Indian Institute of Technology Bombay

Department of Mechanical Engineering

ME 781: Semester End Examination

November 17, 2016: 9:30 AM – 12:30 PM

Maximum Marks: 40

Important Note:

There are three Parts to this question paper. All questions of a Part **SHOULD BE** answered together.

PART - A

1. a) Clearly differentiate between supervised and un-supervised learning, their objectives and applications. List at least three techniques as examples of each. (1)

   b) Given a data set with thousands of observations and tens of features explain the specific steps, in sequence, that you will take to analyse such a data set. (1)

   c) Explain "Map-Reduce" with an example. When is the Map-Reduce technique useful? (1)

   d) What is "dimensionality reduction" and when is it required? Name and explain one technique that helps to achieve this goal. (1)

   e) Clearly bring out the differences between the following: "Maximal Margin Classifier", "Support Vector Classifier" and "Support Vector Machines". (1)

2. In a certain experiment, the response variable Y is dependent on two features X1 and X2. Y is classified as HIGH/LOW. Following is a sample data set, based on which the Support Vectors need to be decided:

ObsNo	X1	X2	Y
1	1.0	3.0	HIGH
2	1.5	4.0	HIGH
3	2.0	2.0	HIGH
4	2.5	3.0	HIGH
5	3.0	3.5	HIGH
6	4.0	4.0	HIGH
7	2.0	1.0	LOW
8	3.0	1.5	LOW
9	3.5	1.0	LOW
10	4.0	3.0	LOW

   a) Using coordinates X1 and X2, plot the observations (points) and show their Y classification (2)
      - Identify and circle the points that are the Support Vectors in your assessment
      - Explain your decision

   b) Sketch the optimal separating hyper-plane and derive its equation in the form: (2)
      - $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$
      - Express the HIGH and LOW regions using the equation of this hyper-plane

   c) Will such a boundary plane always exist? Justify your answer with supporting plots. (1)

3. The following observations are related to two features X1 and X2:

   $X1 = \{2/\sqrt{3}, 4/\sqrt{3}, 6/\sqrt{3}, 12/\sqrt{3}, 16/\sqrt{3}, 20/\sqrt{3}, 40/\sqrt{3}, 48/\sqrt{3}, 56/\sqrt{3}\}$

   $X2 = \{0.5, 1.0, 1.5, 3.0, 4.0, 5.0, 10.0, 12.0, 14.0\}$

   The task is to identify clusters using the hierarchical clustering technique. Answer the following as part of your attempt to carry out this task.

   a) Create a scatter plot (approximately to scale) (1)

   b) Based on the scatter plot detect and explain the nature of relationship between X1 and X2. Derive the parameters of this relationship (2)

   c) Using the above relationship, and the basic concepts of dimensionality reduction, simplify the feature values. List the simplified values of X1, X2 (2)

   d) Using the simplified feature values, walk through the process of the (4)

	hierarchical clustering algorithm, stage-by-stage, to build up the dendrogram using "complete linkage".	
	e) Plot the dendrogram to illustrate the formation and constitution of the clusters	(1)
	PART - B	
4.	a) Explain in brief the concept of type 1 and type 2 errors in the context of a control chart	(1)
	b) Define Average Run Length for an in-control process. How will this definition differ for an out-of-control process?	(1)
	c) If the probability of type 1 error is 0.0027 and that for type 2 error is 0.45, calculate the corresponding run lengths.	(1)
5.	Using your own numerical example, answer the following based on the method discussed in the class for control charts. Make and state the necessary assumptions	
	a) How will you determine the length of a control cycle	(1)
	b) How will you calculate the cost associated with a control chart	(1)
	PART - C	
6.	Wireless cell phone provider wants to see the relation between telephone bill of its customer as a function of the operating system of their cell phone. The cell phone provider identifier three possible operating systems, i.e. Apple OS, Android and others. If a linear regression model is to be fitted then identify how you would have the predictors designed so that they can predict the telephone bill. Write a simple algebraic expression and explain the various predictor variables. The wireless service provider later wants these predictors to be used in predicting the fraction of bill being used in data service. In such a case can you use a linear regression model? If not, then which would be a better model to predict the fraction of bill being used in data services and why? Write a generic expression for the same.	(1) (2)
7.	The Indian gene has a specific expression which makes 1 in every 10,000 Indians susceptible to a specific heart disease. An accurate test (99% accurate) to diagnose this is developed and tested on a patient. The results of the tests turns out to be positive (i.e. the result suggests that the patient has the disease). What is the probability that the patient actually has the disease?	(3)
8.	Let A={(x,y) : $1<=x^2+y^2<=2$}and B ={(x,y): $0<=x^2+y^2<=2$}. Let C = B-A. Then find $(A \oplus C)$ where $(M \oplus N)$={m+n: m $\subset M$ *and* $n \subset N$).	(2)
9.	If x is a random variate which is uniformly distributed between 0 and 1; then how would y be distributed where, $y=a+be^x$.	(2)
10.	State whether the following is true or false: a. K-Nearest Neighbor regression is a good model if the training data is large and clustered in one region. b. In K-Nearest Neighbors regression it is possible to have very precise predictions by choosing a large enough k. c. For a given data, a non-linear model will be more accurate than a linear-model. d. If two predictor variables are strongly correlated, then only one of them is needed in a linear regression model. e. Outliers in a training data are best handled by ignoring that data point.	(5)