CSCI 5523 **Homework 2** **Due: March 11, 2019**

Please clearly mention your name, Student ID and Section(1 for morning, 2 for evening or UNITE) on your submission.

Question 1

You are given a classifier that attempts to predict whether it will rain tomorrow (+) or not (-). The confusion matrix below gives the results of this classification algorithm on a sample of 1000 consecutive days:

actual/ predicted	+	-
+	30	20
-	50	900

(i) Compute the accuracy, precision, recall, and F measure for the confusion matrix.
(Compute precision, recall, and the F-measure with respect to + class only.)

(ii) Which of these metrics is a poor indicator of the overall performance of your algorithm? Which of these metrics is a good indicator of the overall performance? Give a one sentence reason why this is the case?

(iii) Construct a trivial rule-based model which gives better accuracy than the classification algorithm above.

Question 2

You are given a task to evaluate how well a new fire mapping algorithm works. The fire mapping algorithm is a Bayesian classifier which labels all the locations into two classes, burned and unburned. To evaluate the algorithm, two regions are tested. The confusion matrices of these two regions are given in Table 1 and Table 2.

Data set 1		Predicted class	
		Burned	Unburned
Actual Class	Burned	30	20
	Unburned	10	40

Data set 2		Predicted class	
		Burned	Unburned
Actual Class	Burned	30	20
	Unburned	1000	4000

a) Calculate the TNR, FPR, Precision and Recall of M for the "burned" class for both these data sets.

b) Is there a difference in their values for the two data sets? If so, what characteristic of the data sets (that are used to derive the above contingency tables) lead to the differences between the values of (TPR,FPR) and (Precision, Recall) that you observe above.

c) Compute Accuracy and F-measure with respect to 'burned' class for dataset 2.

Question 3

Consider a data set with instances belonging to one of two classes - positive(+) and negative(-). A classifier was built using a training set consisting of equal number of positive and negative instances. Among the training instances, the classifier has a recall of ***50%*** on the positive class and a recall of ***95%*** on the negative class.

The trained classifier is now tested on two data sets. Both have similar data characteristics as the training set. The first data set has 1000 positive and 1000 negative instances. The second data set has 100 positive and 1000 negative instances.

A. Draw the expected confusion matrix summarizing the ***expected*** classifier performance on the two data sets.

B. What is the accuracy of the classifier on the training set? Compute the precision, TPR and FPR for the two test data sets using the confusion matrix from part A. Also report the accuracy of the classifier on both data sets.

C.In the scenario where the class imbalance is pretty high, how are precision and recall better metrics in comparison to overall accuracy? What information does precision capture that recall doesn't?

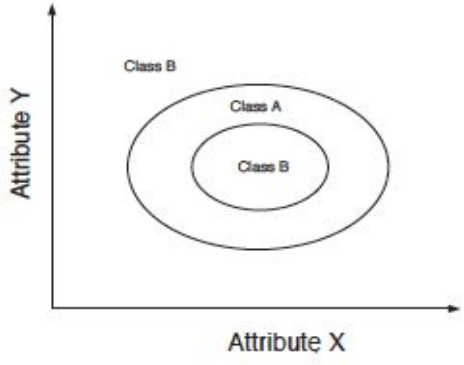
Question 4

Both Minimum Description Length (MDL) and the pessimistic error estimate are techniques used for incorporating model complexity. State one similarity and one difference between them in the context of decision trees.

Question 5

Given the data sets shown in Figure 2, explain how the decision tree, naïve Bayes (NB), and k-nearest neighbor (k-NN) classifiers would perform on these data sets.

Instance	Distinguishing Attributes						Noise Attributes									Class Label
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	
1	■							■					■			■	Class A
2	■						■				■						
3	■									■				■			
4	■								■			■					
5		■													■		
6		■					■						■				
7		■								■						■	
8		■												■			
9			■					■									
10			■								■				■		
11			■				■						■				
12			■						■			■				■	
13				■						■							Class B
14				■			■								■		
15				■							■			■			
16				■					■				■			■	
17					■										■		
18					■						■						
19					■			■									
20					■		■										
21						■				■			■				
22						■		■							■		
23						■					■					■	
24						■			■					■			



(a) Synthetic data set 1. **(b) Synthetic data set 2.**

Figure 1: Data sets for Question 5

Question 6

Answer True or False and briefly explain.

(i) ANN is able to handle redundant attributes.

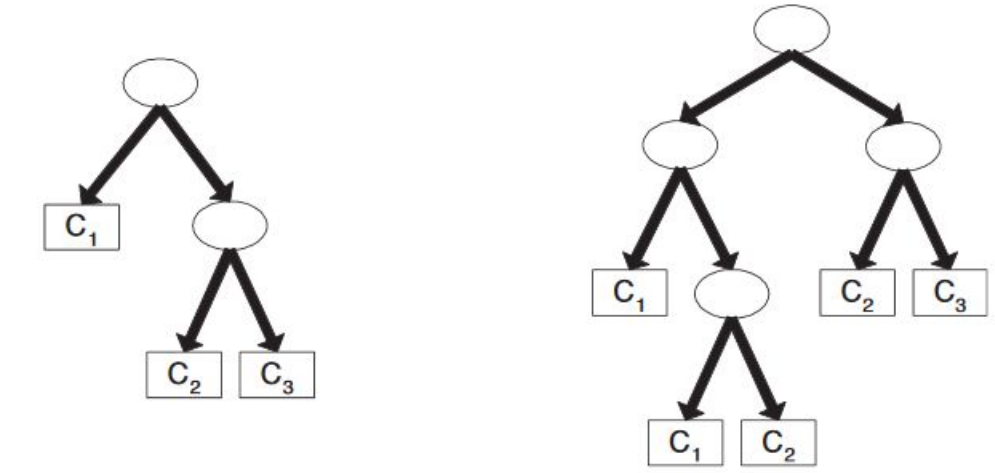
(ii) SVM is particularly effective for categorical data compared to other techniques such as decision trees.

(iii) SVM and Neural Network always produce the same decision boundary for a given data set with two classes.

Practice questions

Question 1

Consider the decision trees shown in Figure Assume they are generated from a data set that contains 16 binary attributes and 3 classes, **C1**, **C2**, and **C3**. Compute the total description length of each decision tree according to the minimum description length principle.



(a) Decision tree with 7 errors

(b) Decision tree with 4 errors

Figure 2 Decision Trees for Practice Question 1

- The total description length of a tree is given by:

$$\textbf{Cost(tree, data)} = \textbf{Cost(tree)} + \textbf{Cost(data|tree)}$$

- Each internal node of the tree is encoded by the ID of the splitting attribute. If there are **m** attributes, the cost of encoding each attribute is log**2(m)** bits.
- Each leaf is encoded using the ID of the class it is associated with. If there are **k** classes, the cost of encoding a class is log**2(k)** bits.

- **Cost(tree)** is the cost of encoding all the nodes in the tree. To simplify the computation, you can assume that the total cost of the tree is obtained by adding up the costs of encoding each internal node and each leaf node.
- **Cost(data|tree)** is encoded using the classification errors the tree commits on the training set. Each error is encoded by log**2(n)** bits, where **n** is the total number of training instances.

Which decision tree is better, according to the MDL principle?

Question 2

Consider the dataset shown in Table 1.1.

Instance	A	B	C	Class
1	0	0	1	-
2	1	0	1	+
3	0	1	0	-
4	1	0	0	-
5	1	0	1	+
6	0	0	1	+
7	1	1	0	-
8	0	0	0	-
9	0	1	0	+
10	1	1	1	+

a. Estimate the conditional probabilities for P(A = 1|+), P(B = 1|+), P(C = 1|+), P(A = 1|−), P(B = 1|−), and P(C = 1|−).

b. Use the conditional probabilities in part (a) to predict the class label for a test sample (A = 1, B = 1, C = 1) using the naïve Bayes approach.

c. Compare P(A = 1, B = 1|Class = +) against P(A = 1|Class = +) and P(B = 1|Class = +). Are the variables conditionally independent given the class?

d. Let us consider the data instance (A=1, B=1, C=1). Compute the probability of this instance belonging to Class = + using
   i. no attributes(i.e. calculate prior probability)

ii. attribute A [P(Class = +|A=1)]
iii. attributes A and B [P(Class = +|A=1, B=1)]
iv. attributes A, B and C [P(Class = +|A=1, B=1, C=1)]

Comment on the change in probability values as we proceed from (i) to (iv).
Now, consider the data shown in Table 1.2.

e. Estimate the conditional probabilities for P(A = 1|+), P(B = 1|+), P(C = 1|+), P(A = 1|−), P(B = 1|−), and P(C = 1|−) using table 1.2.
f. For a new data instance, **x** = (A = 1, B = 1, C = 1), compute the posterior probabilities, P(+|x) and P(-|x) using the naıve Bayes approach.
g. What kind of problems will you encounter in predicting the class of **x** using the posterior probabilities computed in (f) and how can you resolve them?

Question 3

a) Suppose you are given a data set consisting of nominal attributes, such as color, which takes values such as red, blue, green etc. Can you use this data set directly to train an SVM? If not, how will you transform these attributes into a representation that can be used to train an SVM?

b) List one key similarity and one key difference between support vector machines and artificial neural networks.

Question 4

Consider the following classification methods: Decision Trees, RIPPER, Support Vector Machines, Naïve Bayes, k-Nearest Neighbor and Artificial Neural Network. For each of the following scenarios, state the choice of classifier that is well suited and another one that is poorly suited. Give a brief explanation to support your answer. (In some cases, maybe more than one can be well or poorly suited.)
(a) The number of attributes is large and many of them are uninformative (i.e., they provide no discriminative power individually or in combination with other attributes).
(b) There are attributes that are not discriminative individually but their combination provides.
(c) Computation time for model building is to be minimized.

We strongly encourage you to do questions:

2, 3, 4, 6, 7,

from book "Introduction to Data Mining, Second Edition, P.N. Tan, M. Steinbach, A. Karpatne, V. Kumar.", chapter 5, section 10 (5.10)