


CMR INSTITUTE OF TECHNOLOGY		USN							
Internal Assesment Test - II									
Sub:	Data Mining with Business Intelligence						Code:	20MCA252	
Date:	22.09.2021	Duration:	90 mins	Max Marks:	50	Sem:	II	Branch:	MCA
Answer Any One FULL Question from each part.									
								Ma rks	OBE
									CO

Part - I

1 (a) What is Rule based classifier? Explain Sequential Covering Algorithm.

[10]	CO3	L2
[10]	CO2	L2

OR

2 (a) What is anti monotone property of support? Explain Apriori Algorithm with example.

Part – II

3 (a) Explain Hunt Algorithm for decision tree induction.

[5]	CO3	L2
[5]	CO3	L2
[10]	CO3	L2

(b) What are node impurity measures? Explain with example.

OR

4 (a) Explain Back propagation with Neural Network.

PART - III

5 (a) Explain Data mining for business Applications like Balanced Scorecard, ClickstreamMining.

[10]	CO4	L2
[10]	CO3	L2

OR

6 (a) What is Classification and Prediction? Explain the issues regarding classification and Prediction.

Part – IV

7 (a) Explain in detail Linear and nonlinear regression, Logistic Regression.

[10]	CO3	L2
[10]	CO2	L3

OR

8 (a) What is concept description? Discuss Data Generalization and summarization-based characterization.

Part – V

9 (a) Explain Improved Apriori algorithm – Incremental ARM – Associative Classification – Rule Mining.

[10]	CO2	L2
[10]	CO4	L2

OR

10(a) What is Big data Business Analytics? Also discuss about Data Analytics Life Cycle.

1. Ans: Rule-Based Classifier

- Classify records by using a collection of “if...then...” rules
 - Rule: (*Condition*) → *y*
 - where
 - *Condition* is a conjunctions of attributes
 - *y* is the class label
 - *LHS*: rule antecedent or precondition
 - *RHS*: rule consequent
 - Examples of classification rules:
 - (Blood Type=Warm) ∧ (Lay Eggs=Yes) → Birds
 - (Taxable Income < 50K) ∧ (Refund=Yes) → Evade=No

Rule-based Classifier (Example)

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

R1: (Give Birth = no) \wedge (Can Fly = yes) \rightarrow Birds

R2: (Give Birth = no) \wedge (Live in Water = yes) \rightarrow Fishes

R3: (Give Birth = yes) \wedge (Blood Type = warm) \rightarrow Mammals

R4: (Give Birth = no) \wedge (Can Fly = no) \rightarrow Reptiles

R5: (Live in Water = sometimes) \rightarrow Amphibians

Application of Rule-Based Classifier

- A rule r **covers** an instance x if the attributes of the instance satisfy the condition of the rule

R1: (Give Birth = no) \wedge (Can Fly = yes) \rightarrow Birds

R2: (Give Birth = no) \wedge (Live in Water = yes) \rightarrow Fishes

R3: (Give Birth = yes) \wedge (Blood Type = warm) \rightarrow Mammals

R4: (Give Birth = no) \wedge (Can Fly = no) \rightarrow Reptiles

R5: (Live in Water = sometimes) \rightarrow Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	warm	no	yes	no	?
grizzly bear	warm	yes	no	no	?

The rule R1 covers a hawk => Bird

The rule R3 covers the grizzly bear => Mammal

Direct Method: Sequential Covering

Extracts the rules one class at a time for data sets having more than two classes.

Criterion for choosing class depend on the factor such as class prevalence.

- Start from an empty decision list, R .
- Grow a rule using the Learn-One-Rule function
- Add the new rule to the bottom of the decision list
- Remove training records covered by the rule
- Repeat Step (2) and (3) until stopping criterion is met

Sequential Covering Algorithm:

- Let E be the training records and A be the set of attribute-value pairs, $\{(A_j, V_j)\}$.
- Let Y_o be an ordered set of classes $\{y_1, y_2, \dots, y_k\}$
- Let $R = \{ \}$ be the initial rule list.
- for each class $y \in Y_o - \{y_k\}$ do
- while stopping condition is not met do
- $r \leftarrow \text{Learn-One-Rule}(E, A, y)$.

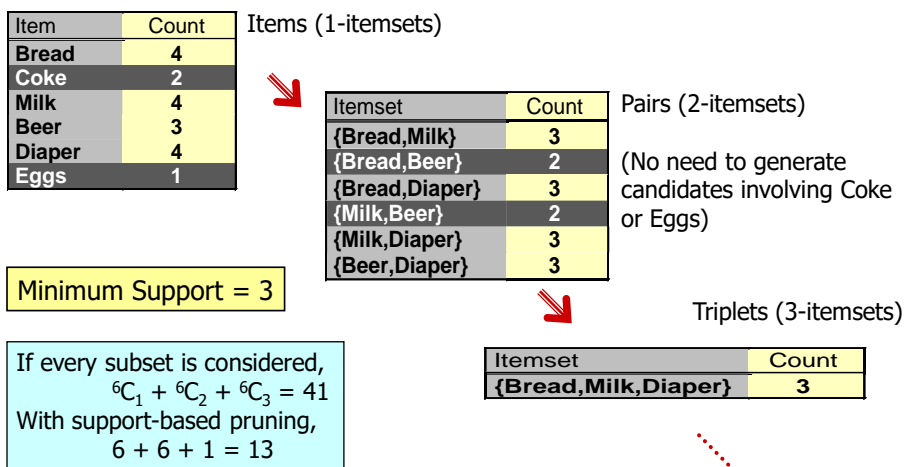
- 7: Remove training records from E that are covered by r.
- 8: Add r to the bottom of the rule list: R->RVr.
- 9: end while
- 10: end for
- 11: Insert the default rule, {}->yk, to the bottom of the rule list R

2. Ans Apriori principle:

- If an itemset is frequent, then all of its subsets must also be frequent
 - Apriori principle holds due to the following property of the support measure:

$$\forall X, Y : (X \subset Y) \Rightarrow s(X) \geq s(Y)$$
- Support of an itemset never exceeds the support of its subsets
- This is known as the anti-monotone property of support

Illustrating Apriori Principle



Apriori Algorithm

Method:

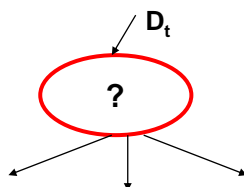
- Let k=1
- Generate frequent itemsets of length 1
- Repeat until no new frequent itemsets are identified
 - ◆ Generate length (k+1) candidate itemsets from length k frequent itemsets
 - ◆ Prune candidate itemsets containing subsets of length k that are infrequent
 - ◆ Count the support of each candidate by scanning the DB
 - ◆ Eliminate candidates that are infrequent, leaving only those that are frequent

3a. Ans

General Structure of Hunt's Algorithm

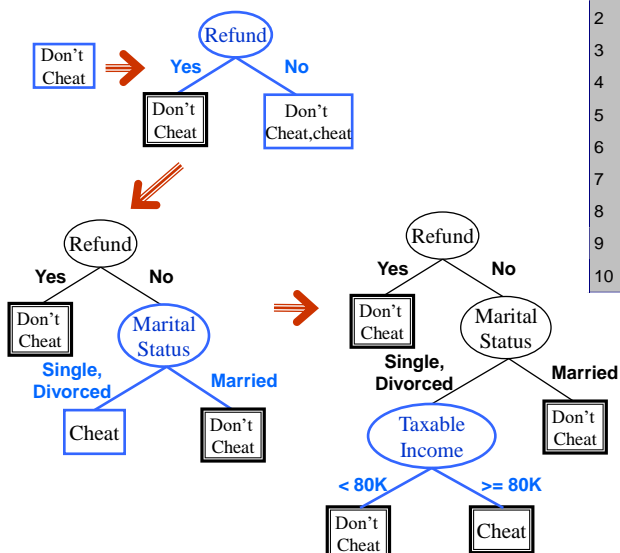
- Let D_t be the set of training records that reach a node t
- General Procedure:
 - If D_t contains records that belong the same class y_t , then t is a leaf node labeled as y_t
 - If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Hunt's Algorithm

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



3 b. Ans: Measures of Node Impurity

- Gini Index
- Entropy
- Misclassification error

Measure of Impurity: GINI

- Gini Index for a given node t :

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

(NOTE: $p(j|t)$ is the relative frequency of class j at node t).

- Maximum ($1 - 1/n_c$) when records are equally distributed among all classes, implying least interesting information
- Minimum (0.0) when all records belong to one class, implying most interesting information

C1	0
C2	6
Gini=0.000	

C1	1
C2	5
Gini=0.278	

C1	2
C2	4
Gini=0.444	

C1	3
C2	3
Gini=0.500	

Examples for computing GINI

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

C1	0
C2	6

$P(C1) = 0/6 = 0$ $P(C2) = 6/6 = 1$
Gini = $1 - P(C1)^2 - P(C2)^2 = 1 - 0 - 1 = 0$

C1	1
C2	5

$P(C1) = 1/6$ $P(C2) = 5/6$
Gini = $1 - (1/6)^2 - (5/6)^2 = 0.278$

C1	2
C2	4

$P(C1) = 2/6$ $P(C2) = 4/6$
Gini = $1 - (2/6)^2 - (4/6)^2 = 0.444$

Splitting Based on GINI

- Used in CART, SLIQ, SPRINT.
- When a node p is split into k partitions (children), the quality of split is computed as,

$$GINI_{split} = \sum_{i=1}^k \frac{n_i}{n} GINI(i)$$

where, n_i = number of records at child i ,
 n = number of records at node p .

Alternative Splitting Criteria based on INFO

- Entropy at a given node t :

$$Entropy(t) = -\sum_j p(j|t) \log p(j|t)$$

(NOTE: $p(j|t)$ is the relative frequency of class j at node t).

- Measures homogeneity of a node.
 - ◆ Maximum ($\log n_c$) when records are equally distributed among all classes implying least information
 - ◆ Minimum (0.0) when all records belong to one class, implying most information
- Entropy based computations are similar to the GINI index computations

Splitting Criteria based on Classification Error

- Classification error at a node t :

$$Error(t) = 1 - \max_i P(i | t)$$

- Measures misclassification error made by a node.
 - ◆ Maximum ($1 - 1/n_c$) when records are equally distributed among all classes, implying least interesting information
 - ◆ Minimum (0.0) when all records belong to one class, implying most interesting information

4 Ans: Backpropagation:

A neural network learning algorithm

Started by psychologists and neurobiologists to develop and test computational analogues of neurons

A neural network: A set of connected input/output units where each connection has a weight associated with it
During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples

The inputs to the network correspond to the attributes measured for each training tuple

Inputs are fed simultaneously into the units making up the input layer

They are then weighted and fed simultaneously to a hidden layer

The number of hidden layers is arbitrary, although usually only one

The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction

The network is feed-forward in that none of the weights cycles back to an input unit or to an output unit of a previous layer

From a statistical point of view, networks perform nonlinear regression: Given enough hidden units and enough training samples, they can closely approximate any function. Also referred to as connectionist learning due to the connections between units.

First decide the network topology: # of units in the input layer, # of hidden layers (if > 1), # of units in each hidden layer, and # of units in the output layer

Normalizing the input values for each attribute measured in the training tuples to [0.0—1.0]

One input unit per domain value, each initialized to 0

Output, if for classification and more than two classes, one output unit per class is used

Once a network has been trained and its accuracy is unacceptable, repeat the training process with a different network topology or a different set of initial weights

5. Ans:

Balanced Score Card: The term balanced scorecard (BSC) refers to a [strategic management performance metric](#) used to identify and improve various internal business functions and their resulting external outcomes. Used to measure and provide feedback to organizations, balanced scorecards are common among companies in the United States, the United Kingdom, Japan, and Europe. Data collection is crucial to providing quantitative results as managers and executives gather and interpret the information. Company personnel can use this information to make better decisions for the future of their organizations.

- A balanced scorecard is a performance metric used to identify, improve, and control a business's various functions and resulting outcomes.
- The concept of BSCs was first introduced in 1992 by David Norton and Robert Kaplan, who took previous metric performance measures and adapted them to include nonfinancial information.
- BSCs were originally developed for for-profit companies but were later adapted for use by nonprofits and government agencies.
- The balanced scorecard involves measuring four main aspects of a business: Learning and growth, business processes, customers, and finance.
- BSCs allow companies to pool information in a single report, to provide information into service and quality in addition to financial performance, and to help improve efficiencies.

ClickStream Analysis:

On a Web site, clickstream analysis (also called clickstream analytics) is the process of collecting, analyzing and reporting aggregate data about which pages a website visitor visits -- and in what order. The path the visitor takes through a website is called the clickstream.

There are two levels of clickstream analysis, traffic analytics and e-commerce analytics. Traffic analytics operates at the server level and tracks how many pages are served to the user, how long it takes each page to load, how often the user hits the browser's back or stop button and how much data is transmitted before the user moves on. E-commerce-based analysis uses clickstream data to determine the effectiveness of the site as a channel-to-market. It's concerned with what pages the shopper lingers on, what the shopper puts in or takes out of a shopping cart, what items the shopper purchases, whether or not the shopper belongs to a [loyalty program](#) and uses a coupon code and the shopper's preferred method of payment.

6. Ans:

Classification

predicts categorical class labels (discrete or nominal)

classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data

Prediction

models continuous-valued functions, i.e., predicts unknown or missing values

Typical applications

Credit approval

Target marketing

Medical diagnosis

Fraud detection

Issues: Data Preparation

Data cleaning

Preprocess data in order to reduce noise and handle missing values

Relevance analysis (feature selection)

Remove the irrelevant or redundant attributes

Data transformation

Generalize and/or normalize data

Issues: Evaluating Classification Methods

Accuracy

classifier accuracy: predicting class label

predictor accuracy: guessing value of predicted attributes

Speed

time to construct the model (training time)

time to use the model (classification/prediction time)

Robustness: handling noise and missing values

Scalability: efficiency in disk-resident databases

Interpretability

Understanding and insight provided by the model

Other measures, e.g., goodness of rules, such as decision tree size or compactness of classification rules.

7.Ans:

Linear regression is used for finding linear relationship between target and one or more predictors. There are two types of linear regression- Simple and Multiple.

Simple Linear Regression

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

Logistic regression is known and used as a linear classifier. It is used to come up with a *hyperplane* in feature space to separate observations that belong to a class from all the other observations that do *not* belong to that class. The decision boundary is thus *linear*. Robust and efficient implementations are readily available (e.g. scikit-learn) to use logistic regression as a linear classifier.

While logistic regression makes core assumptions about the observations such as IID (each observation is independent of the others and they all have an identical probability distribution), the use of a linear decision boundary is *not* one of them.

Logistic regression is an exercise in predicting (regressing to — one can say) discrete outcomes from a continuous and/or categorical set of observations. Each observation is independent and the probability p that an observation belongs to the class is some (& same!) function of the features describing that observation. Consider a set of n observations $[x_i, y_i; Z_i]$ where x_i, y_i are the feature values for the i th observation. Z_i equals 1 if the i th observation belongs to the class, and equals 0 otherwise. The likelihood of having obtained n such observations is simply the product of the probability $p(x_i, y_i)$ of obtaining each one of them separately.

8.Ans:

Data can be associated with classes or concepts. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived via (1) data characterization, by summarizing the data of the class under study (often called the target class) in general terms, or (2) data discrimination, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes), or (3) both data characterization and discrimination.

Concept description: » Characterization: provides a concise and succinct summarization of the given collection of data » Comparison: provides descriptions comparing two or more collections of data

Data generalization » A process which abstracts a large set of task-relevant data in a database from a low conceptual levels to higher ones.

Conceptual levels Approaches: •Data cube approach(OLAP approach) •Attribute-oriented induction approach

Characterization: Data Cube Approach

- Perform computations and store results in data cubes
- Strength
 - » An efficient implementation of data generalization
 - » Computation of various kinds of measures
 - e.g., count(), sum(), average(), max()
 - » Generalization and specialization can be performed on a data cube by *roll-up* and *drill-down*
- Limitations
 - » handle only dimensions of *simple nonnumeric data* and measures of *simple aggregated numeric values*.
 - » Lack of intelligent analysis, can't tell which dimensions should be used and what levels should the generalization reach

- Proposed in 1989 (KDD '89 workshop)
- Not confined to categorical data nor particular measures.
- How it is done?
 - » Collect the task-relevant data(*initial relation*) using a relational database query
 - » Perform generalization by attribute removal or attribute generalization.
 - » Apply aggregation by merging identical, generalized tuples and accumulating their respective counts.
 - » Interactive presentation with users.

9. Ans:

Several refinements have been proposed that focus on reducing the number of database scans, the number of candidate itemsets counted in each scan, or both. Partition - based Apriori is an algorithm that requires only two scans of the transaction database. The database is divided into disjoint partitions, each small enough to fit into available memory. In a first scan, the algorithm reads each partition and computes locally frequent itemsets on each partition. In the second scan, the algorithm counts the support of all locally frequent itemsets toward the complete database. If an itemset is frequent with respect to the complete database, it must be frequent in at least one partition. That is the heuristics used in the algorithm. Therefore, the second scan through the database counts itemset ' s frequency only for a union of all locally frequent itemsets. This second scan directly determines all frequent itemsets in the database as a subset of a previously defined union.

As the database size increases, sampling appears to be an attractive approach to data mining. A sampling - based algorithm typically requires two scans of the database. The algorithm first takes a sample from the database and generates a set of candidate itemsets that are highly likely to be frequent in the complete database. In a subsequent scan over the database, the algorithm counts these itemsets ' exact support and the support of their negative border. If no itemset in the negative border is frequent, then the algorithm has discovered all frequent itemsets. Otherwise, some superset of an itemset in the negative border could be frequent, but its support has not yet been counted. The sampling algorithm generates and counts all such potentially frequent itemsets in subsequent

database scans. Because it is costly to find frequent itemsets in large databases, incremental updating techniques should be developed to maintain the discovered frequent itemsets (and corresponding association rules) so as to avoid mining the whole updated database again.

CMAR is a classification method adopted from the FP growth method for generation of frequent itemsets. The main reason we included CMAR methodology in this chapter is its FP growth roots, but there is the possibility of comparing CMAR accuracy and efficiency with the C4.5 methodology. ASSOCIATIVE-CLASSIFICATION METHOD

Suppose data samples are given with n attributes (A_1, A_2, \dots, A_n) . Attributes can be categorical or continuous. For a continuous attribute, we assume that its values are discretized into intervals in the preprocessing phase. A training data set T is a set of samples such that for each sample there exists a class label associated with it. Let $C = \{c_1, c_2, \dots, c_m\}$ be a finite set of class labels. In general, a pattern $P = \{a_1, a_2, \dots, a_k\}$ is a set of attribute values for different attributes $(1 \leq k \leq n)$. A sample is said to match the pattern P if it has all the attribute values given in the pattern. For rule $R: P \rightarrow c$, the number of data samples matching pattern P and having class label c is called the support of rule R , denoted $\text{sup}(R)$. The ratio of the number of samples matching pattern P and having class label c versus the total number of samples matching pattern P is called the confidence of R , denoted as $\text{conf}(R)$. The association - classification method (CMAR) consists of two phases: 1. rule generation or training, and 2. classification or testing.

10. Ans: Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency. Big data has one or more of the following characteristics: high volume, high velocity or high variety.

Big Data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processed. To address the distinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and repurposing data. The upcoming sections explore a specific data analytics lifecycle that organizes and manages the tasks and activities associated with the analysis of Big Data. From a Big Data adoption and planning perspective, it is important that in addition to the lifecycle, consideration be made for issues of training, education, tooling and staffing of a data analytics team.

The Big Data analytics lifecycle can be divided into the following nine stages, as shown in [Figure 3.6](#):

1. Business Case Evaluation
2. Data Identification
3. Data Acquisition & Filtering
4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Results

Stage 1

