CS422-01: Homework #4

Vijay K. Gurbani

Sujad Patel

October 11, 2022

Part I
Exercises

1.1 Chapter 3

Exercise 2

2.(a) Compute the Gini index for the overall collection of training examples.

C0 class : 10, C1 class : 10
Total collection : 20

$$\Rightarrow 1 - (P(C0 \mid class)^2 + P(C1 \mid class)^2)$$
$$\Rightarrow 1 - \left(\frac{10}{20}\right)^2 - \left(\frac{10}{20}\right)^2$$
$$\Rightarrow 1 - 2 \times \left(\frac{1}{2}\right)^2$$
$$= 0.5$$

2.(b) Gini index for the overall collection of training examples is 0.5

2.(b) Compute the Gini index for the Customer ID *attribute.*

Since each customer ID is unique, the Gini index for each ID is 0, therefore the Gini index for Customer ID attribute is *0.*

2.(c) Compute the Gini index for the Gender *attribute.*

Male : 10, Female : 10
Gini Index for Male:

$$\Rightarrow 1 - (P(C0 \mid Male)^2 + P(C1 \mid Male)^2)$$
$$\Rightarrow 1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2$$
$$\Rightarrow \left(\frac{25 - 13}{25}\right)$$
$$\Rightarrow \frac{12}{25} = 0.48$$

Gini index for Female:

$$\Rightarrow 1 - (P(C0 \mid Female)^2 + P(C1 \mid Female)^2)$$
$$\Rightarrow 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2$$
$$\Rightarrow \frac{12}{25} = 0.48$$

Gini index for Gender attribute:

$$\left(\frac{Male}{Male+Female} \times Male\ gini\ index\right) + \left(\frac{Female}{Male+Female} \times Female\ gini\ index\right)$$

$$\Rightarrow \frac{10}{20} \times 0.48 + \frac{10}{20} \times 0.48$$

$$= 0.48$$

Gini Index for *Gender* attribute *0.48*

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

Family : 4, Sports : 8, Luxury : 8
Gini index for Family:

$$\Rightarrow 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = \left(\frac{10}{16}\right)^2 = 0.375$$

Gini index for Sports:

$$\Rightarrow 1 - \left(\frac{8}{8}\right)^2 - \left(\frac{0}{8}\right)^2 = 0$$

Gini index for Luxury:

$$\Rightarrow 1 - \left(\frac{1}{8}\right)^2 - \left(\frac{7}{8}\right)^2 = \left(\frac{14}{64}\right)^2 = 0.219$$

Gini index for *Car Type* attribute:

$$\Rightarrow \frac{4}{20} \times 0.375 + \frac{8}{20} \times 0 + \frac{8}{20} \times 0.219 = 0.16$$

Gini Index for *Car Type* attribute attribute *0.16*

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

Small : 5, Medium : 7, Large : 4, Extra Large : 4
Gini index for Small:

$$\Rightarrow 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = \left(\frac{12}{25}\right)^2 = 0.48$$

Gini index for Medium:

$$\Rightarrow 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = \left(\frac{24}{49}\right)^2 = 0.489$$

Gini index for Large:

$$\Rightarrow 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = \left(\frac{8}{16}\right)^2 = 0.5$$

Gini index for Extra Large:

$$\Rightarrow 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = \left(\frac{8}{16}\right)^2 = 0.5$$

Gini index for *Shirt Size* attribute:

$$\Rightarrow \frac{5}{20} \times 0.48 + \frac{7}{20} \times 0.489 + \frac{4}{20} \times 0.5 + \frac{4}{20} \times 0.5 = 0.4912$$

Gini Index for *Car Type* attribute attribute *0.49*

2.(f) Which attribute is better, Gender, Car Type*, or* Shirt Size*?*

Lower Gini index is preferred in making decision tree, because it represents that all the elements belong to a specific class. Therefore, *Car Type* is the best attribute.

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

Since each Customer ID is unique, including new ones, there is no reason to use it for prediction. Usually, a function creates these unique ID's.

Exercise 3

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

Positive(+) : 4, Negative(-) : 5
Entropy:

$$\Rightarrow -\left(\frac{4}{9}\right)\log_2\left(\frac{4}{9}\right) - \left(\frac{5}{9}\right)\log_2\left(\frac{5}{9}\right) = 0.9911$$

Entropy of the collection with training examples respect to class attribute is 0.991.

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

Entropy of *a1*:

a1	*T*	*F*
+	3	1
-	1	4

$$\Rightarrow -\frac{4}{9}\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) - \frac{5}{9}\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.7618$$

Information Gain for *a1*:

$$Gain(a1) = Entropy(collection) - Entropy(a1)$$

$$\Rightarrow Gain(a1) = 0.9911 - 0.7618$$

$$= 0.2292$$

Information Gain for a1 is 0.2292
Entropy of *a2*:

a2	*T*	*F*
+	2	2
-	3	2

$$\Rightarrow -\frac{4}{9}\left(\frac{2}{4}\log_2\frac{2}{4}+\frac{2}{4}\log_2\frac{2}{4}\right)-\frac{5}{9}\left(\frac{3}{5}\log_2\frac{3}{5}+\frac{2}{5}\log_2\frac{2}{5}\right)=0.9839$$

Information Gain for *a2*:

$$\Rightarrow Gain(a2) = 0.9911 - 0.9839$$
$$= 0.0072$$

Information Gain for a2 is 0.0072

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

**Gain = 0.9911 - Weighted Entropy(after split).*

Sorted Values	1	3	4	5	6	7
Splits (midpoints)	2	3.5	4.5	5.5	6.5	7
•	$\leq \mid >$	$\leq \mid >$	$\leq \mid >$	$\leq \mid >$	$\leq \mid >$	$\leq \mid >$
+	1|3	1|3	2|2	3|1	4|0	4|0
-	0|5	1|4	1|4	3|2	3|2	5|0
Entropy	0.8484	0.9885	0.9183	0.9839	0.9728	0.8889
Gain	0.1427	0.0026	0.0728	0.0072	0.0183	0.1022

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

Higher Information gain is preferred in making the best split, therefore *a1* has the highest Information gain.

3.(e) What is the best split (between a1 *and* a2 *) according to the miss-classification error rate?*

The best split a1 since its error rate is lower,$\left(\frac{2}{9} < \frac{4}{9}\right)$.

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

Gini for *a1*:

$$\Rightarrow \frac{4}{9}\left[1-\left(\frac{1}{4}\right)^2-\left(\frac{3}{4}\right)^2\right]+\frac{5}{9}\left[1-\left(\frac{4}{5}\right)^2-\left(\frac{1}{5}\right)^2\right]=0.3444$$

Gini for *a2*:

$$\Rightarrow \frac{4}{9}\left[1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right]+\frac{5}{9}\left[1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2\right]=0.4889$$

Since a1 has the lowest Gini, it is the best attribute to split.

Exercise 5

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

Entropy of all examples: Positives(+): 4, Negatives(-): 6

$$\Rightarrow -\left(\frac{4}{10}\right)\log_2\left(\frac{4}{10}\right) - \left(\frac{6}{10}\right)\log_2\left(\frac{6}{10}\right) = 0.9710$$

Entropy of *A*:

A	*T*	*F*
+	4	0
-	3	3

$$\Rightarrow -\frac{7}{10}\left(\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}\right) - \frac{3}{10}\left(\frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3}\right) = 0.6896$$

Information Gain for *A*:

$$\Rightarrow Gain(A) = 0.9710 - 0.6896$$
$$= 0.2816$$

Information Gain for A is 0.28

Entropy of *B*:

B	*T*	*F*
+	3	1
-	1	5

$$\Rightarrow -\frac{4}{10}\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) - \frac{6}{10}\left(\frac{1}{6}\log_2\frac{1}{6} + \frac{5}{6}\log_2\frac{5}{6}\right) = 0.7145$$

Information Gain for *B*:

$$\Rightarrow Gain(B) = 0.9710 - 0.7145$$
$$= 0.2565$$

Information Gain for B is 0.26

Information gain of A is higher than B, therefore, I would choose A attribute.

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

Gini of all examples:

$$\Rightarrow 1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 = 0.48$$

Gini of *A*:

$$\Rightarrow \frac{7}{10}\left[1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2\right] + \frac{3}{10}\left[1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2\right] = 0.3428$$

Gini gain of A = 0.48 - 0.3428 = 0.14 Gini of *B*:

$$\Rightarrow \frac{4}{10}\left[1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right]+\frac{6}{10}\left[1-\left(\frac{1}{6}\right)^2-\left(\frac{5}{6}\right)^2\right]=0.3166$$

Gini gain(X) = Gini(all) - Gini(X)
Gini gain of B = 0.48 - 0.3166 = 0.16
Since Gini gain of B is higher than A, attribute B will be chosen.

5.(c) The 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.

Yes, their gains are different, according to (a) and (b), therefore indicating that they use different metrics and favor different attributes.

1.2 Chapter 4

Exercise 18

18.(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?

There are equal number of instances from both classes. Therefore, Error rate is:

$$P(error) = P(error|+) \times (+)instances + P(error|-) \times (-)instances.$$

$$\Rightarrow (0.5 \times 0.5) + (0.5 \times 0.5)$$

$$P(error) = 0.50$$

18.(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.

$$\Rightarrow (0.8 \times 0.5) + (0.2 \times 0.5)$$

$$P(error) = 0.50$$

18.(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?

Since the negative class is also predicted as positive, the error rate will be negative/N. In this case, it is *0.33* (one-third of instances are negative).

18.(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.

$$\Rightarrow \left(0.33 \times \frac{2}{3}\right) + \left(0.67 \times \frac{1}{3}\right)$$

$$P(error) = 0.50$$

1.3 Multiclass Classification

Binary confusion Matrix

Iris-dataset	•	**References**		
•		*Setosa*	*Versicolor*	*Virginica*
Prediction	*Setosa*	10	0	0
Prediction	*Versicolor*	0	10	1
Prediction	*Virginica*	0	0	9

Q. Compute the Sensitivity, Specificity, and Precision for each class.

Setosa Matrix	•	**References**	
•		*Setosa*	*Versicolor, Virginica*
Prediction	*Setosa*	10	0
Prediction	*Versicolor, Virginica*	0	20

Setosa:
TP = 10, TN = 20, FN = 0, FP = 0.

$$Sensitivity = \left(\frac{TP}{TP + FN}\right) \Rightarrow \frac{10}{10} = 1.0$$

$$Specificity = \left(\frac{TN}{TN + FN}\right) \Rightarrow \frac{20}{20} = 1.0$$

$$Precision = \left(\frac{TP}{TP + FP}\right) \Rightarrow \frac{10}{10} = 1.0$$

Versicolor Matrix	•	**References**	
•		*Versicolor*	*Setosa, Virginica*
Prediction	*Versicolor*	10	1
Prediction	*Setosa, Virginica*	0	19

Versicolor:
TP = 10, TN = 19, FN = 0, FP = 1.

$$Sensitivity = \frac{10}{10} = 1.0$$

$$Specificity = \frac{19}{19} = 1.0$$

$$Precision = \frac{10}{11} = 0.91$$

Virginica Matrix	•	**References**	
•		*Virginica*	*Setosa, Versicolor*
Prediction	*Virginica*	9	0
Prediction	*Setosa, Versicolor*	1	20

Virginica:
TP = 9, TN = 20, FN = 1, FP = 0.

$$Sensitivity = \frac{9}{10} = 0.90$$

$$Specificity = \frac{20}{21} = 0.95$$

$$Precision = \frac{9}{9} = 1.0$$