DATA MINING AND KNOWLEDGE DISCOVERY
(CSEN 4144)

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) Suppose your model is overfitting. Which of the following is NOT a valid way to try and reduce the overfitting?
(a) Increase the amount of training data.
(b) Improve the optimisation algorithm being used for error minimisation.
(c) Decrease the model complexity.
(d) Reduce the noise in the training data.

(ii) _____ predicts future trends & behaviors, allowing business managers to make proactive, knowledge-driven decisions.
(a) Data warehouse (b) Data mining
(c) Datamarts (d) Metadata.

(iii) _____ is the goal of data mining.
(a) To explain some observed event or condition
(b) To confirm that data exists
(c) To analyze data for expected relationships
(d) To create a new data warehouse.

(iv) Classification rules are extracted from _____.
(a) root node (b) decision tree (c) siblings (d) branches.

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

(vi) You are given data about seismic activity in the United States, and you want to predict the magnitude of the upcoming earthquake. This can be considered as an example of which of the following methods?

(a) Supervised learning (b) Unsupervised learning
(c) Feature Selection (d) Dimensionality reduction.

(vii) How will you counter over-fitting in decision tree?
(a) By pruning the longer rules
(b) By creating new rules
(c) Both by pruning the longer rules' and 'By creating new rules'
(d) None of the above options.

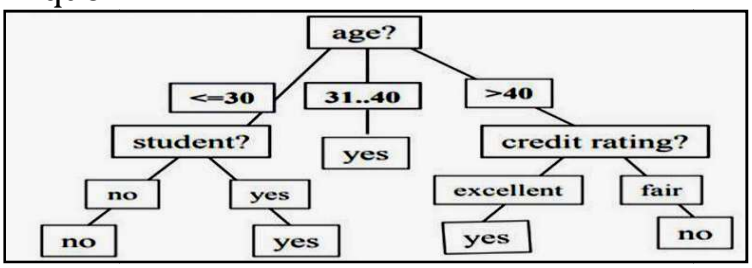
(viii) After SVM learning, each Lagrange multiplier α_i takes either zero or non-zero value. What does it indicate in each situation?
(a) A non-zero α_i indicates the data point i is a support vector, meaning it touches the margin boundary.
(b) A non-zero α_i indicates that the learning has not yet converged to a global minimum.
(c) A zero α_i indicates that the data point i has become a support vector data point, on the margin.
(d) A zero α_i indicates that the learning process has identified support for vector i.

(ix) Which is needed by K-means clustering?
(a) Defined distance metric (b) Number of clusters
(c) Initial guess as to cluster centroids (d) All of these.

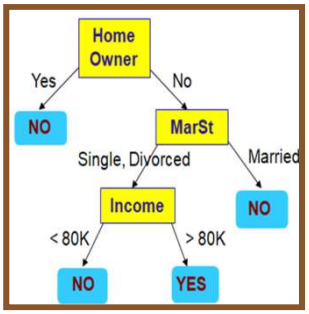
(x) What do you mean by generalization error in terms of the SVM?
(a) How far the hyperplane is from the support vectors
(b) How accurately the SVM can predict outcomes for unseen data
(c) The threshold amount of error in an SVM
(d) All of the above.

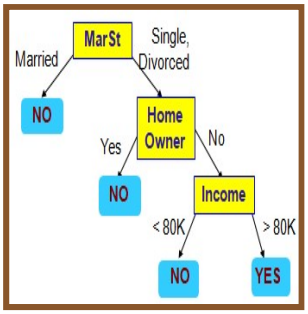
Group – B

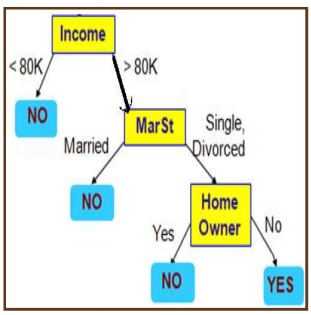
2. (a) Extract a rule-based system from the decision tree given below. Use rule-based ordering technique.



(b)







Derive a rule based system based on the above decision trees discarding redundant rules.

6 + 6 = 12

3. (a) Define Information gain and gain in the Gini index.

(b) Consider the training examples shown in Table 1 for a binary classification problem.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Calculate the information gain when splitting on Gender, Car Type and Shirt Size. Which attribute would the decision tree induction algorithm choose?

3+ 9 = 12

Group – C

4. (a) Explain how to compute and maximize the margin in SVM?

(b) Justify the assertion: 'A support vector can reside inside a margin'.

(c) What is the use of kernel function in Support Vector Machine?

5 + 3 + 4 = 12

5. (a) Briefly explain the Bayes theorem in the context of classification

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

Body Temperature	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 given data set of species classification, find the class label of the species called Salmon having following attribute values:

Body Temperature	Species Name	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates
Cold	Salmon	no	yes	no	no	no

4 + 8 = 12

Group – D

6.

Table . Market basket transactions.

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

Consider the market basket transactions shown in Table.

(i) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

(ii) What is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?

(iii) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

(iv) Find an itemset (of size 2 or larger) that has the largest support.

(v) Find a pair of items, a and b, such that the rules {a} → {b} and {b} → {a} have the same confidence.

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

7. (a) Define support and confidence in mining frequent pattern.

(b) With an example, briefly explain the apriori algorithm.

Table .1. Example of market basket transactions.

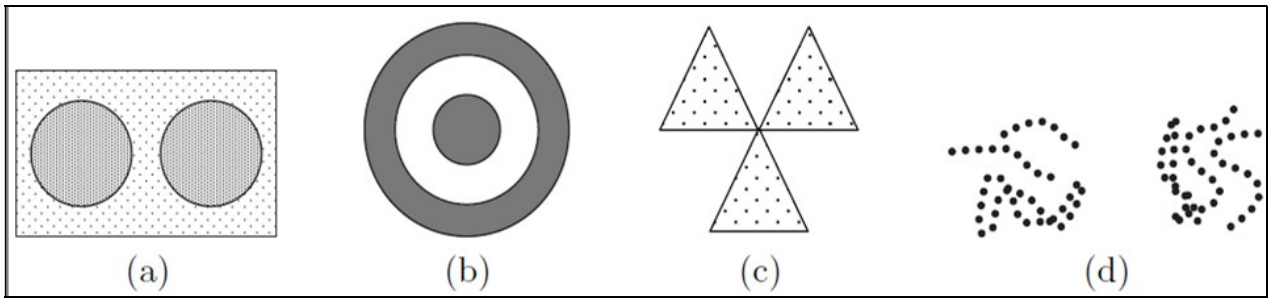
Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

(c) Construct the FP-tree for the data set shown in Table 1 and find all frequent item-sets using FP-growth approach considering 2 as the minimum support count.

2 + 3 + 7 = 12

Group – E

8.



(a) Identify the clusters in the above figure using the center-, contiguity-, and density based definitions. Also indicate the number of clusters for each case and give a brief indication of your reasoning.

(b) Suppose that for a data set
- there are m points and K clusters,
- half the points and clusters are in "more dense" regions,
- half the points and clusters are in "less dense" regions, and
- the two regions are well-separated from each other.

For the given data set, which of the following should occur in order to minimize the squared error when finding K clusters:

(i) Centroids should be equally distributed between more dense and less dense regions.

(ii) More centroids should be allocated to the less dense region.

(iii) More centroids should be allocated to the denser region.

Note: Do not get distracted by special cases or bring in factors other than density. However, if you feel the true answer is different from any given above, justify your response.

6 + 6 = 12

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

(b) Describe the DBSCAN Algorithm.

(c) Describe 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.

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

3 + 3 + 4 + 2 = 12

Department & Section	Submission link:
CSE A+B+C	https://classroom.google.com/c/MTIyMjE3OTk4ODA2/a/MjY1MDE0NTgyMTIy/details