Birla Institute of Technology & Science-Pilani, Hyderabad Campus
1st Semester 2011-2012
Data Mining (CS / IS C415)
Test I (Makeup)

Type: Closed Book Time: 50 mins Max Marks:60 Date: 28/9/2011

PART – A (5 Marks)

1. What principle is used for selecting attributes when building decision trees using the ID3-algorithm (the one you used in the lab)?
a) Select attributes which minimize the tree height
b) Select attributes which maximizes the hypothesis space size
c) Select attributes which have few possible values
d) Select attributes which gives the best information gain

2. What is the advantage of using kernels in support vector machines?
a) They tend to maximize the margins
b) They will reduce the number of support vectors
c) They reduce the risk of getting stuck in local minima
d) They make it possible to do non-linear separations

3. What happens when the number of nodes in the hidden layer of a two-layer perceptron is increased?
a) It will be capable of representing more complicated decision boundaries
b) It will be less prone to overfitting
c) It will generalize better
d) It will converge faster

4. Suppose that X_1, ..., Xm are categorical input attributes and Y is categorical output attribute. Suppose we plan to learn a decision tree without pruning, using the standard algorithm.
a.1 (True or False) : If Xi and Y are independent in the distribution that generated this dataset, then X_i will not appear in the decision tree.

5. For each term (a–c) in the left list, find the explanation from the right list which best describes how the term in used in data mining.

a) Linearly separable data	(1) The amount of information needed to predict an experiment
b) Entropy	(2) Training data point, not part of the concept
c) Negative example	(3) Training with an unknown sample
	(4) Concept which can be learned by a one-layered neural network

PART – B

1. Draw a decision tree which implements the following concept: **(6 Marks)**

$$(a \wedge b \wedge c) \vee (\neg a \wedge \neg c)$$

where a, b, and c are binary (boolean) attributes.

2. The table below describes a training dataset where each sample has four attributes (a_1,a_2, a_3, and a_4) and a corresponding class (c). **(6 Marks)**

   a) What is the entropy for this dataset?

   b) What is the expected information gain from measuring a_1?

a1	a2	a3	a4	class
0	0	1	0	+
0	1	1	1	+
0	0	0	0	-
1	1	0	0	-

3. Consider the data in the following table and model a neural network that can classify the person as overweight or not. **(6 Marks)**

ID	Height	Weight	Class
1	1.6	50	NO
2	1.7	60	NO
3	1.9	70	NO
4	1.5	70	YES
5	1.7	80	YES
6	1.6	90	YES

4.Consider a classifier (hypothesis) *h* that has been evaluated on a test data set of *n* = 100 examples. The observed error is 0.17, i.e. *h* misclassifies 17 test examples. **(6 Marks)**

(a) What is the standard deviation and the (two-sided) 95% confidence interval for the true classification error (the error on the entire population)?

(b) How many test examples would you need to assure that the width of the (two-sided) 95% confidence interval of the true error will be at most 0.1?

5. Assume that you have 2 features (or could even be the raw signal!) from speech signals recorded from 2 'up' and 2 'down' classes Now, you are asked to used KNN to classify whether a test pattern, Y(3,7), i.e. X1=3 and X2=7 is 'up' or 'down'. Compute Euclidean distance of Y (3, 7) to each training set. (**11 Marks)**

	Features		
Pattern No	**X1**	**X2**	**Speech Class**
T1	**7**	**7**	**Up**
T2	**7**	**4**	**Up**
T3	**3**	**4**	**Down**
T4	**1**	**4**	**Down**