uOttawa

L'Université canadienne
Canada's university

Ottawa-Carleton Institute for Computer Science
School of Information Technology and Engineering
CSI 5387: Concept Learning and Data Mining
Final Examination Fall 2009

Instructor: Dr. Stan Matwin

Closed Text Exam; Time: 3 hrs. Total points: 100

Write all your answers in the exam booklet. Use one booklet for rough work, and the other for proper answers. Calculators are allowed (but NOT laptops).

Good luck, and have a nice Christmas!

Name:

Student #

Indicate your home university: OTTAWA / CARLETON

Q.	MAX	OBTAINED	Q.	MAX	OBTAINED
1a	3		7a	6	
1b	8		7b	6	
2	6		8a	3	
3	7		8b	3	
4	4		8c	4	
5a	1		9a	2	
5b	6		9b	3	
5c	3		9c	3	
5d	3		9d	4	
6a	3		10	5	
6b	5		11a	4	
			11b	4	
			12	4	
TOTAL	100				

1. [Decision trees] You are given a dataset S concerning opinions of moviegoers. Each instance describes a movie seen by a person, and the class attribute is the Likes/Not. Attributes are:

Age:	Young, Old
Sex:	Male, Female
Violence in the movie :	Yes/No

Age	Sex	Violence	Likes/Not
Y	M	N	N
Y	F	Y	L
Y	M	N	N
O	F	N	L
O	M	Y	L
O	F	Y	L
Y	F	N	N
Y	M	Y	L
O	M	Y	L
O	F	Y	N

a) `(3pts)` give the value of info(S)
b) `(8pts)` which attribute will be chosen as the root? Why? (you can use the calculator here)

2. [1R] `(6pts)` Give (draw) the 1R classifier obtained from this training data

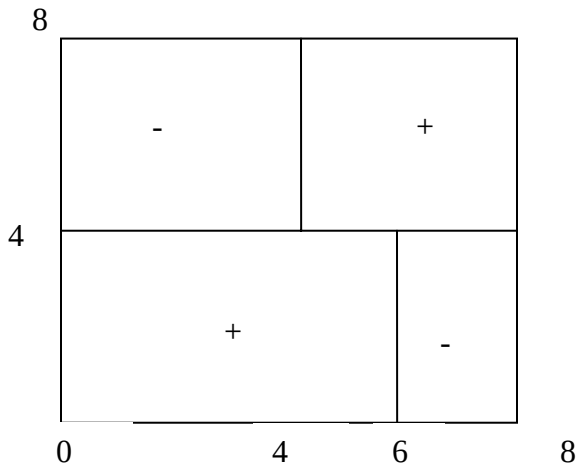
3. [Naïve Bayes] `(7pts)` Suppose the following new instance of the data from Question 1 is to be classified by an Naïve Bayes classifier:

Age	Sex	Violence	Likes/Not
O	M	N	?

The NB formula is $v_{NB} = \text{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$

Show the calculations and the predicted class. If one of the probabilities is computed from the data as =0, use a simple form of Laplace smoothing: add 1 to the numerator and to the denominator of the fraction representing this conditional probability.

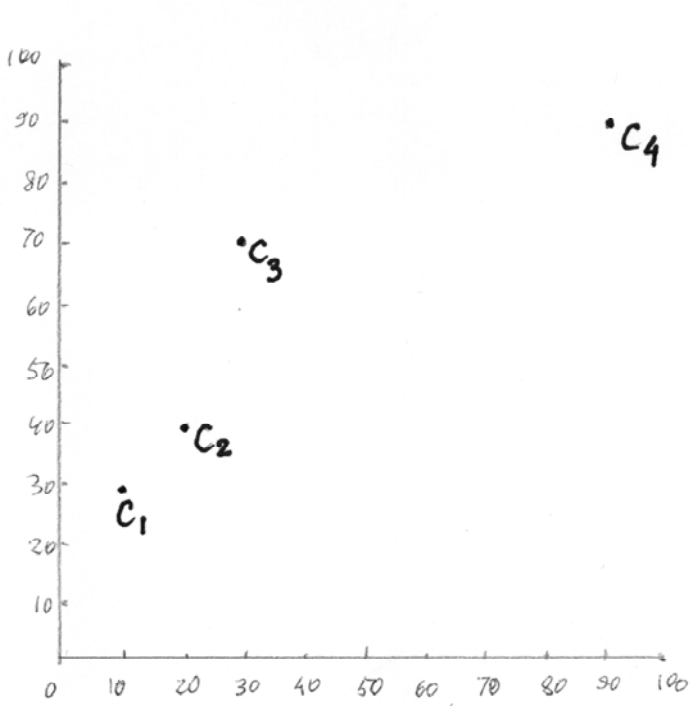
4. [Decision trees]. (4pts) Suppose a data set has two real-valued attributes, x $\in$[0,8], y [0,8], and the class is {+,-}. Given the following partition of the instance space, what tree does it result from?



draw the tree with representing two attributes, *x* corresponding to x-axis, *y* corresponds to y-axis. Use only '>' for tests, assume that if the test is true the right branch is followed.

5. [ROC]. Some data set produced the following classifiers C_1(10,30), C_2 (20,40), C_3(30,70), C_4(90,90) in the ROC space:

TP rate (in %)



FP rate (in %)

a) (1pt) draw the convex ROC curve

b) (6pts) describe (give a formula for) a classifier corresponding to FP rate = 20%? (if you need it, the cosine law is $c^2 = a^2 + b^2 - 2ab\cos\gamma$ where *a*, *b*, *c* are sides of a triangle, and γ is the angle opposite the side *c*.)

c) (3pts) assume pos:neg ratio 1:1. Which classifier will you use? Why?

d) (3pts) For the classifier identified in c) above, what will be its expected accuracy? Justify.

6. [Evaluation]. Explain why

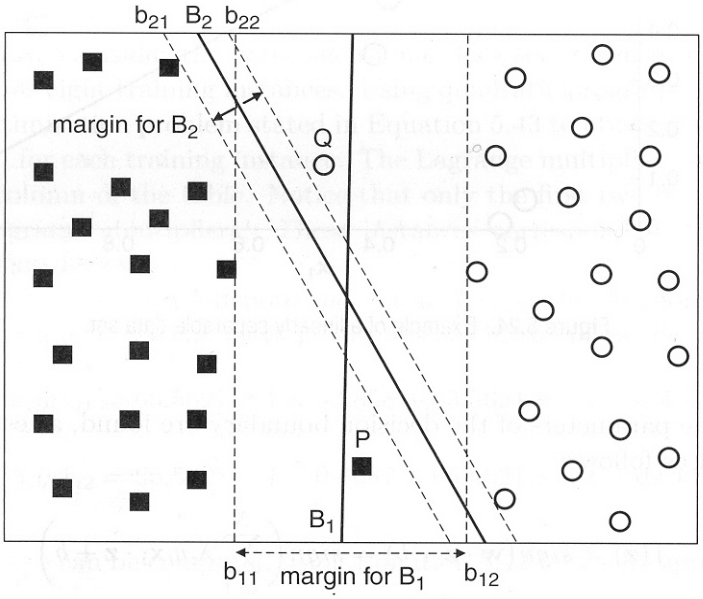
a. (3pts) accuracy is not a good measure for imbalanced data
b. (5pts) why evaluating a classifier in the ROC space IS a good measure even for imbalanced data. Refer to the confusion matrix in your explanation.

7. [SVM]. Φ: $R^2 \rightarrow R^6$ is the mapping from the input space (D = R^2) to the feature space (FS = R^6) defined as follows

$$(x_1, x_2) \rightarrow (z_1, z_2, z_3, z_4, z_5, z_6) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$$

a) (6pts) show that Φ above defines a polynomial kernel of the second degree 2.

b) (6pts) consider the data below, where the two classes are denoted by black and white squares. Consider two candidate decision boundaries, B_1 and B_2. What is the main difference between the two boundaries (1p.)? Which one should be chosen, and why (2)? What mechanism in the SVM will make this choice (3)?



8. [Association rules – Apriori]. The following table database of 10 transactions is given:

Assume the following database of transactions is given (I symbol is omitted in from of the item symbols for simplicity):

T00	1 2 3
T10	4 5
T20	2 3 4 5
T30	1 2 4 5
T40	1 3 5
T50	2 3 4
T60	2 4 5
T70	4 5
T80	3 4
T90	3 5

a) (3pts) give all frequent 2-item sets with support ≥ 0.3
b) (3pts) give all frequent 3-item sets with support ≥ 0.3
c) (4pts) from the item sets identified in c), give an association rule with confidence > 0.7.
In c) show how you obtained the rule and the confidence. In (a) and (b), show which candidate itemsets are rejected due to the monotonicity, and which due to insufficient support.

9. [General] Some Machine Learning methods rely on modifying the distribution during the learning process. This is often achieved by **duplicating** instances.

a) (2pts) name **one** ML method that modifies the distribution of instances in the training data during the training phase.

b) (3pts) Are or are not the decision trees sensitive to adding multiple copies of the same instance in the training set. Justify your answer BRIEFLY (max. 2 sentences)
c) (3pts) Is or are not the naïve Bayesian classifier sensitive to having multiple copies of the same instance in the training set. Justify your answer BRIEFLY (max. 2 sentences)
d) (4pts) Is or is not the SVM classifier sensitive to having multiple copies of the same instance in the training set. Justify your answer BRIEFLY (max. 2 sentences)

10. [Theory- PAC] (5pts) Consider the following learning task T: instances are real numbers, and hypotheses are intervals in R^1. Show that VC dimension (T) = 2. Hint: show that it is at least 2, and that it is < 3.

11. [Clustering] Recall the *k*-means and the Expectation Maximization clustering methods. Explain

a. (4pts) Where does the *k*-means algorithm spend most of its computational effort?
b. (4pts) What does the "expectation' refer to, and what is being 'maximized"?

12. [Combining classifiers] (4pts) In the ECOC method of combining *k* binary classifiers to classify into one of *n* classes, $k > n$, if the edit distance between each of the *n* binary vectors encoding the *k* classifiers is at least *d*, the classifier combined by ECOC can correct $(d-1)/2$ errors. Why?