


CMR INSTITUTE OF TECHNOLOGY	USN <input type="text"/>	
-----------------------------------	--------------------------	---

Internal Assessment Test - III

Sub:	Data Warehousing & Data Mining	Code:	13MCA442
Date:	1.06.2017	Duration:	90 mins
		Max Marks:	50
		Sem:	IV
		Branch:	MCA

Answer Any FIVE FULL Questions

		Marks	OBE	
			CO	RBT
1(a)	List and explain the different data mining techniques used during the data preprocessing.	[10]	CO3	L2
2(a)	List and explain the other evaluation criteria for classification methods.	[10]	CO4	L2
3(a)	What is anti monotone property of support? Explain Apriori Algorithm with example.	[10]	CO3	L2
4(a)	Explain the types of attributes.	[5]	CO3	L2
(b)	What is nearest neighbor classifier? Explain.	[5]	CO3	L2

5(a)	Explain the FASMI characteristics of OLAP system.	[5]	CO1	L2
(b)	Explain Hierarchical Algorithm for clustering.	[5]	CO4	L2
6(a)	Explain types of clusters.	[10]	CO4	L2
7(a)	What is DBSCAN algorithm for clustering? Explain the limitations also.	[6+4]	CO4	L2
8(a)	Explain the statistical Approach for Outlier Analysis.	[10]	CO4	L2

Course Outcomes		PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8
CO1:	Describe the designing of Data Warehousing so that it can be able to solve the root problems.	1	-	3	2	-	-	3	3
CO2:	Understanding of the value of data mining in solving real-world problems.	2	2	3	2	-	-	2	3
CO3:	Understanding of foundational concepts underlying data mining.	2	3	3	1	-	-	3	3
CO4:	Understanding of algorithms commonly used in data mining tools.	2	-	3	2	-	-	3	3
CO5:	To develop further interest in research and design of new Data Mining techniques.	1	-	2	-	-	-	3	3

Cognitive level	KEYWORDS
L1	List, define, tell, describe, identify, show, label, collect, examine, tabulate, quote, name, who, when, where, etc.
L2	summarize, describe, interpret, contrast, predict, associate, distinguish, estimate, differentiate, discuss, extend
L3	Apply, demonstrate, calculate, complete, illustrate, show, solve, examine, modify, relate, change, classify, experiment, discover.
L4	Analyze, separate, order, explain, connect, classify, arrange, divide, compare, select, explain, infer.
L5	Assess, decide, rank, grade, test, measure, recommend, convince, select, judge, explain, discriminate, support, conclude, compare, summarize.

PO1 - Apply *knowledge*; PO2 - *Problem analysis*; PO3 - *Design/development of solutions*; PO4 - team work; PO5 - *Ethics*; PO6 - Communication; PO7- *Business Solution*; PO8 – Life-long learning;

Q1(a). List and explain the different data mining techniques used during the data preprocessing.

Ans: Sampling without replacement

- As each item is selected, it is removed from the population

Sampling with replacement

Objects are not **Aggregation: Combining two or more attributes (or objects) into a single attribute (or object)**

Purpose

- **Data reduction**
 - ◆ **Reduce the number of attributes or objects**
- **Change of scale**
 - ◆ **Cities aggregated into regions, states, countries, etc**
- **More “stable” data**
 - ◆ **Aggregated data tends to have less variability**

Sampling: Sampling is the main technique employed for data selection.

- **It is often used for both the preliminary investigation of the data and the final data analysis.**

Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.

Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

Types of Sampling

Simple Random Sampling

- There is an equal probability of selecting any particular item
- Two types
- removed from the population as they are selected for the sample.
 - ◆ In sampling with replacement, the same object can be picked up more than once

Stratified sampling

- Split the data into several partitions; then draw random samples from each partition

Dimensionality Reduction

Purpose:

- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise
 - 1 Techniques
- Principle Component Analysis
- Singular Value Decomposition
- Others: supervised and non-linear techniques

Feature Subset Selection

Another way to reduce dimensionality of data

- 1 Redundant features
 - duplicate much or all of the information contained in one or more other attributes
 - Example: purchase price of a product and the amount of sales tax paid
- 1 Irrelevant features
 - contain no information that is useful for the data mining task at hand
 - Example: students' ID is often irrelevant to the task of predicting students' GPA

Techniques:

- **Brute-force approach:**
- Try all possible feature subsets as input to data mining algorithm
- **Embedded approaches:**
- Feature selection occurs naturally as part of the data mining algorithm
- **Filter approaches:**
- Features are selected before data mining algorithm is run
- **Wrapper approaches:**
- Use the data mining algorithm as a black box to find best subset of attributes

Discretization and Binarization

- **Binarization** : converting continuous and discrete attributes into one or more binary attributes.
- For m categorical values use [0,m-1] values to assign
- Convert each of these m integers to a binary numbers.
- **Discretization:** process of transforming a continuous attribute into a categorical attribute.
- Two types:
- Supervised – using class information
- Ex. Entropy based discretization
- Unsupervised- not using class information
- Two types: Equal width and equal Frequency

Q2(a) List and explain the other evaluation criteria for classification methods.

Ans: **Other Evaluation Criteria for Classification Methods**

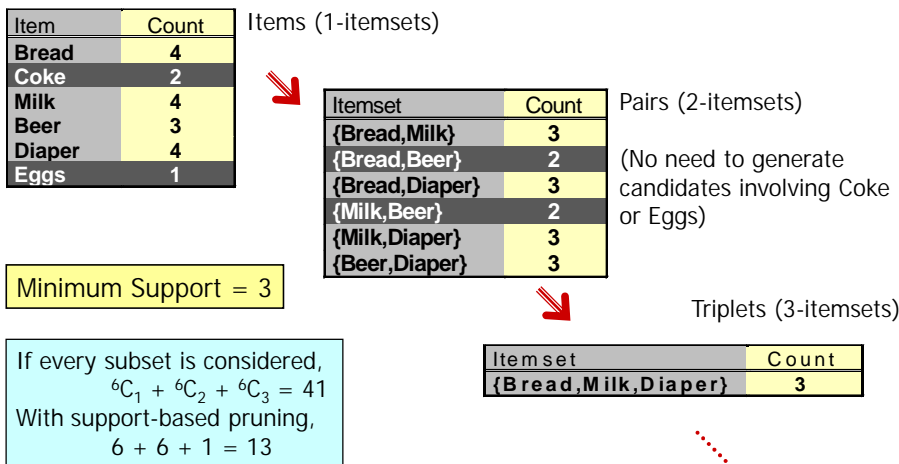
- Speed
- Robustness
- Scalability
- Interpretability
- Goodness of the model
- Flexibility
- Time complexity
- Speed includes
 - Time or computation cost of constructing a model
 - Time required to learn to use the model.
 - Aim- to minimize both times.
- Robustness
 - Method be able to produce good results in spite of some errors and missing values in datasets.
- Scalability
 - Method continues to work efficiently for large disk-resident databases as well.
- Interpretability
 - End-user be able to understand and gain insight from the results produced by the classification method.
- Goodness of the model
 - It needs to fit the problem that is being solved.

Q3(a). What is anti monotone property of support? Explain Apriori Algorithm with example.

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: $\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$
- 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

Q4(a). Explain the types of attributes.

1 Ans: There are different types of attributes

- Nominal
 - ◆ Examples: ID numbers, eye color, zip codes
- Ordinal
 - ◆ Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
- Interval

- ◆ Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- Ratio
 - ◆ Examples: temperature in Kelvin, length, time, counts

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: { <i>male, female</i> }	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, { <i>good, better, best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

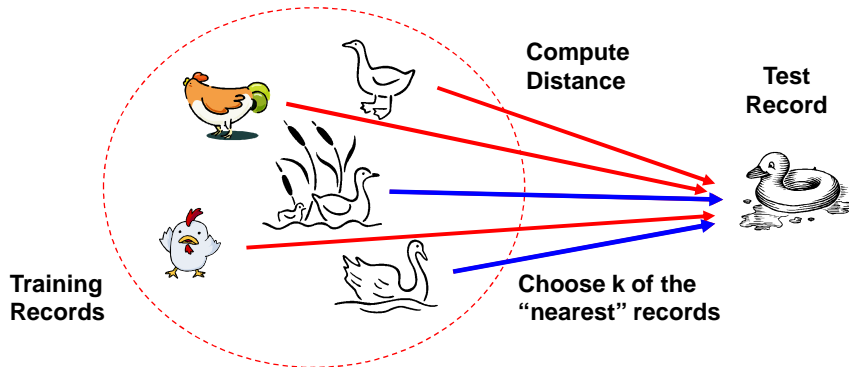
Q4(b). What is nearest neighbor classifier? Explain.

Ans: **Nearest Neighbor Classifiers:**

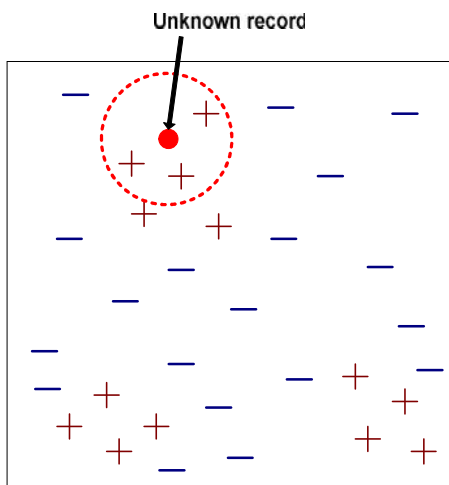
Nearest Neighbor Classifiers

- Basic idea:

- If it walks like a duck, quacks like a duck, then it's probably a duck



Nearest-Neighbor Classifiers



- Requires three things
 - The set of stored records
 - Distance Metric to compute distance between records
 - The value of k , the number of nearest neighbors to retrieve
- To classify an unknown record:
 - Compute distance to other training records
 - Identify k nearest neighbors
 - Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

Nearest Neighbor Classification

- Compute distance between two points:
 - Euclidean distance

$$d(p, q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Determine the class from nearest neighbor list
 - take the majority vote of class labels among the k-nearest neighbors
 - Weigh the vote according to distance
 - ◆ weight factor, $w = 1/d^2$

Q5(a). Explain the FASMI characteristics of OLAP system.

Ans: **FASMI Characteristics**

- *Fast*: OLAP queries should be answered very quickly, perhaps within seconds.
- *Analytic*: An OLAP system must provide rich analytic functionality and it is expected that most OLAP queries can be answered without any programming.
- *Shared*: An OLAP system is a shared resource although it is likely to be accessed only by a select group of managers. Being a shared system, an OLAP system should provide adequate security for confidentiality as well as integrity.
- *Multidimensional*: It must provide a multidimensional conceptual view of the data. A dimension often has hierarchies that show parent/child relationships between the members of a dimension. The multidimensional structure should allow such hierarchies.
- *Information*: OLAP system usually obtain information from a data warehouse. The system should be able to handle a large amount of input data.

Q5(b). Explain Hierarchical Algorithm for clustering.

1 Ans: Two main types of hierarchical clustering

- Agglomerative:
 - ◆ Start with the points as individual clusters
 - ◆ At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
- Divisive:
 - ◆ Start with one, all-inclusive cluster
 - ◆ At each step, split a cluster until each cluster contains a point (or there are k clusters)
 - 1 Traditional hierarchical algorithms use a similarity or distance matrix
- Merge or split one cluster at a time

Agglomerative Clustering Algorithm

- More popular hierarchical clustering technique

Basic algorithm is straightforward
Compute the proximity matrix
Let each data point be a cluster

Repeat

Merge the two closest clusters
Update the proximity matrix

Until only a single cluster remains

Key operation is the computation of the proximity of two clusters

Different approaches to defining the distance between clusters distinguish the different algorithms

Q6(a). Explain types of clusters.

Ans: Types of Clusters:

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function

- Well-Separated Clusters:
 - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.
- Center-based
 - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster
 - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most “representative” point of a cluster
- Contiguous Cluster (Nearest neighbor or Transitive)
 - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.
- Density-based
 - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
 - Used when the clusters are irregular or intertwined, and when noise and outliers are present.
- Shared Property or Conceptual Clusters
 - Finds clusters that share some common property or represent a particular concept.

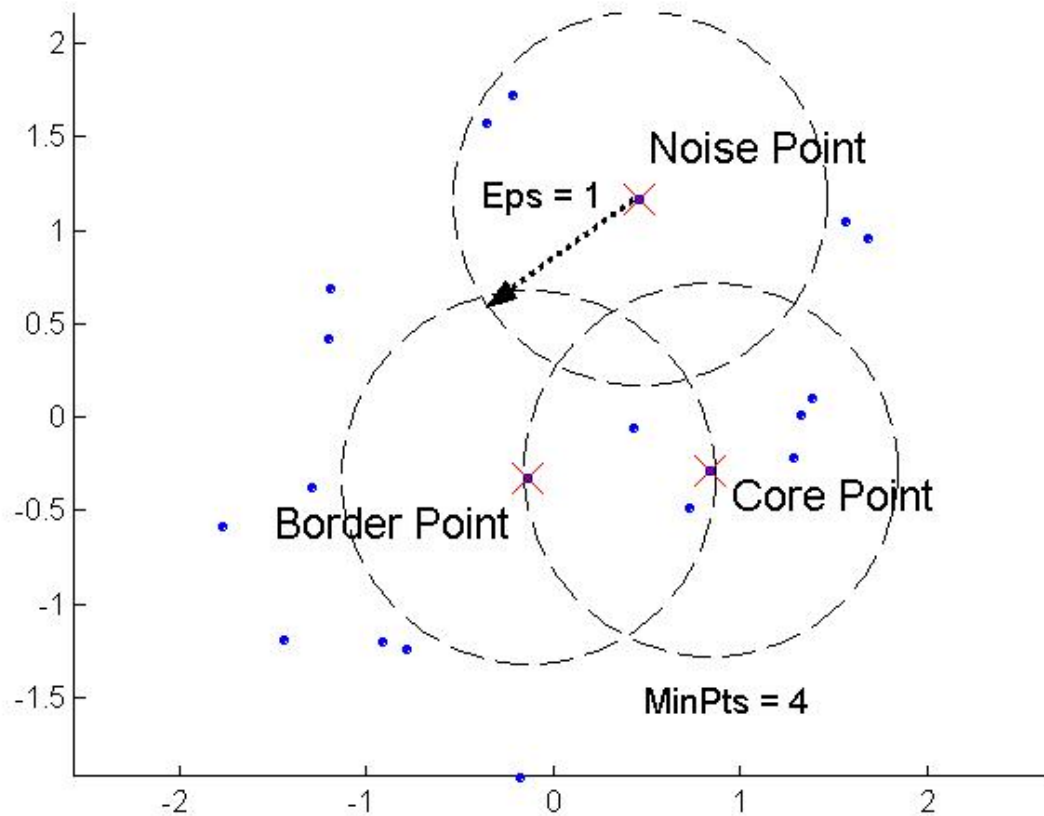
$O(N^2)$ space since it uses the proximity matrix.

- N is the number of points.
- $O(N^3)$ time in many cases
- There are N steps and at each step the size, N^2 , proximity matrix must be updated and searched
- Complexity can be reduced to $O(N^2 \log(N))$ time for some approaches

Q7(a). What is DBSCAN algorithm for clustering? Explain the limitations also.

1 Ans: DBSCAN is a density-based algorithm.

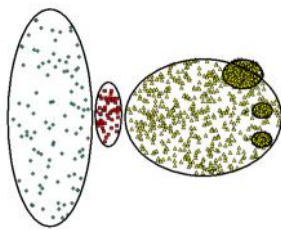
- Density = number of points within a specified radius (Eps)
- A point is a core point if it has more than a specified number of points (MinPts) within Eps
 - ◆ These are points that are at the interior of a cluster
- A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point
- A noise point is any point that is not a core point or a border point.



DBSCAN Algorithm

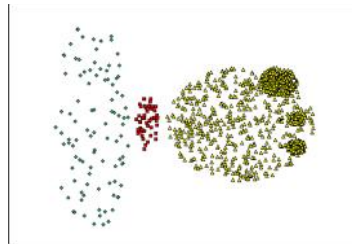
- Label all points as core, border, or noise points.
- Eliminate noise points
- Put an edge between all core points that are within Eps of each other.
- Make each group of connected core points into a separate cluster.
- Assign each border point to one of the clusters of its associated core points.

When DBSCAN Does NOT Work Well

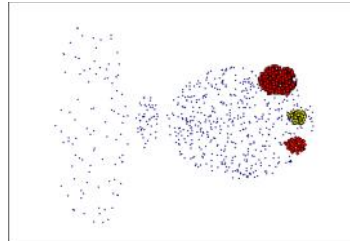


Original Points

- Varying densities
- High-dimensional data



(MinPts=4, Eps=9.75).

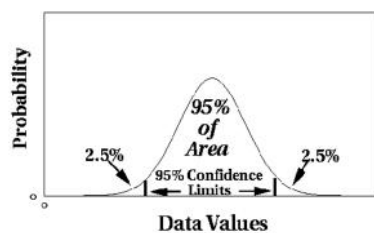


(MinPts=4, Eps=9.92)

Q8(a). Explain the statistical Approach for Outlier Analysis.
Ans:

Statistical Approaches

- Assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Apply a statistical test that depends on
 - Data distribution
 - Parameter of distribution (e.g., mean, variance)
 - Number of expected outliers (confidence limit)



Statistical-based – Likelihood Approach

- Assume the data set D contains samples from a mixture of two probability distributions:
 - M (majority distribution)
 - A (anomalous distribution)
- General Approach:

- Initially, assume all the data points belong to M
- Let $L_t(D)$ be the log likelihood of D at time t
- For each point x_t that belongs to M, move it to A
 - Let $L_{t+1}(D)$ be the new log likelihood.
 - Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
 - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

Statistical-based – Likelihood Approach

- Data distribution, $D = (1 - \lambda) M + \lambda A$
- M is a probability distribution estimated from data
 - Can be based on any modeling method
 - A is initially assumed to be uniform distribution
- Likelihood at time t:

$$L_t(D) = \prod_{i=1}^N P_D(x_i) = \left((1 - \lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1 - \lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

Limitations of Statistical Approaches

- Most of the tests are for a single attribute
- In many cases, data distribution may not be known
- For multi-dimensional data, it may be difficult to estimate the true distribution