B.TECH/CSE/5TH SEM/CSEN 3132/2020

DATA MINING & KNOWLEDGE DISCOVERY (CSEN 3132)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

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

Group – A (Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) Bayesian classifier is
- (a) A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory
- (b) Any mechanism employed by a learning system to constrain the search space of a hypothesis
- (c) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation
- (d) None of these.

(ii) Parameters for association rule mining are
(a) Confidence and Item set (b) Confidence and Item Count
(c) Support and Item count (d) Support, confidence and Item count.

(iii) You are given data about seismic activity in Japan, and you want to predict a magnitude of the next earthquake, this is an example of
(a) Dimensionality Reduction (b) Supervised Learning
(c) Unsupervised Learning (d) Reinforcement Learning.

(iv) Data Mining is
- (a) The actual discovery phase of a knowledge discovery process
- (b) The stage of selecting the right data for a Knowledge Discovery process
- (c) A subject-oriented integrated time variant non-volatile collection of data in support of management
- (d) None of the Above.

(v) 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.

(vi) Which of the following is finally produced by Hierarchical Clustering?
(a) Final estimate of cluster centroids
(b) Tree showing how close things are to each other
(c) Assignment of each point to clusters
(d) All of the Mentioned.

(vii) Slack variable is applicable for
(a) SVM (b) Bayes classifier
(c) K-means (d) None of the above.

(viii) In a picture, where 7 cats and 10 dogs are present, your dog detection algorithm has detected 9 entities, out of which only 6 are dogs and remaining are cats. What is the precision of your algorithm?
(a) 0.6 (b) 0.66 (c) 6/17 (d) 0.9.

(ix) K-means clustering does not suffer from
(a) selection of distance metric (b) selection of k
(c) initial guess as to cluster centroids (d) None of the above.

(x) DBScan uses k-nearest neighbour distance to find the parameter
(a) eps (b) minpts
(c) Core points (d) Noise points.

Group – B

2. (a) Define Gini Index and gain in Gini index.

(b) Construct (induct) a decision tree using gain in Gini index from the data provided in the following table. Consider the Gender as the class label.

Sl No	Over 170CM	Eye	Hair length	Gender
1	No	Blue	Short	Male
2	Yes	Brown	Long	Female
3	No	Blue	Long	Female
4	No	Blue	Long	Female
5	Yes	Brown	Short	Male
6	No	Blue	Long	Female
7	Yes	Brown	Short	Female
8	Yes	Blue	Long	Male

2 + 10 = 12

3. Write short notes on ***any three*** of the following topics: **(3 × 4) = 12**
(i) Data space and Feature space (ii) Classification and Prediction
(iii) Under fitting and over fitting (iv) Rule Based Classification.

Group – C

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

(b) Consider the following dataset of species classification table:

Body Tem-pera-ture	Species Name	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates	Class Label
Warm	Human	yes	no	no	yes	no	Mammal
Cold	Python	no	no	no	no	yes	Reptile
Warm	Whale	yes	yes	no	no	no	Mammal
Cold	Frog	no	semi	no	yes	yes	Amphibian
Cold	Komodo	no	no	no	yes	no	Reptile
Warm	Bat	yes	no	yes	yes	yes	Mammal
Warm	Pigeon	no	no	yes	yes	no	Bird
Warm	Cat	yes	no	no	yes	no	Mammal
Cold	Leopard	yes	yes	no	no	no	Fish
Cold	Turtle	no	semi	no	yes	no	Reptile
Warm	Penguin	no	semi	no	yes	no	Bird
Warm	Porcupine	yes	no	no	yes	yes	Mammal
Cold	Eel	no	yes	no	no	no	Fish
Cold	Salamander	no	semi	no	yes	yes	Amphibian

Using Naive Bayes Classifier on the above data set of species classification, find the class label of the species called Salmon having following attribute values:

Body Tempe-rature	Species Name	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates
Cold	Salmon	no	yes	no	no	no

(1 + 3) + 8 = 12

5. (a) Suppose a support vector machine for separating pluses from minuses finds a plus support vector at the point x_1= (1, 0), a minus support vector at x_2= (0, 1). You are to determine values for the classification vector w and the threshold value b.

(b) Construct the Lagrangian for the primal optimization problem in finding the support vectors for a two-class linearly separable classification problem.

3 + 9 = 12

Group – D

6. (a) Define support and confidence in mining frequent pattern mining. Are these measures symmetric? Justify your answer.

(b) In a fast food restaurant there are 5 different food items (viz., M1, M2, M3, M4 and M5) available in their menu. Food items ordered in 9 different online transactions/orders are given in the table below:

Order/Transaction Id	Ordered Food Items
Order: 1	M1, M2, M5
Order: 2	M2, M3
Order: 3	M2, M4,M5
Order: 4	M1, M3
Order: 5	M2, M3
Order: 6	M1, M2, M3, M5
Order: 7	M1, M2, M4,M5
Order: 8	M1, M2, M3
Order: 9	M1, M3

(i) Compute the support for item-sets {M5}, {M2, M4} and {M2, M4, M5} by treating each transaction ID as a market basket.

(ii) Use the above results to compute the confidence for the association rules {M2, M4} → {M5} and {M5} → { M2, M4}.

(4 + 2) + (3 + 3) = 12

7. You are given the transaction data shown in the Table below

Txn Id	List of Items
1	a, b, d
2	a, b, c
3	b, f
4	a, d
5	b, c
6	a, b, d, e
7	a, b, d, f
8	a, c, e
9	a, b, f
10	a, c, e, f

(i) Construct the FP-growth tree.

(ii) Find the frequent item sets assuming minpts = 3.

(iii) Find at least 6 (if possible) association rules.

(6 + 3 + 3) = 12

Group – E

8. (a) Define minimum distance and maximum distances between two clusters.

(b) Construct the dendrograms for the following proximity matrix using both minimum distance and maximum distance. Show all the steps.

	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

2 + 10 = 12

9. (a) Define Core point, Border Point and Noise point in the perspective of DBSCAN clustering algorithm.

(b) Describe the DBSCAN Algorithm.

(c) Explain why DBSCAN does not work well for the data having varying density.

(d) Describe, in brief, a methodology to select the values of the parameters, viz., eps (the radius) and the minpts (the minimum points)) of the DBSCAN Algorithm.

3 + 3 + 2 + 4 = 12

Department & Section	Submission link:
CSE	https://classroom.google.com/c/MTQxNTUwMDI4NzA5/a/Mjc0MDU3MDU5MjAx/details