

d. *Repeat the previous analysis assuming that the classifier predicts each test instance to be positive class with probability 2/3 and negative class with probability 1/3.*

- true positives: ⅔ * ⅔
- true negatives: ⅓ * ⅓
- false negatives: ⅓ * ⅔
- false positives: ⅔ * 1/3

error rate is $\frac{2+2}{9} = \frac{4}{9} = 0.4444$