Nirma University

Institute of Technology

Semester End Examination (IR), May - 2016

B. Tech. in Computer Engineering / Information Technology, Semester-VI

CE623 Machine Learning

Roll / Exam No. _____ Supervisor's initial with date _____

Time: 3 hours Max. Marks: 100

Instructions:
1. Attempt all questions.
2. Figures to right indicate full marks.
3. Use section-wise separate answer book.
4. Make suitable assumption wherever necessary.

Q.1 Answer the following questions. **[18]**

(a) The following are 8 data points that show the relationship between the number of fishermen and the amount of fish (in thousand pounds) they can catch a day. [12]

Number of Fishermen	Fish Caught
18	39
14	9
9	9
10	7
5	8
22	35
14	36
12	22

According to this data set, find the linear relation function between the number of fishermen and the amount of fish caught.

(b) Differentiate between linear regression and logistic regression using appropriate examples. [06]

Q.2 Answer the following questions. **[16]**

(a) What kind of problem can occur when polynomial regression is used for machine learning? How to overcome such problem(s)? [06]

OR

(a) What is fuzzy clustering? How the membership of the data sample is calculated using fuzzy C-means clustering? [06]

(b) Apply k-means clustering algorithm to create 2 clusters from the data samples given below. Use initial centroids as point 1 and point 4. Show the result for one epoch only. [10]

Point	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Q.3 Answer the following questions. **[16]**

(a) Use Naïve Bayesian Classifier for training which uses previous patients' records of symptoms and diagnosis as given below. [10]

Symptoms				Diagnosis
Chills	Runny nose	Headache	Fever	Flu
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

Can anybody believe that a patient with following symptoms has the flu?

Chills	Runny nose	Headache	Fever
Y	N	Mild	Y

(b) From the below given confusion matrix, find out accuracy, error rate, sensitivity and specificity. [06]

	Predicted class		
Actual class	Classes	Yes	No
	Yes	90	210
	No	140	9560

OR

(b) Give proper example of AdaBoost ensemble method. List out its advantages and disadvantages. [06]

SECTION II

Q.4 Answer the following questions. **[18]**

(a) Consider a travelling salesperson problem. Design the input vector and the fitness function for a genetic algorithm to choose the route for the salesperson. Justify your answer. [06]

(b) The Support Vector Machine is a highly accurate classification method. However, SVM classifiers suffer from slow processing [06]

when training with a large set of data tuples. How this difficulty can be overcome to develop a scalable SVM algorithm for efficient SVM classification in large datasets?

(c) Differentiate between supervised, unsupervised and reinforcement learning. [06]

Q.5 Answer the following questions. **[14]**

(a) Which of the following statement(s) is/are true for item parameter estimation using the expectation maximization (E-M) algorithm? Justify your answer. [04]

a. The E-M algorithm requires that the user determines the starting values for the expectation step. The user can either randomly assign starting values from a constrained distribution or empirically compute them from the data.
b. The expectation step is followed by the maximization step.
c. During the maximization step, the item parameters are estimated followed by structural parameter estimates. These are then re-estimated in an iterative process until the difference between parameter estimates from successive steps is minimal.
d. The efficiency of the E-M estimation is affected by the number of hidden classes.

(b) Related to error there are two possible terms: true error and test error. Define the terms and explain using an example. [04]

(c) Define following terms with appropriate examples. [06]

a. VC Dimension
b. PAC Learning
c. MDL Principle

Q.6 Answer the following questions. **[18]**

(a) Give examples to demonstrate proper use of attribute selection measures information gain, gain ratio and gini index. [06]

OR

(a) Demonstrate various chromosome encoding schemes for evolutionary algorithms with example. [06]

(b) Describe the preprocessing steps to improve accuracy, efficiency and scalability of classification/prediction process. [06]

OR

(b) In classification, what is class imbalance problem? How to deal with it? [06]

(c) What is Q-learning? Give the application where Q-learning can be applied. [06]