

Name:

Enrolment No:



UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, May 2022

Course: B. Tech CSE+AI/ML
Program: Machine Learning
Course Code: CSAI 2001

Semester: IV
Time : 03 hrs.
Max. Marks: 100

Instructions:

SECTION A
(5Qx4M=20Marks)

S. No.		Marks	CO		
Q 1	Define Machine Learning. Write down five application of it.	4	CO2		
Q 2	List down four application of Linear Regression Model with their dependent and independent variable.	4	CO3		
Q 3	Define mean, median, mode and standard deviation with formula.	4	CO1		
Q 4	Time Point	Infosys	TCS	4	CO1
	Jan 2019	6	20		
	Feb 2019	5	10		
	March 2019	4	14		
	April 2019	3	5		
	May 2019	2	5		
It is given the average stock price of Infosys and TCS for five consecutive months. Find it either the stock price are independent to each other or not.					
Q 5	Differentiate between Similarity Metrics and Term Weighting.	4	CO5		

SECTION B
(4Qx10M= 40 Marks)

Q 6	Explain and discuss the architecture of information retrieval system of Google Search Engine.	10	CO5																				
Q 7	Discuss and derive the mathematical proof of linear regression model.	10	CO3																				
Q 8	<p>Transactional data of AllElectronics</p> <table border="1"><thead><tr><th><i>TID</i></th><th><i>List of item IDs</i></th></tr></thead><tbody><tr><td>T100</td><td>I1, I2, I5</td></tr><tr><td>T200</td><td>I2, I4</td></tr><tr><td>T300</td><td>I2, I3</td></tr><tr><td>T400</td><td>I1, I2, I4</td></tr><tr><td>T500</td><td>I1, I3</td></tr><tr><td>T600</td><td>I2, I3</td></tr><tr><td>T700</td><td>I1, I3</td></tr><tr><td>T800</td><td>I1, I2, I3, I5</td></tr><tr><td>T900</td><td>I1, I2, I3</td></tr></tbody></table>	<i>TID</i>	<i>List of item IDs</i>	T100	I1, I2, I5	T200	I2, I4	T300	I2, I3	T400	I1, I2, I4	T500	I1, I3	T600	I2, I3	T700	I1, I3	T800	I1, I2, I3, I5	T900	I1, I2, I3	10	CO2
<i>TID</i>	<i>List of item IDs</i>																						
T100	I1, I2, I5																						
T200	I2, I4																						
T300	I2, I3																						
T400	I1, I2, I4																						
T500	I1, I3																						
T600	I2, I3																						
T700	I1, I3																						
T800	I1, I2, I3, I5																						
T900	I1, I2, I3																						

	Evaluate the most frequent data item set of 3 items using FP-Growth algorithm for the above AllElectronics data set.																																					
Q 9	<p>Discuss Support Vector Machine model. What is essential tuples in SVM and how it is different from other tuples? Formulate the mathematical proof of the SVM model.</p> <p style="text-align: center;">OR</p> <p>Explain KNN algorithm. Why it is also called Lazy Learner? What are the points to be subjected when choosing the value of k? For the below problem predict for the class of David using KNN and assume the value of k=3.</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>Customer</th> <th>Age</th> <th>Income (K)</th> <th>No. of cards</th> <th>Response</th> </tr> </thead> <tbody> <tr> <td>John</td> <td>35</td> <td>35</td> <td>3</td> <td>Yes</td> </tr> <tr> <td>Rachel</td> <td>22</td> <td>50</td> <td>2</td> <td>No</td> </tr> <tr> <td>Ruth</td> <td>63</td> <td>200</td> <td>1</td> <td>No</td> </tr> <tr> <td>Tom</td> <td>59</td> <td>170</td> <td>1</td> <td>No</td> </tr> <tr> <td>Neil</td> <td>25</td> <td>40</td> <td>4</td> <td>Yes</td> </tr> <tr> <td>David</td> <td>37</td> <td>50</td> <td>2</td> <td>?</td> </tr> </tbody> </table>	Customer	Age	Income (K)	No. of cards	Response	John	35	35	3	Yes	Rachel	22	50	2	No	Ruth	63	200	1	No	Tom	59	170	1	No	Neil	25	40	4	Yes	David	37	50	2	?	10	CO4
Customer	Age	Income (K)	No. of cards	Response																																		
John	35	35	3	Yes																																		
Rachel	22	50	2	No																																		
Ruth	63	200	1	No																																		
Tom	59	170	1	No																																		
Neil	25	40	4	Yes																																		
David	37	50	2	?																																		

SECTION-C
(2Qx20M=40 Marks)

Q 10	<p>Suppose that the data mining task is to cluster the following eight points with (x, y) representing location into three clusters:</p> <p style="text-align: center;">A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9):</p> <p>The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively.</p> <p>a) Write down k-means algorithm b) Apply k-means algorithm for the three cluster centers after the first round execution c) Find the final three clusters</p>	8+6+6 =20	CO4
------	---	----------------------	------------

Q 11	<p>a) Write down Bayes' Theorem. Implement Bayesian Classification model on below dataset and try to predict for the given instance: (2+10+3=15 Marks)</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>Outlook</th> <th>Temp</th> <th>Humidity</th> <th>Windy</th> <th>Play Golf</th> </tr> </thead> <tbody> <tr> <td>Rainy</td> <td>Cool</td> <td>High</td> <td>True</td> <td>?</td> </tr> </tbody> </table>	Outlook	Temp	Humidity	Windy	Play Golf	Rainy	Cool	High	True	?		
Outlook	Temp	Humidity	Windy	Play Golf									
Rainy	Cool	High	True	?									

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

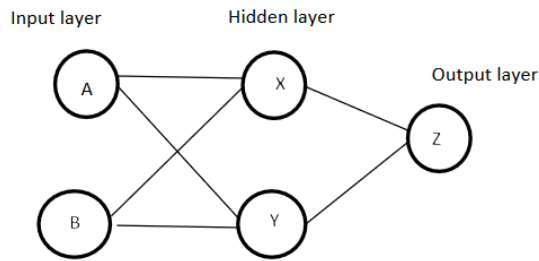
20

CO4

Discuss how to handle the continuous dataset in Bayesian Classification model.

b) Discuss gini index in CART with the formula. How it helped to solve the problem of C4.5? (5 marks)

OR



Input		Output
A	B	Z
0	0	0
0	1	1
1	0	1
1	1	1

Learning rate=0.35

Biases are $\sigma_x = \sigma_y = \sigma_z = 0$. Neural Network of above diagram has two nodes (A,B) in the input layer, two nodes in the hidden layer (X,Y) and one node in the output layer (Z). The values given to weights are taken randomly and will be changed during back propagation iterations. Initial weights of the top input nodes taken at random are 0.4, 0.1. Weights of bottom input node are 0.8 and 0.6. Weights of top hidden node is 0.3 and that of bottom hidden node is 0.9. Assume the number of iterations are two.