

# C S 488/508 Introduction to Data Mining

## Homework 3: Classification

### Objective

In this homework, you will do exercises to understand the *basic* concepts of the decision tree classifier and model evaluation.

### Q1. (30 points) Node impurity questions.

Consider the training examples shown in Table 1 for a binary classification problem.

Instance	$a_1$	$a_2$	$a_3$	Target 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	-

Table 1: This table shows some data

- (a) (5 points) What is the entropy of this collection of training examples?
- (b) (5 points) What is the information gain of  $a_1$  relative to these training examples?
- (c) (10 points) What is the best split (between  $a_1$  and  $a_3$ ) according to the information gain?
- (d) (10 points) What is the best split (between  $a_1$  and  $a_2$ ) according to the Gini index?

### Q2. (25 points) Decision tree construction.

$A$	$B$	$C$	Number of Instances	
			+	-
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

Table 2: Data set for Q2

Table 2 gives a data set with three attributes  $A$ ,  $B$ ,  $C$  and two class labels  $+$ ,  $-$ . Build a two-level decision tree.

- (a) (10 points) According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.
- (b) (CS 508 students only) (15 points) Split the two children of the root node.

### Q3. (45 points) Classification model evaluation.

You are asked to evaluate the performance of two classification models,  $M_1$  and  $M_2$ . The test set you have chosen contains 10 binary attributes, labeled as  $A_1, \dots, A_{10}$ . Table 3 shows the probabilities obtained by applying the models to the test set. Assume that we are mostly interested in detecting instances from the positive class.

<i>Instance</i>	<i>TrueClass</i>	$P(+ A_1, \dots, A_{10}, M_1)$	$P(+ A_1, \dots, A_{10}, M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	-	0.44	0.68
4	-	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	-	0.08	0.38
8	-	0.15	0.05
9	+	0.45	0.01
10	-	0.35	0.04

Table 3: Posterior probabilities by applying two models on one test set

- (20 pts) Plot the ROC curve for both  $M_1$  and  $M_2$  (You should plot them on the same graph. Which model do you think is better. Explain your reasons. **Manually show** the steps for calculating TP, TP, TN, FN, TPR, FPR, and draw ROC curve for  $M_1$ . Write code to draw ROC curves for both  $M_1$  and  $M_2$  in one graph. Compare the ROC curves that you get from your manual calculation and the code output.
- (10 points) For model  $M_1$ , suppose you choose the cutoff threshold to be  $t = 0.5$ . In other words, any test instances whose posterior probability is greater than  $t$  will be classified as a positive example. Compute the precision, recall, and F-measure for the model at this threshold value.
- (15 points) Repeat part (b) for Model  $M_1$  using the threshold  $t = 0.1$ . Which threshold do you prefer,  $t = 0.5$  or  $t = 0.1$ ? Are the results consistent with what you expect from the ROC curve?

#### General requirements

- For questions that are not required to be done manually, you can write code or conduct manual calculation to answer the questions.
- Put the code for all these question to one file. Please properly organize the code to make grading easy.

### Submission instructions

A zipped file `hw-lastname.zip` consisting of all the code and the PDF file.

### Grading criteria

- (1) CS 508 students need to answer all the questions.
- (2) CS 488 students do not need to answer questions marked with **(CS 508 only)** although you have the freedom to work on them. Your scores will be scaled to 100. If CS 488 students answer the questions marked with **(CS 508 only)**, you will not have any points deducted if your answers are wrong; you will not get any extra points either if your answers are correct.
- (3) The score allocation has been put beside the questions.
- (4) Please make sure that you test your code **thoroughly**.
- (5) FIVE points will be deducted if files are not submitted in the required format.