Tania Soutonglang

CS 422-01

October 10, 2022

Homework 4

1.1 Tan, Chapter 3

1.1.1 Exercise 2

Consider the training examples shown in Table 3.5 for a binary classification problem.

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

$$1 - \left[\left(\frac{10\ males}{20\ people}\right)^2 + \left(\frac{10\ females}{20\ people}\right)^2\right] = 1 - (0.25 + 0.25) = 1 - 0.5 = \mathbf{0.5}$$

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

$$1 - \left[\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right] = 1 - (0 + 1) = 1 - 1 = \mathbf{0}$$

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

$$males \rightarrow 1 - \left[\left(\frac{6\ C0}{10\ males}\right)^2 + \left(\frac{4\ C1}{10\ males}\right)^2\right] = 1 - (0.36 + 0.16) = 0.48$$

$$females \rightarrow 1 - \left[\left(\frac{4\ C0}{10\ females}\right)^2 + \left(\frac{6\ C1}{10\ females}\right)^2\right] = 1 - (0.16 + 0.36) = 0.48$$

$$overall \rightarrow 0.48\left(\frac{10\ males}{20\ people}\right) + 0.48\left(\frac{10\ females}{20\ people}\right) = 0.24 + 0.24 = \mathbf{0.48}$$

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

$$family \rightarrow 1 - \left[\left(\frac{1\ C0}{4\ family\ cars}\right)^2 + \left(\frac{3\ C1}{4\ family\ cars}\right)^2\right] = 1 - (0.0625 + 0.5625) = 1 - 0.625 = 0.375$$

$$sports \rightarrow 1 - \left[\left(\frac{8\ C0}{8\ sports\ cars}\right)^2 + \left(\frac{0\ C1}{8\ sports\ cars}\right)^2\right] = 1 - (1 + 0) = 1 - 1 = 0$$

$$luxury \rightarrow 1 - \left[\left(\frac{1\ C0}{8\ luxury\ cars}\right)^2 + \left(\frac{7\ C1}{8\ luxury\ cars}\right)^2\right] = 1 - (0.015625 + 0.765625) = 1 - 0.78125 = 0.21875$$

$$overall \rightarrow 0.375\left(\frac{4\ family}{20\ cars}\right) + 0\left(\frac{8\ sports}{20\ cars}\right) + 0.21875\left(\frac{8\ luxury}{20\ cars}\right) = 0.075 + 0 + 0.0875 = 0.1625 = \mathbf{0.16}$$

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

$$small \rightarrow 1 - \left[\left(\frac{3\ C0}{5\ small}\right)^2 + \left(\frac{2\ C1}{5\ small}\right)^2\right] = 1 - (0.36 + 0.16) = 0.48$$

$$medium \rightarrow 1 - \left[\left(\frac{3\ C0}{7\ medium}\right)^2 + \left(\frac{4\ C1}{7\ medium}\right)^2\right] = 1 - (0.18367 + 0.32653)$$
$$= 1 - 0.51020 = 0.48980$$
$$large \rightarrow 1 - \left[\left(\frac{2\ C0}{4\ large}\right)^2 + \left(\frac{2\ C1}{4\ large}\right)^2\right] = 1 - (0.25 + 0.25) = 1 - 0.50$$
$$= 0.50$$
$$extra\ large \rightarrow 1 - \left[\left(\frac{2\ C0}{4\ xl}\right)^2 + \left(\frac{2\ C1}{4\ xl}\right)^2\right] = 1 - (0.25 + 0.25) = 1 - 0.50$$
$$= 0.50$$
$$overall \rightarrow 0.48\left(\frac{5\ small}{20\ shirts}\right) + 0.48980\left(\frac{7\ medium}{20\ shirts}\right) + 0.50\left(\frac{4\ large}{20\ shirts}\right)$$
$$+ 0.50\left(\frac{4\ extra\ large}{20\ shirts}\right) = 0.12 + 0.17143 + 0.10 + 0.10$$
$$= 0.49143 = \mathbf{0.49}$$

f. Which attribute is better, **Gender**, **Car Type**, or **Shirt Size**?
Car Type because it has the lowest Gini Index.

g. Explain why **Customer ID** should not be used as the attribute test condition even though it has the lowest Gini.
Since every customer gets a new ID, there is no point in using Customer ID has a sorting attribute.

1.1.2 Exercise 3

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

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

+	-
4	5

$$E(total) = -\left(\frac{4}{9}\log_2\frac{4}{9} + \frac{5}{9}\log_2\frac{5}{9}\right) = -(-0.51997 + -0.47111)$$
$$= -(-0.99108) = \mathbf{0.99}$$

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

	+	-
T	3	1
F	1	4

$$E(a1) = -\frac{4}{9}\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) + -\frac{5}{9}\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right)$$
$$= -\frac{4}{9}(-0.81128) + -\frac{5}{9}(-0.72193) = 0.36057 + 0.40107$$
$$= 0.76164$$
$$information\ gain(a1) = 0.99108 - 0.76164 = \mathbf{0.23}$$

	+	-

T	2	3
F	2	2

$$E(a2) = -\frac{5}{9}\left(\frac{2}{5}\log_2\frac{2}{5}+\frac{3}{5}\log_2\frac{3}{5}\right) + -\frac{4}{9}\left(\frac{2}{4}\log_2\frac{2}{4}+\frac{2}{4}\log_2\frac{2}{4}\right)$$

$$= -\frac{5}{9}(-0.97095) + -\frac{4}{9}(-1) = 0.53942 + 0.44444 = 0.98386$$

$$information\ gain(a2) = 0.99108 - 0.98386 = \mathbf{0.0072}$$

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

instance	a3	class	split-point	entropy	information gain
1	1.0	+			
6	3.0	-	2.0	0.8484	0.1427
4	4.0	+	3.5	0.9885	0.0026
3	5.0	-	4.5	0.9183	0.0728
9	5.0	-	0.0	0.9839	0.0072
2	6.0	+	5.5	0.9839	0.0072
5	7.0	-	6.5	0.9728	0.0183
8	7.0	+	0.0	0.9728	0.0183
7	8.0	-	7.5	0.8889	0.1022

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

a1 has the best split since it has the highest information gain.

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

a1 error rate: 2/9

a2 error rate: 4/9

a1 would have the better split.

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

$$a1 \rightarrow \frac{4}{9}\left[1-\left(\frac{3^2}{4}+\frac{1^2}{4}\right)\right]+\frac{5}{9}\left[1-\left(\frac{1^2}{5}+\frac{4^2}{5}\right)\right] = \frac{4}{9}\left(1-\frac{5}{8}\right)+\frac{5}{9}\left(1-\frac{17}{25}\right)$$

$$= \frac{4}{9}\left(\frac{3}{8}\right)+\frac{5}{9}\left(\frac{8}{25}\right) = \frac{1}{6}+\frac{8}{45} = \frac{31}{90} = \mathbf{0.34}$$

$$a2 \rightarrow \frac{4}{9}\left[1-\left(\frac{2^2}{4}+\frac{2^2}{4}\right)\right]+\frac{5}{9}\left[1-\left(\frac{2^2}{5}+\frac{3^2}{5}\right)\right] = \frac{4}{9}\left(1-\frac{1}{2}\right)+\frac{5}{9}\left(1-\frac{13}{25}\right)$$

$$= \frac{4}{9}\left(\frac{1}{2}\right)+\frac{5}{9}\left(\frac{12}{25}\right) = \frac{2}{9}+\frac{4}{15} = \frac{22}{45} = \mathbf{0.49}$$

a1 would have the best split since the Gini index is smaller

1.1.3 Exercise 5

Consider the following data set for a binary class problem.

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

$$E(total) = -\left(\frac{4}{10}\log_2\frac{4}{10} + \frac{6}{10}\log_2\frac{6}{10}\right) = -(-0.52877 + -0.44218)$$
$$= -(-0.97095) = 0.97$$

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

$$E(A = T) = -\frac{7}{10}\left(\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}\right) = -\frac{7}{10}(-0.46135 + -0.52388)$$
$$= -\frac{7}{10}(-0.98523) = 0.69$$
$$E(A = F) = -\frac{3}{10}\left(\frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3}\right) = -\frac{3}{10}(0 + 0) = 0$$
$$information\ gain(A) = 0.97 - (0.69 + 0) = \mathbf{0.28}$$

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

$$E(B = T) = -\frac{4}{10}\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = -\frac{4}{10}(-0.31128 \pm 0.50)$$
$$= -\frac{4}{10}(-0.81128) = 0.32$$
$$E(B = F) = -\frac{6}{10}\left(\frac{1}{6}\log_2\frac{1}{6} + \frac{5}{6}\log_2\frac{5}{6}\right) = -\frac{6}{10}(-0.43083 \pm 0.21920)$$
$$= -\frac{6}{10}(-0.65002) = 0.39$$
$$information\ gain(B) = 0.97 - (0.32 + 0.39) = \mathbf{0.26}$$

The split should be on attribute A.

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

$$total \rightarrow 1 - \left(\frac{4}{10}^2 + \frac{6}{10}^2\right) = 1 - \left(\frac{4}{25} + \frac{9}{25}\right) = 1 - \frac{13}{25} = \frac{12}{25} = 0.48$$

$$A = T \rightarrow 1 - \left(\frac{4^2}{7} + \frac{3^2}{7}\right) = 1 - \left(\frac{16}{49} + \frac{9}{49}\right) = 1 - \frac{25}{49} = 0.48980$$

$$A = F \rightarrow 1 - \left(\frac{0^2}{3} + \frac{3^2}{3}\right) = 1 - (0 + 1) = 1 - 1 = 0$$

$$Gini\ gain \rightarrow 0.48 - \left[\frac{7}{10}(0.48980) + \frac{3}{10}(0)\right] = 0.48 - 0.34286 = \mathbf{0.14}$$

$$B = T \rightarrow 1 - \left(\frac{3^2}{4} + \frac{1^2}{4}\right) = 1 - \left(\frac{9}{16} + \frac{1}{16}\right) = 1 - \frac{5}{8} = 0.375$$

$$B = F \rightarrow 1 - \left(\frac{1^2}{6} + \frac{5^2}{6}\right) = 1 - \left(\frac{1}{36} + \frac{25}{36}\right) = 1 - \frac{13}{18} = 0.27778$$

$$Gini\ gain \rightarrow 0.48 - \left[\frac{4}{10}(0.375) + \frac{6}{10}(0.27778)\right] = 0.48 - (0.15 + 0.16667)$$
$$= 0.48 - 0.31667 = \mathbf{0.16}$$

The node should be split on B.

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's possible that the information gain and the gain in the Gini index favors different attributes even though the measures have similar range and monotonous behavior.

1.2 Tan, Chapter 4, Exercise 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?

$$P(error_+) = 0.50$$
$$P(+) = 0.50$$
$$P(error_-) = 0.50$$
$$P(-) = 0.50$$
$$P(error) = 0.50(0.50) + 0.50(0.50) = \mathbf{0.50}$$

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.

$$P(error_+) = 0.20$$
$$P(+) = 0.50$$
$$P(error_-) = 0.80$$
$$P(-) = 0.50$$
$$P(error) = 0.20(0.50) + 0.80(0.50) = \mathbf{0.50}$$

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?

$$P(error_+) = \frac{1}{3}$$
$$P(+) = 0.50$$
$$P(error_-) = \frac{2}{3}$$
$$P(-) = 0.50$$

$$P(error) = \frac{1}{3}(0.50) + \frac{2}{3}(0.50) = \mathbf{0.50}$$

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.

$$P(error_+) = \frac{1}{3}$$
$$P(+) = \frac{2}{3}$$
$$P(error_-) = \frac{2}{3}$$
$$P(-) = \frac{1}{3}$$
$$P(error) = \frac{1}{3}\left(\frac{2}{3}\right) + \frac{2}{4}\left(\frac{1}{3}\right) = \frac{4}{9} = \mathbf{0.44}$$

1.3 Multiclass Classification

		Actual		
		Setosa	**Versicolor**	**Virginica**
Predicted	**Setosa**	8	0	0
	Versicolor	0	10	1
	Virginica	0	2	9

Versicolor		Actual	
		Versicolor	**Setosa, Virginica**
Predicted	**Versicolor**	10	1
	Setosa, Virginica	2	17

$$Sensitivity = \frac{10}{12} = \mathbf{0.83}$$
$$Specitivity = \frac{17}{18} = \mathbf{0.94}$$
$$Precision = \frac{10}{11} = \mathbf{0.91}$$

Setosa		Actual	
		Setosa	**Versicolor, Virginica**
Predicted	**Setosa**	8	0
	Versicolor, Virginica	0	22

$$Sensitivity = \frac{8}{8} = \mathbf{1.0}$$
$$Specitivity = \frac{22}{22} = \mathbf{1.0}$$

$$Precision = \frac{8}{8} = \mathbf{1.0}$$

Virginica		Actual	
		Virginica	Setosa, Versicolor
Predicted	Virginica	9	2
	Setosa, Versicolor	1	18

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

$$Specitivity = \frac{18}{20} = \mathbf{0.90}$$

$$Precision = \frac{9}{11} = \mathbf{0.82}$$