Sample Midterm Exam. COMP 337

1. Suppose that there are a total of 80 data mining related documents in a library of 1000 documents. Suppose that a search engine retrieves 10 documents after a user enters ‘data mining’ as a query, of which 8 are data mining related documents. What are the precision and recall?

A. precision = 80% and recall = 1%
B. precision = 10% and recall = 8%
C. precision = 8% and recall = 10%
D. precision = 80% and recall = 10%
E. precision = 10% and recall = 1%

2. Let N be the number of number of test data and e be the average error rate. Which of the following statements is correct?

The confidence interval of the true error rate is
A. proportional to e
B. inversely proportional to e
C. proportional to N
D. inversely proportional to N
E. none of the above

3. What is the size of the margin for a support vector machine classifier trained on the dataset {<(-1, -1), F>, <(-1, 1), F>,<(1, -1), T>,<(1, 1), T>}? (In the dataset, <(x1,x2),class> is a data instance with attributes (x1,x2) and class label “class”).
A. 1
B. 2
C. 4
D. 10
E. inifinite
F. 0

4. Consider transforming the following continuous data to a binary-valued attribute using entropy. What is the temperature point with the largest reduction of entropy value?

a) 15.5

b) 16.5

c) 17

d) 19

e) 24

f) 29

Temperature	**Class**
15	F
16	F
18	T
20	F
22	F
25	T
28	T
30	F
31	T

5. Consider the logical OR learning problem in the table. Fill out the following table according to the Perceptron learning rule, assuming the threshold θ is represented as a input X0=-1 with a weight of w0, and a learning rate of 0.1.

x1	x2	t	w0	w1	w2	a=sum(wi*Xi)	output y	error=(t-y)	α=0.1
0	0	0	1	1	1				
0	1	1							
1	0	1							
1	1	1							
0	0	0							
0	1	1							
1	0	1							
1	1	1							

6. Consider the training data in the following table where *Play* is a class attribute. In the table, the *Humidity* attribute has values "L" (for low) or "H" (for high), *Sunny* has values "Y" (for yes) or "N" (for no), *Wind* has values "S" (for strong) or "W" (for weak), and *Play* has values "Yes" or "No".

Humidity	**Sunny**	**Wind**	**Play**
L	N	S	No
H	N	W	Yes
H	Y	S	Yes
H	N	W	Yes
L	Y	S	No

a) (10 marks) Build a conditional probability table for this training data.

b) (5 marks) Is there a zero-frequency problem? Suggest a way to solve it.

c) (10 marks) What is the probability of Play=yes in the following day (Humidity=L, Sunny=N, Wind=W), according to naïve Bayesian rule?