Homework 4 - CS 422 Data Mining

Julen Ferro

October 12, 2022

1 Introduction

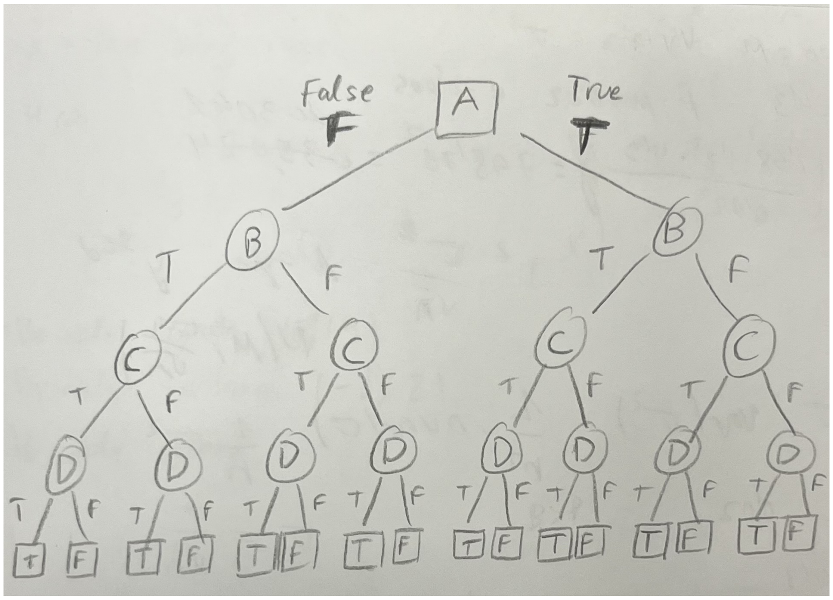
This will be the report for Assignment number 4. It can be noticed that there are some extra exercises performed due to a misunderstanding with different versions of the quoted book. They have been useful in order to catch up with some new concepts.

2 Chapter 3 - Exercises

2.1 Exercise 2

Draw a full decision tree for the odd parity function, where only when the count of True is odd is the class label True, of four Boolean attributes A, B, C, and D. Is it possible to simplify the tree?

A	B	C	D	Class
T	T	T	T	F
T	T	T	F	T
T	T	F	T	T
T	T	F	F	F
T	F	T	T	T
T	F	T	F	F
T	F	F	T	F
T	F	F	F	T
F	T	T	T	T
F	T	T	F	F
F	T	F	T	F
F	T	F	F	T
F	F	T	T	F
F	F	T	F	T
F	F	F	T	T
F	F	F	F	F



2.2 Exercise 2, THE GRADED ONE

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 - $0.5^2 - 0.5^2 = 0.5$

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

The gini for each of the Customer IDvalues is 0, so the sum of them is also 0.

(c) Compute the Gini index for the Gender attribute

The Gini for Male is $1((4/10)^2 + (6/10)^2) = 0.48$
$TheGiniforFemaleisalso1((4/10)^2 + (6/10)^2) = 0.48.$
$Therefore, theoverallGiniforGenderis0.480.5 + 0.480.5 = 0.48.$

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

The gini for Familycar is 1- $(1/4)^2 - (3/4)^2 = 0.375.$
$Sportscaris0becausealltheclassificationispure(thereareonlyCoclasses).$
$Luxurycaris1 - ((1/8)^2 + (7/8)^2) = 0.2188.$
$Theoverallginiis0.375 * 4/20 + 0 * 8/20 + 0.21888/20 = 0.1625$

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

The gini forSmallshirt size is 1 - $((3/5)^2 + (2/5)^2) = 0.48.$
$Mediumshirtsizeis1 - ((3/7)^2 + (4/7)^2) = 0.4898.$
$Largeshirtsizeis0.5(equallydistributedCoandC1).$
$ExtraLargeshirtsizeis0.5.$
$TheoverallginiforShirtSizeattributeis(5 * 0.48 + 7 * 0.489 + 4 * 0.5 + 4 * 0.5)/20 = 0.49$

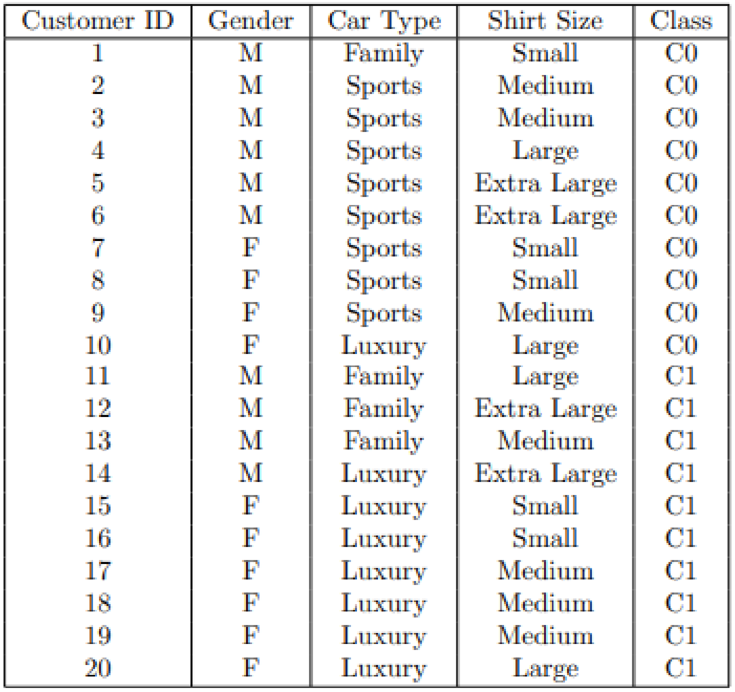
(f) Which attribute is better, Gender, Car Type, or Shirt Size?

The Car Type is the best one because it has the lowest gini among the three attributes. That is because the Co and C1 are classified in a purest way in terms of the attribute Car Type, having one car or another means more than the other attributes in order to pertain to the class 1 or 0.

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

The attribute Customer ID does not serve for the decision tree due to the fact that they will not be two data samples with same ID. That is to say, unlike in the case of two samples having the same kind of cars, every new sample will be assigned a new ID, so it is impossible to find a correlation between a value of the ID and the Class.

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

2.3 Exercise 3, THE GRADED ONE

Consider the training examples shown in Table 4.2 for a binary classification problem. (a) What is the entropy of this collection of training examples with respect to the positive class?

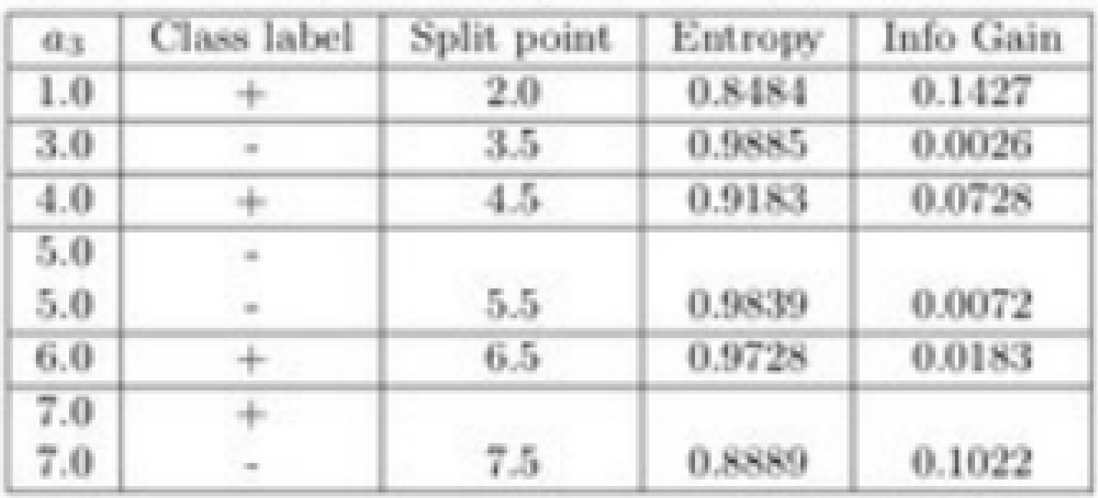
We have 4 positive examples and 5 negative examples out of 4+5 = 9 examples. Hence the probabilities are as follows: P(+) = 4/9 P(-) = 5/9.
Therefore the total entrophy is:
E = -(4/9 * log2 (4/9) + 5/9 * log2 (5/9) = 0.99

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

E(a1) = 4/9 * (-(3/4)*log2(3/4)-(1/4)*log2(1/4)) + 5/9 * (-(1/5)*log2(1/5)-(4/5)*log2(4/5)) = 0.7616.
Hence the IG = 0.99 - 0.76 ? 0.23
E(a2) = 5/9 * (-(2/5)*log2(2/5)-(3/5)*log2(3/5)) + 4/9 * (-(2/4)*log2(2/4)-(2/4)*log2(2/4)) = 0.98.
Hence the IG = 0.99 - 0.98 = 0.0072
We should choose E(a1) as it has a greater IG, due to the lower value of E(a1). More pure. Less impure.

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

The best split for a 3 occurs at split point equals to 2; see table below.



a_3	Class label	Split point	Entropy	Info 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	0.9839	0.0072
6.0	+	6.5	0.9728	0.0183
7.0	+			
7.0	-	7.5	0.8889	0.1022

As we have n different values for a3, we make (n-1) different split boundary points estimations for a3. After having perform all tha calculations, we decide that the best boundary for splitting a3 is the a3 = 2.0 because has the smaller entrophy value, that is to say, sorts better the positive and negative cases in different leaves.

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

a1 produces the best split.

e) What is the best split (between a 1 and a 2) according to the classification error rate?

a 1 : error rate = 2/9.
a 2 : error rate = 5/9.
a 1 produces the best split.

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

Gini a1 = 0.34
Gini a2 = 0.49
As before calculated with the entrophy, a1 is a better split.

2.4 Exercise 5 - THE GRADED ONE

Q5. 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 origin = -0.4*log2(0.4)-0.6*log0.6 = 0.971

IG = E origin - E(A) = 0.97 - (7/10) * 0.98 - 0 = 0.28

IG = E origin - E(B) = 0.97 - (4/10) * 0.81 - 0.6 * 0.65 = 0.25

A will be chosen because it has a greater IG.

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

If we do the same exercise as (a), but calculating the impurity by using the Gini indexes we get the next Informatin Gains.

IG(A) = 0.137

IG(B) = 0.16

Therefore, in this case we will choose B. The Gini index says B , whereas the entrophy tells us to split in terms of A. This is really interesting. Two variables that measure the same (impurity) by using different analytical expressions.

(c) Figure 4.13 shows that entropy and the Gini index are both monotonously increasing on the range [0, 0.5] and they are both monotonously 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, even though these measures have similar range and monotonous behavior, their respective gains, , which are scaled differences of the measures, do not necessarily behave in the same way, as illustrated by the results in parts (a) and (b).

a and b serve as examples of that statement. Even though Gini and Entrophy measure the same property, impurity, for the same input data could end up giving different insights. That is due to the difference in the mathematical expressions, which weights differently some events than others, as for instance, having in one of the subtree isolated obersvations as it happens in the A True and False observations. The log entrophy function gives more importance to this event for example, than the Gini index does.

3 Chapter 4 - Exercises

3.1 Exercise 18 - THE GRADED ONE

Q18. 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 records 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 records in the data and the decision tree classifier predicts every test record to be positive. What is the expected error rate of the classifier on the test data?

(P(error) = P(error—+) * P(+) + P(error—-) * P(-) = 0.50 * 0.50 + 0.50 * 0.50 = 0.50)

Or looking at the big picture. We will get half positive and half negative classes, and the prediction will always be positive, so we will fail the half of the times. 50 percent.

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

(P(error) = P(error — +) P(+) + P(error— -) P(-) = 0.2 * 0.5 + 0.8 * 0.5 = 0.1 + 0.4 = 0.5

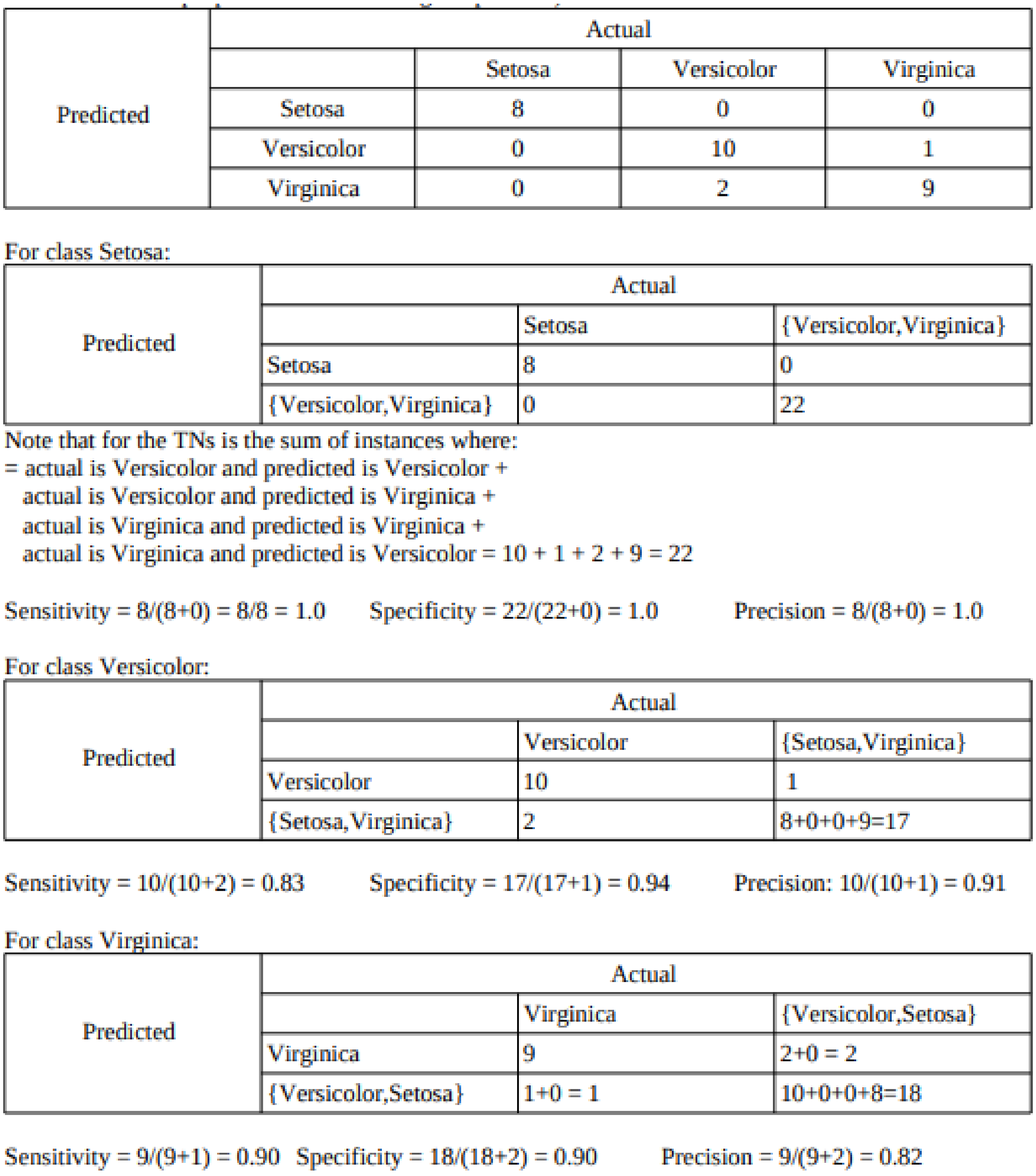
(c) Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test record to be positive?

0.66 of being correct. Hence, 0.33 * 100 = 33 percent of failing.

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

(P(error) = P(error—+) * P(+) + P(error—-) * P(-) = (0.33 *0.67)+(0.67*0.33) = 0.4422 = 44 percent.

4 Exercise 1.3 - Multiclass - THE GRADED ONE



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

For class Setosa:

		Actual	
		Setosa	{Versicolor,Virginica}
Predicted	Setosa	8	0
	{Versicolor,Virginica}	0	22

Note that for the TNs is the sum of instances where:
= actual is Versicolor and predicted is Versicolor +
actual is Versicolor and predicted is Virginica +
actual is Virginica and predicted is Virginica +
actual is Virginica and predicted is Versicolor = 10 + 1 + 2 + 9 = 22

Sensitivity = 8/(8+0) = 8/8 = 1.0 Specificity = 22/(22+0) = 1.0 Precision = 8/(8+0) = 1.0

For class Versicolor:

		Actual	
		Versicolor	{Setosa,Virginica}
Predicted	Versicolor	10	1
	{Setosa,Virginica}	2	8+0+0+9=17

Sensitivity = 10/(10+2) = 0.83 Specificity = 17/(17+1) = 0.94 Precision: 10/(10+1) = 0.91

For class Virginica:

		Actual	
		Virginica	{Versicolor,Setosa}
Predicted	Virginica	9	2+0 = 2
	{Versicolor,Setosa}	1+0 = 1	10+0+0+8=18

Sensitivity = 9/(9+1) = 0.90 Specificity = 18/(18+2) = 0.90 Precision = 9/(9+2) = 0.82

The figure given by the professor in which the tables are plotted has been taken in order to let the student spend time in the explanation of the exercise, rather than in trying to write down a Latex table. Thank you very much for the resources.

Explanation:

Given the first table in which we can find the confussion matrix taking into account the three species independently, we can derive other three tables in which in each one, one out of the three elements will be isolated, and the other two ones will be trated as other factor (not two).

Second table: isolating SETOSA. Sensitivity = 8/(8+0) = 8/8 = 1.0 Specificity = 22/(22+0) = 1.0 Precision = 8/(8+0) = 1.0

The key for this exercise is the next one. For instance, in this second table, SETOSAa is isolated, so the same values of ACTUAL and PREDICTED for SETOSA are taken from teh first table. Now, Virginica and Versicolor are taken as a same group, so even if the classification was mistaken, if the prediction was not SETOSA, those misclassifications will not be taken as ERRORs like FALSE VERSICOLOR or FALSE VIRGINICA.

Third table: isolating VERSICOLOR. Sensitivity = 10/(10+2) = 0.83 Specificity = 17/(17+1) = 0.94 Precision = 10/(10+1) = 0.91

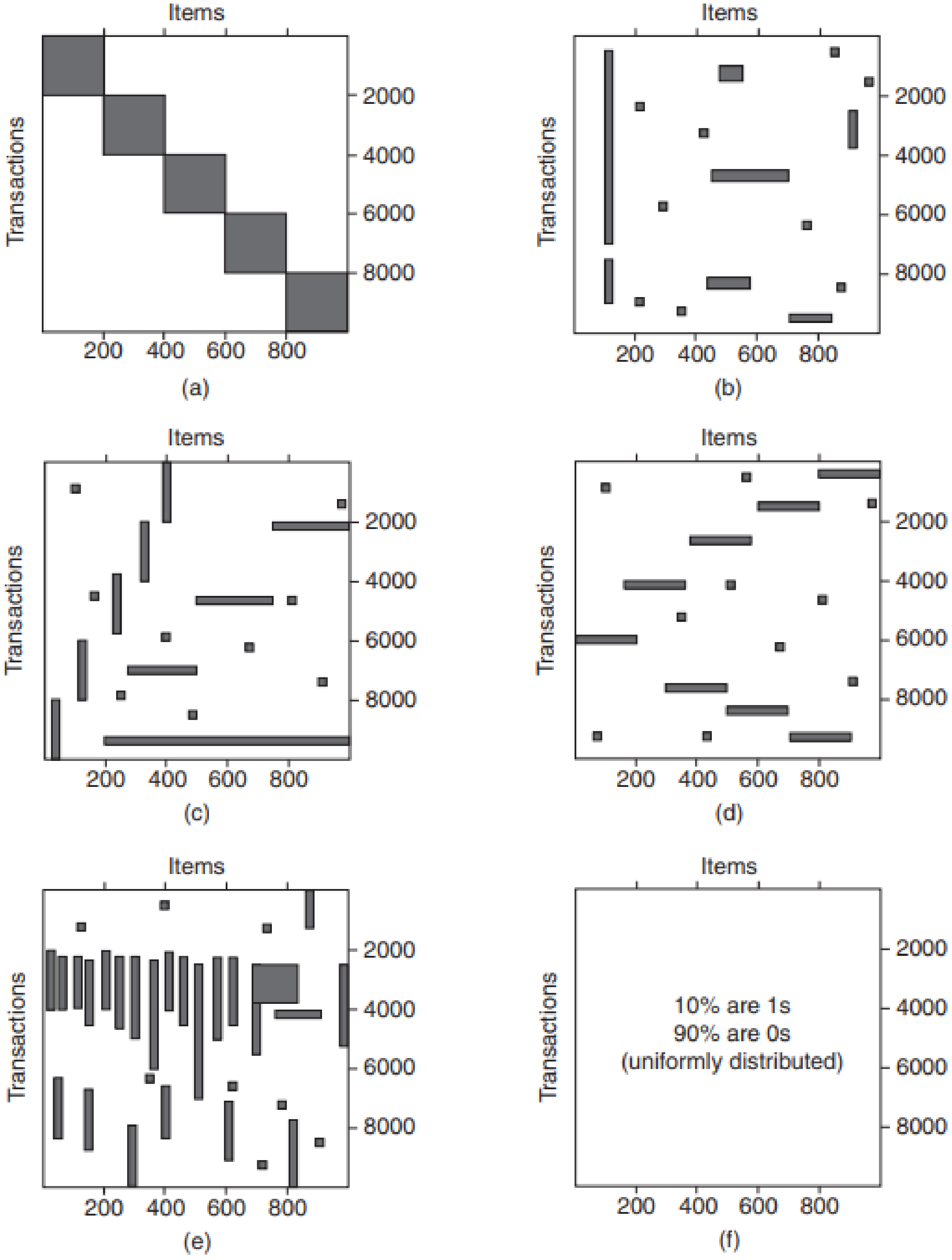
Forth table: isolating VIRGINICA. Sensitivity = 9/(9+1) = 0.90 Specificity = 18/(18+2) = 0.90 Precision = 9/(9+2) = 0.82

But in the last table, we can see that some misclasifications will continue being misclasifications even though we form groups of variables. That is due to the fact that, for example, if VIRGINICA is isolated, a VIRGINICA has beenmisclassified as VERSICOLOR, which is in the other group of factors.

5 Chapter 4 - Exercises

5.1 Exercise 18

Answer the following questions using the data sets shown in Figure 4.34. Note that each data set contains 1000 items and 10,000 transactions. Dark cells indicate the presence of items and white cells indicate the absence of items. We will apply the Apriori algorithm to extract frequent itemsets with minsup = 10 percent (i.e., itemsets must be contained in at least 1000 transactions).



(a) (b) (c) (d) (e) (f)

Figure 4.34. Figures for Exercise 18.

Figure for the exercise 18

(a) Which data set(s) will produce the most number of frequent itemsets?

Data set (e) because it has to generate the longest frequent itemset along with its subsets.

(b) Which data set(s) will produce the fewest number of frequent itemsets?

Data set (d) which does not produce any frequent itemsets at 10

(c) Which data set(s) will produce the longest frequent itemset?

Data set (e).

(d) Which data set(s) will produce frequent itemsets with highest maximum support?
Data set (b)

(e) Which data set(s) will produce frequent itemsets containing items with wide-varying support levels (i.e., items with mixed support, ranging from less than 20 percent to more than 70 percent)?
Data set (e