PATTERN RECOGNITION
(CSEN 4233)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) Which of the following is not linearly separable?
(a) NAND (b) XNOR (c) NOR (d) All of these

(ii) Factors which affect the performance of learning system do not include?
(a) Representation scheme used (b) Training scenario
(c) Type of feedback (d) Good data structures.

(iii) Decision trees are appropriate for the problems where
(a) Attributes are both numeric and nominal
(b) Target function takes on a discrete number of values
(c) Data may have errors
(d) All of the mentioned.

(iv) In an unsupervised learning
(a) Specific output values are given (b) Specific output values are not given
(c) No specific Inputs are given (d) Both inputs and outputs are given.

(v) The back-propagation algorithm learns a globally optimal neural network with hidden layers.
(a) Always True (b) Always False
(c) Mostly True (d) Mostly False

(vi) DBSCAN cannot be used (with high accuracy) for datasets that are
(a) Convex (b) Uniform density
(c) Non-uniform density (d) None of the above.

(vii) The K-Means algorithm terminates when
(a) a user-defined minimum value for the summation of squared error differences between instances and their corresponding cluster centre is seen.
(b) the cluster centres for the current iteration are identical to the cluster centres for the previous iteration.
(c) the number of instances in each cluster for the current iteration is identical to the number of instances in each cluster of the previous iteration.
(d) the number of clusters formed for the current iteration is identical to the number of clusters formed in the previous iteration.

(viii) The goal in Bayes classifier is to predict class label using,
(a) posterior probability (b) prior probability
(c) likelihood (d) evidence.

(ix) In PCA, what are the principal components?
(a) Features of the dataset (b) Eigenvalues of the covariance matrix
(c) Eigenvectors of the covariance matrix (d) Data points in the dataset.

(x) The penalty term for the Ridge regression is denoted by
(a) the square of the magnitude of the coefficients (b) the square root of the magnitude of the coefficients
(c) the absolute sum of the coefficients (d) the sum of the coefficients.

Fill in the blanks with the correct word

(xi) Automated vehicle is an example of _____ learning.

(xii) Neural Networks are complex _____functions with many parameters.

(xiii) _____clustering techniques starts with all records in one cluster and then try to split that cluster into small pieces.

(xiv) Decision Trees algorithm will always tries to minimize _____ at each level of the tree

(xv) The Independent Component analysis try to predict those _____variables which are revealed as observed variable in the dataset.

Group – B

2. (a) What is min-max normalisation of data? How characteristics of the raw data set changes after applying min-max normalisation on it? *[(CO1)(CO3)(CO5)(Understand/LOCQ)]*

(b) (i) Determine using 1- way ANOVA test on the following dataset, whether the attribute/feature Salary of an Employee is dependent on his/ her Stream in Graduation degree or not, at the 0.05 significance level (α). The F-distribution table is provided for your reference.

(ii) State the null hypothesis and alternative hypothesis in the ANOVA test of the current problem for doing feature selection?

DATASET

Employee No	Stream of Graduation degree	Salary
1	Science	80000
2	Engineering	85000
3	Science	75000
4	Engineering	60000
5	Management	75000
6	Science	95000
7	Engineering	50000
8	Management	57000
9	Management	95000
10	Science	40000

F-distribution where α = 0.05

V_1 \ V_2	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27

(CO1)(CO5)(Apply/IOCQ)]

(2 + 2) + (6 + 2) = 12

3. (a) Is K-NN a linear or non-linear classifier- Justify your answer. Explain why high dimensionality of data will affect the performance of K- NN classifier. *[(CO1,CO3,CO5)(Analyse/LOCQ)]*

(b) (i) Determine using chi-square test on the following dataset, whether the attribute/feature indicating the presence of Medical insurance is dependent on income level or not, at the 0.05 significance level. . The chi-square distribution table is provided for your reference.

(ii) State the null hypothesis and alternative hypothesis in the chi-square test of the current problem for doing feature selection?

Person ID	Income Level	Medical Insurance
1	High	Yes
2	Medium	Yes
3	Low	No
4	High	Yes
5	Medium	No
6	Low	Yes
7	High	Yes
8	Medium	Yes
9	Low	No

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89

(CO1)(CO5)(Evaluate/IOCQ)]

(2 + 2) + (6 + 2) = 12

Group - C

4. (a) Why naïve Bayesian classification is called naïve? Briefly outline the major ideas of naïve Bayesian classification.

[(CO3)(Analyse/IOCQ)]

(b) Use Naïve Bayes' classifier to predict whether a baloon defined by the tuple (Colour = Purple, Size = Small, Act = Dip and Age = Child) is inflated or not. The training data is as follows:

Table – I

Sl No	Color	Size	Act	Age	Inflated
1	Yellow	Small	Stretch	Child	T
2	Yellow	Small	Stretch	Child	T
3	Yellow	Small	Stretch	Child	T
4	Yellow	Small	Stretch	Child	T
5	Yellow	Small	Stretch	Adult	T
6	Yellow	Small	Stretch	Child	F
7	Purple	Large	Dip	Adult	F
8	Purple	Large	Dip	Child	F
9	Purple	Small	Stretch	Adult	T
10	Purple	Small	Stretch	Child	F
11	Purple	Small	Dip	Adult	T
12	Purple	Small	Dip	Child	T
13	Purple	Large	Stretch	Adult	F
14	Purple	Large	Stretch	Child	F
15	Purple	Large	Dip	Adult	F
16	Purple	Large	Dip	Child	T

[(CO3)(Analyse/LOCQ)]

4 + 8 = 12

5. (a) Define Information Gain. *[(CO1)(Remember/LOCQ)]*

(b) Consider the data provided in Table – I of Question 4 (b). Calculate the information gain when splitting on different attributes. Which attribute would the decision tree induction algorithm choose? *[(CO4)(Analyse/LOCQ)]*

3 + 9 = 12

Group - D

6. (a) Define minimum distance and maximum distances between two clusters. *[(CO5)(Understand/LOCQ)]*

(b) Construct the dendrogram for the following proximity matrix using both minimum distance and maximum distances.

	P1	P2	P3	P4	P5
P1	1.00	0.90	0.10	0.65	0.20
P2	0.90	1.00	0.70	0.60	0.50
P3	0.10	0.70	1.00	0.40	0.30
P4	0.65	0.60	0.40	1.00	0.80
P5	0.20	0.50	0.30	0.80	1.00

[(CO4)(Apply/LOCQ)]

4 + 8 = 12

7. (a) Explain the process of selecting the parameters Eps (radius that defines the neighbourhood of a point) and MinPts (minimum number of points in the neighbourhood of the core point) in DBSCAN. *[(CO4)(Apply/IOCQ)]*

(d) Describe the Expectation Maximization Clustering Algorithm along with necessary explanation. *[(CO2,CO4,CO6)(Describe/IOCQ)]*

4 + 8 = 12

Group - E

8. (a) Discuss the difference between Principal Component Analysis and Independent Component Analysis in the context of Feature/ Dimension Reduction *[(CO1,CO4,CO5)(Analyse/LOCQ)]*

(b) Explain how PCA maximizes variance in the data in a new dimension space. *[(CO1,CO4,CO5,CO6)(Analyse/IOCQ)]*

6 + 6 = 12

9. Write notes on the following topics:

(i) Linear Discriminant Analysis can be used for Dimension Reduction

(ii) Lasso Regression based Feature Selection. *[(CO1,CO4,CO5,CO6)(Understand/HOCQ)]*

(6 + 6) = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	48	40	12

Course Outcome (CO):

After the completion of the course students will be able to

1. Learn and understand feature, pattern and the problem of pattern recognition.
2. Understand and describe the difference between supervised and unsupervised learning.
3. Understand and apply pattern recognition algorithm that utilizes supervised learning.
4. Understand and apply pattern recognition algorithm that utilizes unsupervised learning.
5. Analyze pattern recognition algorithms and techniques.
6. Design simple pattern recognition systems.

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.*