Roll No. | | | | | | | | | | |

ANNA UNIVERSITY (UNIVERSITY DEPARTMENTS)

B.E. / B. Tech (Full Time) - END SEMESTER EXAMINATIONS, DECEMBER 2023

MINOR DEGREE ON DATA SCIENCE
(Bio Medical/ Bio Tech/Geo Informatics/ Petroleum)
Fifth Semester

CSM508 MACHINE LEARNING FOR DATA SCIENCE

(Regulation 2019)

Time: 3 hours

Max.Marks: 100

CO 1	To understand the basic concepts of machine learning
CO 2	To understand and build supervised learning models
CO 3	To understand neural network and learn combination of classifiers
CO 4	To understand and build unsupervised learning models.
CO 5	To design and analysis of probabilistic graphical models

BL – Bloom's Taxonomy Levels

(L1 - Remembering, L2 - Understanding, L3 - Applying, L4 - Analysing, L5 - Evaluating, L6 - Creating)

PART- A (10 x 2 = 20 Marks)

(Answer all Questions)

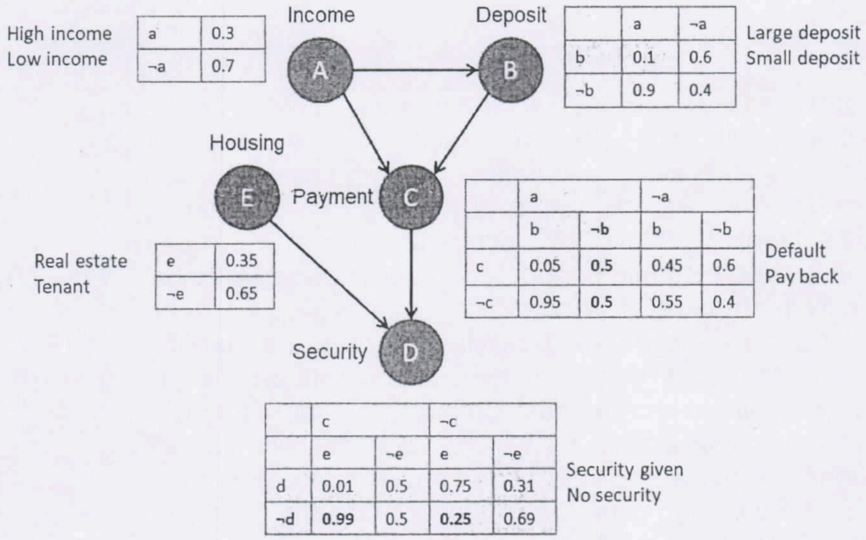
Q. No	Questions	Marks	CO	BL
1	Write any two differences between supervisory and unsupervisory learning methods	**2**	1	L1
2	What is the role of bias and variance in prediction	**2**	2	L2
3	Write an application each for linear and logistic regression	**2**	2	L3
4	What is the importance of kernel function?	**2**	2	L2
5	Show the adjusted weights to learn NOR in perceptron with initial weights are 0 and bias is 1	**2**	3	L3
6	Write any two hyperparameters in neural network based learning	**2**	3	L1
7	Find the clusteroid {(2,10),(4,9), (8,4),(5,8),(6,4),(7,5)}	**2**	4	L4
8	Write any two desirable properties of clustering.	**2**	4	L2
9	Why 'Reinforced' learning is referred as neither supervisory nor unsupervisory learning?	**2**	5	L4
10	What is "hidden" in Hidden Markov Model?	**2**	5	L2

PART- B (5 x 13 = 65 Marks)

Q. No	Questions	Marks	CO	BL
11 (a) (i)	Explain the various data pre-processing techniques with example	**7**	1	L1
(ii)	Consider the task of "guessing question paper" as "Tough/ Easy / Moderate" from the past set of question papers with various factors, like, subject complexity, Type of setter, and others. Identify the features, input, target class and sample training & testing instances	**6**	1	L3

<table>
<tr><td colspan="5">OR</td></tr>
<tr><td>11.b</td><td>
<table>
<tr><td></td><td>Signal violations in last year</td><td>No signal violation in last year</td></tr>
<tr><td>Number of persons used phone while driving</td><td>65</td><td>170</td></tr>
<tr><td>Number of persons NOT used phone while driving</td><td>105</td><td>355</td></tr>
</table>
Find the probability for
i) number of persons have no signal violation last year given that person was not used phone
ii) number of persons used phone while driving and involved in signal violations
iii) signal violations</td><td>5
5
3</td><td>1
1
1</td><td>L3
L3
L3</td></tr>
<tr><td>12 (a) (i)</td><td>The monthly average prices of petrol is given below from April to October. Predict the petrol price in December using linear regression.
<table>
<tr><td>Month</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td></tr>
<tr><td>Average price of petrol in Rs.</td><td>77</td><td>78</td><td>81</td><td>80</td><td>82</td><td>83</td></tr>
</table></td><td>6</td><td>2</td><td>L3</td></tr>
<tr><td>(ii)</td><td>Consider the following training data set.
<table>
<tr><td>Data</td><td>Class</td></tr>
<tr><td><b><i>aa</i></b></td><td>A</td></tr>
<tr><td><b><i>bb</i></b></td><td>A</td></tr>
<tr><td><b><i>ba</i></b></td><td>A</td></tr>
<tr><td><b><i>bb</i></b></td><td>B</td></tr>
</table>
Classify the testing patterns, “<b><i>aaba, bccbba, abb</i></b>” into class A or B using Naïve Bayes classifier</td><td>7</td><td>2</td><td>L4</td></tr>
<tr><td colspan="5">OR</td></tr>
<tr><td>12 (b) (i)</td><td>Classify for a test case <6,> into positive or negative using, KNN classifier by weighted voting, where k=3
<table>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td></tr>
<tr><td>X</td><td>3</td><td>8</td><td>4</td><td>2</td><td>7</td><td>5</td><td>3</td><td>9</td><td>4</td><td>6</td></tr>
<tr><td>Y</td><td>5</td><td>4</td><td>8</td><td>7</td><td>4</td><td>3</td><td>6</td><td>8</td><td>9</td><td>5</td></tr>
<tr><td>Class</td><td>P</td><td>P</td><td>N</td><td>N</td><td>P</td><td>P</td><td>P</td><td>N</td><td>N</td><td>P</td></tr>
</table></td><td>6</td><td>2</td><td>L3</td></tr>
<tr><td>(ii)</td><td>Identify the support vector and draw an optimal hyperplane for set of points : (3,1), (3,-1), (6,1), (6-1) attached to positive class and (1,0), (0,1), (0, -1), (-1,0) of negative class</td><td>7</td><td>2</td><td>L4</td></tr>
<tr><td>13 (a) (i)</td><td>Consider a perceptron to represent the Boolean function AND with the initial weights w1=0.2, w2=-0.2, learning 0.2 and bias 0.4. Show a perceptron that performs this function and updated the weights it gives the desired output.</td><td>7</td><td>3</td><td>L3</td></tr>
<tr><td>(ii)</td><td>Consider a self organizing feature map consists of four training samples and two output units. Train this SOFM network by determining the class memberships of the input data.
Training samples : x1:(1,0,1,0), x2:(0,1,1,0),
x3:(1,0,0,0) & x4:(1,1,1,1)
Initial Weight matrix : Unit 1 [0.2 0.5 0.4 0.6]</td><td>6</td><td>3</td><td>L4</td></tr>
</table>

<table>
<tr><td>13 (b) (i)</td><td>Why “XOR” logic gate cannot be trained to lean using single layer perceptron? How it can be overcome with multilayer perceptron? Explain with illustration</td><td><b>7</b></td><td>3</td><td>L3</td></tr>
<tr><td>ii)</td><td>Explain how a multilayer perceptron can be used to train a Boolean expression with four variables and single output</td><td><b>6</b></td><td>3</td><td>L4</td></tr>
<tr><td>14 (a) (i)</td><td>Perform clustering using Agglomeration by considering i) simple link & ii) Maximum link on the following set of points.
A (1, 1), B(2, 3), C(3, 5), D(4,5), E(6,6), and F(7,5)
Show the dendogram separately for each.</td><td><b>13</b></td><td>4</td><td>L4</td></tr>
<tr><td colspan="5" align="center"><b>OR</b></td></tr>
<tr><td>14 (b) (i)</td><td>Apply k-means algorithm on given data to form 3 clusters using Euclidean distance. Consider B,D,G are initial cluster heads, respectively. Show the new centroids at each iteration.
<table>
<tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td></tr>
<tr><td>X</td><td>3</td><td>8</td><td>4</td><td>2</td><td>7</td><td>5</td><td>3</td><td>9</td><td>4</td><td>6</td></tr>
<tr><td>Y</td><td>5</td><td>4</td><td>8</td><td>7</td><td>4</td><td>3</td><td>6</td><td>8</td><td>9</td><td>5</td></tr>
<tr><td>Class</td><td>P</td><td>P</td><td>N</td><td>N</td><td>P</td><td>P</td><td>P</td><td>N</td><td>N</td><td>P</td></tr>
</table></td><td><b>13</b></td><td>4</td><td>L4</td></tr>
<tr><td>15 (a) (i)</td><td>Consider the following Bayesian Belief Network and answer:



<table>
<tr><td>High income</td><td>a</td><td>0.3</td></tr>
<tr><td>Low income</td><td>¬a</td><td>0.7</td></tr>
</table>

<table>
<tr><td></td><td></td><td>a</td><td>¬a</td></tr>
<tr><td>Large deposit</td><td>b</td><td>0.1</td><td>0.6</td></tr>
<tr><td>Small deposit</td><td>¬b</td><td>0.9</td><td>0.4</td></tr>
</table>

<table>
<tr><td></td><td></td><td>a</td><td></td><td>¬a</td><td></td></tr>
<tr><td></td><td></td><td>b</td><td>¬b</td><td>b</td><td>¬b</td></tr>
<tr><td>Default</td><td>c</td><td>0.05</td><td>0.5</td><td>0.45</td><td>0.6</td></tr>
<tr><td>Pay back</td><td>¬c</td><td>0.95</td><td>0.5</td><td>0.55</td><td>0.4</td></tr>
</table>

<table>
<tr><td>Real estate</td><td>e</td><td>0.35</td></tr>
<tr><td>Tenant</td><td>¬e</td><td>0.65</td></tr>
</table>

<table>
<tr><td></td><td></td><td>c</td><td></td><td>¬c</td><td></td></tr>
<tr><td></td><td></td><td>e</td><td>¬e</td><td>e</td><td>¬e</td></tr>
<tr><td>Security given</td><td>d</td><td>0.01</td><td>0.5</td><td>0.75</td><td>0.31</td></tr>
<tr><td>No security</td><td>¬d</td><td>0.99</td><td>0.5</td><td>0.25</td><td>0.69</td></tr>
</table>

- The customer has a high income (a), a small deposit (¬b), no security (¬d), and owns real estate property (e). What is the probability that a customer with these characteristics will default or pay back the loan?
- What is the probability for non-payment?</td><td><b>8
(5+3)</b></td><td>5</td><td>L5,L6</td></tr>
<tr><td>(ii)</td><td>Find the following from the given transition probabilities & initial probabilities as
Sunny(S) : 0.7, Cloudy(C) : 0.2, Rainy(R) : 0.1

- If today is sunny what is the probability of R,C,C,R,C,S ?
- What is the probability of having continuous 3 days Cloudy followed by Rainy ?</td><td><b>5
(3+2)</b></td><td>5</td><td>L2</td></tr>
</table>

		Tomorrow		
Today		Sunny	Cloudy	Rainy
	Sunny	0.4	0.3	0.3
	Cloudy	0.5	0.2	0.3
	Rainy	0.1	0.2	0.7

OR

15 (b) (i)	State and explain principal component analysis to reduce the two feature sets: {{ 2, 3, 5, 7,9} {1, 4, 0, 6, 2}	**8**	5	L5
(ii)	Describe Hidden Markov Model with a typical case study	**5**	5	L2

PART- C (1 x 15 = 15 Marks)

(Q.No.16 is compulsory)

Q. No	Questions	Marks	CO	BL
16.	Consider an application of “buying a new computer”, based on the list of hardware specifications that best suit the needs for the cheapest option available. Build a machine learning model that can estimate the price of a computer system by taking into account its various features. Identify sample basic computer dataset which can help to develop a price estimation model that can analyze historical data and identify patterns and trends in the relationship between computer specifications and prices. By training a machine learning model on this data, the model can learn to make accurate predictions of prices for new or unseen computer components. Apply any two suitable Machine learning algorithms which can effectively capture complex relationships between features and prices, leading to more accurate price estimates and compare. Justify with i) set of features, ii) set of training & testing dataset. Suggest associated probability, if any based on the chosen ML technique iii)application(workout) of chosen algorithm on the dataset.	**15** (5+5+5)	4,5	5, 6

-----------------*-----------------