Javier-EchavarrenSuarez-HW3

1.1 Tan, Chapter 3

Exercise 2, 3, 5.

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

Table 3.5. Data set for Exercise 3.

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

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

$$Gini\ Index := 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

$$Gini\ Index_a = 1 - \sum_{i=0}^{c-1} p_i(t)^2 = 1 - \sum_{i=0}^{2} p_i(t)^2 = 1 - 2 \cdot (1/2)^2 = 1 - 2 \cdot 0.25 = 0.5$$

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

We take the mean of the Gini index for each Customer ID

$$Gini\ Index_{Customer\ ID,i} = 1 - \sum_{i=0}^{c-1} p_i(t)^2 = 1 - \sum_{i=0}^{2} p_i(t)^2 = 1 - (0)^2 - (1)^2 = 1 - 1 = 0$$

As they are all 0, the Gini Index for the Customer ID attribute is also 0.

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

$$Gini\ Index_{Gender,M} = 1 - \sum_{i=0}^{c-1} p_i(t)^2 = 1 - \sum_{i=0}^{2} p_i(t)^2 = 1 - (0.6)^2 - (0.4)^2$$
$$= 1 - 0.36 - 0.16 = 0.48$$

$$Gini\ Index_{Gender,F} = 1 - \sum_{i=0}^{c-1} p_i(t)^2 = 1 - \sum_{i=0}^{2} p_i(t)^2 = 1 - (0.6)^2 - (0.4)^2$$
$$= 1 - 0.36 - 0.16 = 0.48$$

$$Gini\ Index_{Gender} = \frac{1}{2} \cdot (Gini\ Index_{Gender,M} + Gini\ Index_{Gender,F}) = \frac{1}{2}(0.48 + 0.48)$$
$$= 0.48$$

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

$$Gini\ Index_{Car\ Type,Family} = 1 - \sum_{i=0}^{c-1} p_i(t)^2 = 1 - \sum_{i=0}^{2} p_i(t)^2 = 1 - (0.25)^2 - (0.75)^2$$
$$= 0.375$$

$$Gini\ Index_{Car\ Type,Sports} = 1 - \sum_{i=0}^{c-1} p_i(t)^2 = 1 - \sum_{i=0}^{2} p_i(t)^2 = 1 - (1)^2 - (0)^2 = 0$$

$$Gini\ Index_{Car\ Type,Luxury} = 1 - \sum_{i=0}^{c-1} p_i(t)^2 = 1 - \sum_{i=0}^{2} p_i(t)^2 = 1 - (0.125)^2 - (0.875)^2$$
$$= 0.21875$$

$$Gini\ Index_{Car\ Type} = \sum_{i=1}^{s} \frac{n_s}{n_T} \cdot Gini\ Index_s$$

$$Gini\ Index_{Car\ Type} = \frac{4}{20} \cdot Gini\ Index_{Cartype,Family} + \frac{8}{20} \cdot Gini\ Index_{Cartype,Sports} + \frac{8}{20}$$
$$\cdot Gini\ Index_{Cartype,Family}$$

$$Gini\ Index_{Car\ Type} = \frac{4}{20} \cdot 0.375 + \frac{8}{20} \cdot 0 + \frac{8}{20} \cdot 0.21875 = 0.1625$$

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

$$Gini\ Index_{Shirt\ Size,Small} = 1 - \sum_{i=0}^{c-1} p_i(t)^2 = 1 - \sum_{i=0}^{2} p_i(t)^2 = 1 - (0.6)^2 - (0.4)^2 = 0.48$$

$$Gini\ Index_{Shirt\ Size,Medium} = 1 - \sum_{i=0}^{c-1} p_i(t)^2 = 1 - \sum_{i=0}^{2} p_i(t)^2 = 1 - (0.4286)^2 - (0.5714)^2$$
$$= 0.489$$

$$Gini\ Index_{Shirt\ Size,Large} = 1 - \sum_{i=0}^{c-1} p_i(t)^2 = 1 - \sum_{i=0}^{2} p_i(t)^2 = 1 - (0.5)^2 - (0.5)^2 = 0.5$$

$$Gini\ Index_{Shirt\ Size,Extra\ Large} = 1 - \sum_{i=0}^{c-1} p_i(t)^2 = 1 - \sum_{i=0}^{2} p_i(t)^2 = 1 - (0.5)^2 - (0.5)^2$$
$$= 0.5$$

$$Gini\ Index_{Shirt\ Size} = \sum_{i=1}^{s} \frac{n_s}{n_T} \cdot Gini\ Index_s$$

$$Gini\ Index_{Shirt\ Size} = \frac{5}{20} \cdot 0.48 + \frac{7}{20} \cdot 0.489 + \frac{4}{20} \cdot 0.5 + \frac{4}{20} \cdot 0.5 = 0.4914$$

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

The best attribute to make the classification is Car type, as it has the lowest Gini.

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

It shouldn't be used as the attribute test for two reasons:

First, because it is logical that Customer ID has no relation to the studied variable and won't serve as a predictor.

Second, because the leaves containing each of the Customer IDs only contain one sample, and it is too few in order to consider the model reliable.

1.2 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?

$$\boldsymbol{Error\ rate = \frac{number\ of\ predicting\ errors}{number\ of\ predictions}}$$

According to this case, the number of predicting errors will be $\frac{n}{2}$, and the number of predictions will be n, and that way, our Error rate is 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.

In this case, we must use the same equation, but taking into account the different cases

$$P(prediction = -|Attribute = +) = 0.2$$

$$P(prediction = +|Attribute = -) = 0.8$$

$$P(Attribute = +) = P(Attribute = -) = 0.5$$

Using Baye's theorem

$$Error\ rate\ = 0.2 \cdot 0.5 + 0.8 \cdot 0.5 = 0.5$$

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 errors are going to occur when the instances contain the negative class, 1/3 of the time, because if that happens, the predictor will predict positive. On the contrary, when the data label is +, the predictor will predict positive and it will not count as error.

The expected will be 1/3=0.333

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.

We have to discuss the different cases as we did in b.

$$P(prediction = -|Attribute = +) = 1/3$$

$$P(prediction = +|Attribute = -) = 2/3$$

$$P(Attribute = +) = \ 2/3$$

$$P(Attribute = -) = \ 1/3$$

Using Baye's theorem

$$Error\ rate\ = 1/3 \cdot 2/3 + 2/3 \cdot 1/3 = 4/9 = \ 0.444$$

1.3 Multiclass classification

Using the confusion matrix from multiclass.Rmd notebook (from Lecture 7), create a binary-class confusion matrix using the "one-vs-many" strategy for each class. Then, for each class, compute the sensitivity, specificity and precision to two decimal places. Show all work, including the binary class confusion matrices.

We compute the binary-class confusion matrix from the multiclass confusion matrix and we use the following equations to compute the sensitivity, specificity and precision. We show the calculations made in excel in the following figure:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1				Reference										
2			setosa	versicolor	virginica									
3	Predictio	setosa	10	0	0									
4		versicolor	0	10	1									
5		virginica	0	0	9									
6														
7														
8			Confusion Matrix					Confusion Matrix					Confusion Matrix	
9			Actual Class					Actual Class					Actual Class	
10		Setosa	Positive	Negative		Versicolor		Positive	Negative		Virginica		Positive	Negative
11	Predicted Class	Positive	=C3	=SUM(D3:E3)		Predicted Class	Positive	=D4	=C4+E4		Predicted Class	Positive	=E5	=SUM(C5:D5)
12		Negative	=SUM(C4:C5)	=SUM(D4:E5)			Negative	=D3+D5	=C3+E3+C5+E5			Negative	=E3+E4	=SUM(C3:D4)
13														
14		Sensitivity	=C11/(C11+C12)				Sensitivity	=H11/(H11+H12)				Sensitivity	=M11/(M11+M12)	
15		Specificity	=D12/(D11+D12)				Specificity	=I12/(I11+I12)				Specificity	=N12/(N11+N12)	
16		Precision	=C11/(C11+D11)				Precision	=H11/(H11+I11)				Precision	=M11/(M11+N11)	
17														

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

$$Sensitivity = \frac{TP}{TP + FP}$$

Where TP= True positives, TN= True negatives, FP= False positives, FN= False negatives

		Confusion Matrix: Setosa	
Setosa		Actual Class	
		Positive	Negative
Predicted Class	Positive	10	0
	Negative	0	20
	Sensitivity	1,00	
	Specificity	1,00	
	Precision	1,00	

		Confusion Matrix: Versicolor	
Versicolor		Actual Class	
		Positive	Negative
Predicted Class	Positive	10	1
	Negative	0	19
	Sensitivity	1,00	
	Specificity	0,95	
	Precision	0,91	

		Confusion Matrix: Virginica	
		Actual Class	

Virginica		Positive	Negative
Predicted Class	Positive	9	0
	Negative	1	20
	Sensitivity	0,90	
	Specificity	1,00	
	Precision	1,00	