CSE 5243
Instructor: Jason Van Hulse
Homework 3

Due Date: Thursday, 2/14 at 11:59 pm (upload the solutions to Carmen)

- Please work on this assignment as an individual, not in a group.
- A 15% penalty all homework handed in up to 24 hours late; homework will not be accepted more than 24 hours late
- Show your work

Question #1: Given the below confusion matrix for classifier *C,* compute the accuracy rate, error rate, true positive, false positive, precision and F-measure.

		Predicted Class		
		+	-	Total
Actual	+	350	122	472
Class	-	344	670	1014
	Total	694	792	1486

Question #2: Consider the training examples shown in the following table for a binary classification problem and answer the following questions:

	a_1	a_2	a_3	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?
b) What are the information gains for attributes a_1 and a_2 relative to these training examples?
c) For a_3, which is a continuous attribute, find the binary split which maximizes the information gain.
d) If you were building a decision tree with a single (binary) split, using the *information gain* measure, which attribute would you split on? Draw the decision tree and show the number of positive and negative examples in each leaf node.

e) Would your choice of optimal split in part d) change if you used the *gain ratio* instead of the information gain?

Question #3: Using the dataset given in the table below, build a 2-level decision tree [one split from the root node, followed by 1 additional split from each child node] using the Gini index as the splitting criteria. You should use multiway splits for each attribute. Draw the decision tree and show the number of class C0 and C1 records in each leaf node. Show your work by computing the Gini index of the possible splits. You do not need to worry about pruning the tree that you've built.

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

Question #4: You have built a decision tree model and would like to estimate the generalization performance of the model. Before building the model, you set aside a test dataset with 1200 records. The predicted class label for each of the 1200 records is calculated, and the results are summarized in the table below. The number of actual positive and negative class examples are aggregated for each score range based on the posterior probability P(+ | x). For example, there are a total of 131 test records with P(+ | x) >= 0.9, 118 of which are actually positive and 13 are actually negative.

a) Given a decision threshold of 0.5, write the confusion matrix and calculate the accuracy of the model.
b) Change the decision threshold to 0.7, write the confusion matrix and calculate the accuracy. Compare the TPR and FPR of this choice to those obtained in part a.
c) Using the data in the table, draw the ROC curve. The curve should have 10 points, one for each row in the table.

Table for Exercise #4

Score Range P(+ l x)	# + Class Examples in the score range	# - Class Examples in the score range	Total Examples in each score range
[0.9 - 1]	118	13	131
[0.8 - 0.9)	94	24	118
[0.7 - 0.8)	81	35	116
[0.6 - 0.7)	74	42	116
[0.5 - 0.6)	51	48	99
[0.4 - 0.5)	34	62	96
[0.3 - 0.4)	25	99	124
[0.2 - 0.3)	11	105	116
[0.1 - 0.2)	9	122	131
[0 - 0.1)	3	150	153
Total	500	700	1200

Note that some problems are taken from the textbook *Introduction to Data Mining*.