Karunya University
(Karunya Institute of Technology and Sciences)
(Declared as Deemed to be University under Sec.3 of the UGC Act, 1965)

End Semester Examinations - 2015-16 Even Semester - May 2016

14CS3074 Advanced Data mining

Set A

Time : 3 hrs
Total Marks: 100

1. a. Use 'DIVIDE-and-CONQUER' approach and determine the root node and one more subsequent node for the following dataset. (15)

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes

b. Explain in detail about linear classification using Winnow (5)

OR

2. a. What is Naïve Bayes Theorem? Use the dataset from Question 1a and determine the class output of test instance : 'Rainy, Mild, High, True' using Naive Bayes classifier. (10)

b. What is "instance based representation"? What is the purpose of KD-tree in searching for an instance? How will you construct KD-tree? Illustrate using an example. (10)

3. a. What is the use of sparse data type? Give an example. (3)

b. Employee table has the following attributes: EmpId, Name, Designation, Date of Joining, Address and Salary. Create ARFF file for the employee data. (5)

c. What is 'association rule' and how will you derive the rules? Use the dataset listed in Question 1a and generate frequent itemsets with coverage >= 2 and derive samples rules from one of the frequent itemsets. (12)

OR

4. a. Explain 'REPLICATED SUBTREE PROBLEM' with example. (8)

b. What is the purpose of data cleansing? In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. (6)

c. Consider the following dataset: 10, 12, 13, 20, 11, 7, 15, 16, 13. (a) Smooth these data using a bin depth of 3 using different binning techniques. (6)

5. a. How will you determine the correlation relationship between two categorical attributes using chi-square test? (10)

b. What is data transformation? What are the methods available for transforming the data? (4)

c. Use the following dataset: 20, 22, 13, 40, 35, 54, 15, 16, 36, 19, 19, 35, 22, 45, 27, 20, 21, 45, 25, 52, 28, 30, 33, 25, 33, 26, 40. (a) Use min-max normalization to transform the value 12 into the range [0.0, 1.0], (b) Use normalization by decimal scaling to transform the value 12, (c) Use Z-score normalization to transform the value 12. (6)

OR

6. a. What is the purpose of wavelet transform? Use 1D Haar Wavelet Transform and transform the following dataset. Explain the steps involved. (10)

15	10	5	6	9	12	13	7

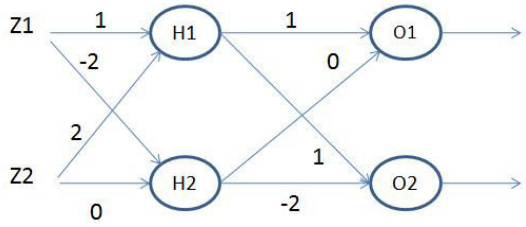
b. What is the purpose of Principal Component Analysis? Explain the steps involved in determining eigen values and eigen vectors. (10)

7. Explain the k-means clustering algorithm. How can we modify the k-means algorithm to reduce its sensitivity to outliers? (20)

OR

8. a.What is a classifier? Compare and contrast 'Multiple Layer Perceptron' and 'Support Vector Machine'. (5)

b. Consider the following Multiple Layer Perceptron network architecture. The network is presented with an input vector $x = [1.0, 3.0]^T$, desired output, $t = [0.95, 0.05]^T$. Update the weight for one iteration using unipolar sigmoidal function and learning rate parameter =1. (15)



9. a. "Bootstrap" method is called as "0.632 bootstrap". Why? How will you estimate the error in this method. (6)

b. Evaluate the output (shown below) of 2-class predictor using the following metrics: TP rate, FP rate, Precison, Recall, F-measure and Overall success rate. (8)

		Predicted Class	
		a	b
Actual Class	a	90	20
	b	15	21

c. Write about "Text Mining". (6)

Wishing you All the Best
