JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST-3 EXAMINATION - 2022

B.Tech. VII Semester (CSE, IT)

COURSE CODE (CREDITS): 19B1WCI731 (2) MAX. MARKS: 35

COURSE NAME: Computational Data Analysis

COURSE INSTRUCTORS: Dr. Ekta Gandotra MAX. TIME: 2 Hours

Note: *All questions are compulsory. Marks are indicated against each question in square brackets.*

Q1. a. What is a dendrogram in hierarchical clustering? How to get the optimal number of clusters using a dendrogram? [3] CO3

b. Consider the following distance matrix for 6 objects. Using single linkage Agglomerative hierarchical clustering, show the first two merge steps (to form clusters). [3] CO3

	A	B	C	D	E	F
A	0.00					
B	0.71	0.00				
C	5.66	4.95	0.00			
D	3.61	2.92	2.24	0.00		
E	4.24	3.54	1.41	1.00	0.00	
F	3.20	2.50	2.50	0.50	1.12	0.00

Q2. a. Explain Adaboost algorithm with the help of an example. [4] CO5

b. Consider the following data pertaining to two books: [2] CO5

Width (X_1)	Thickness (X_2)	Weight (Y)
8	1.8	4.4
8	0.8	2.7

Which of the following two linear hypothesis functions results in more over-fitting and why?

$$Y_{pred} = -3.94 + 0.18\ X_1 + 0.34\ X_2 \quad \text{----(1)}$$

$$Y_{pred} = 2843 - 957\ X_1 + 300\ X_2 \quad \text{----(2)}$$

Q3. *a.* Plot the sigmoid function $1/(1 + e^{-wX})$ vs. $X \in R$ for the weight $w \in \{1, 5, 100\}$. Use these plots to argue why a solution with large weights can cause logistic regression to overfit. (*Note: A qualitative sketch is sufficient.*) [3] CO2

b. List at least four differences between L1 and L2 regularization. [3] CO5

P.T.O.

Q4. a. Elucidate a method used to evaluate the quality of clustering models with the help of an example. [3] CO3

b. Consider the following distance matrix for the data points S1, S2... S8. Label these (as core, border and noise points) to form clusters using DBSCAN algorithm. Take Epsilon = 3.5 and MinPts = 3. [3] CO3

	S1	S2	S3	S4	S5	S6	S7	S8
S1	0.00							
S2	4.24	0.00						
S3	4.47	5.10	0.00					
S4	3.16	4.00	1.41	0.00				
S5	2.00	5.83	4.00	3.16	0.00			
S6	1.00	3.61	5.00	3.61	3.00	0.00		
S7	6.08	3.61	3.61	3.61	6.71	6.00	0.00	
S8	2.00	3.16	2.83	1.41	2.83	2.24	4.12	0.00

Q5. a. Explain the principle of the gradient descent algorithm using a labeled diagram. [3] CO1

b. What are the objectives of feature selection methods? Consider the following dataset of training examples: [3] CO4

A	B	Class Label
T	T	C0
T	T	C0
T	F	C1
F	F	C0
F	T	C1
F	T	C1

Find the information gain of attribute **B** relative to these training examples?

Q6. a. Give the importance of eigen values and eigen vectors in PCA? Suggest a method to calculate the variance captured by each principal component. [3] CO4

b. Suppose 1,000 patients get tested for flu; out of them, 900 are actually healthy and 100 are actually sick. For the sick people, a test was positive for 62 and negative for 38. For the healthy people, the same test was positive for 18 and negative for 882. Construct a confusion matrix for the data and compute the accuracy. [2] CO1