II. B.Tech - I - Semester End Examinations (Regular)

DATA MINING

Common for IT, CSD, CSM.

Part - A

1. a) Defne Data Integration

Data integration is the merging of data from multiple data stores. It can help reduce and avoid redundancies and inconsistencies in the resulting data set. It can improve the accuracy and speed of the subsequent data mining process.

b) What are data objects and Attributes types.

A data object represents an entity - in a sales database, the objects may be customers, store items and sales. Data objects are typically described by attributes.

Examples : samples, instances, data points.

Attribute types :- An attribute is a data field representing a characteristic or feature of a data object. The type of attributes are 1. Nominal attributes 2. Binary Attributes and 3. Ordinal Attributes.

c. Explain Market Basket Analysis.

It is a typical example of frequent itemset mining is market basket analysis. This process analyzes customer buying habits by finding association between the different items that customer place in their shopping baskets. which contains items that are frequently purchased togethres by customers.

d. What is FP Growth?

We can mine the complete set of frequent itemsets without such a costly candidate generation procers. In FP growth tree, it compresses the database representing frequent items into a frequent pattern tree or FP-tree, which retains the itemset-association information.

e. What is FP Define Neural Network.

A neural network is a method in artificial intellgence (AI) that teaches computers to process data in a way that is inspired by the human brain. It uses interconnected nodes or neurons in a layered structure that resembles the human brain.

i) Describe Temporal Mining.

Temporal mining is the process of discovering hidden patterns, trends and relationships in time dependent data to understand how things evolve over time, using techniques from machine learning. It focuses on patterns that occur over time, unlike traditional mining that might look at static data, aiming to uncover insight for forecasting, anomaly detection and trend analysis in dynamic system.

ii) Defn Pattern Detection.

Pattern detection is the process of identifying meaningful regularities, structures, or recurring arrangements in data, enabling systems to classify, predict or make decisions by recognizing familiar features or sequence. for interpreting images, text data used in daily life. It finds hidden relationships from simple correlations to complex, multi-variable interactions.

f) What is the use of k-Nearest Neighbor classifies?

The KNN classifier, classifies new data points by finding their 'k' closest neighbors in the training data and assigning the common class among them. It uses for classification problems, Regression, Ranking/ Recommendation system and in pattern recognition.

g) List out the clustering methods.

clustering is the proces of partitioning a set of data objects into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other cluster.

clustering methods are

1. Partitioning methods
2. Hierarchical methor : Agglomerative or divisive.
3. Density-based method
4. Grid-based method.

h) List out the types of Outlier Analysis.

1) Global outlier. : it deviates significantly from the rest of the dataset.

2) Contextual outliers : — Its depends on the time and location.

3) Collective Outliers : if the objects as a whole deviate significantly from the entire data set.

Part-B.

2. What is Data Reprocessing? Explain need for data preprocessing and its techniques.

Data preprocessing is the essential process of cleaning, transforming, and organizing raw data into a usable format, making it suitable for analysis.

The major steps involved in data preprocessing, namely data cleaning, data integration, data reduction and data transformation.

1) Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies. The noisy data can cause confusion for the mining procedure, resulting in unreliable output. The useful preprocessing step is to run your datata through some data cleaning methods.

2) Data integration: It would involve integrating multiple databases, data cubes or files. Yet some attributes representing a given concept may have different databases, causing inconsistencies and

redundancies. In data cleaning, steps must be taken to help avoid redundancies during data integration. Typically, data cleaning and data integration are performed as a pre-procesing step when preparing data.

3. Data Reduction: It is a technique that can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. Mining on the reduced data set should be more efficient yet produce the same analytical results. Data reduction strategies include dimensionality reduction, numerosity reduction and data compression.

4. Data Transformation and Data Discretization.

Data transformation include the following:

1. Smoothing which works to remove noise from the data.
2. Attribute Construction, where new attributes are constructed and added from the given set of attributes.
3. Aggregation: It- Constructs a data cube for data analysis at multiple abstraction levels.
4. Normalization, where the attribute data are scaled so as to fall within a smaller range.

3. Explain KDD process in detail. (10)

Data mining should have been more appropriately named "knowledge mining from data", which is unfortunately somewhat long. Knowledge mining may not reflect the emphasis on mining from large amounts of data.

The term knowledge discovery from data or KDD, while others view data mining as merely an essential step in the process of knowledge discovery.

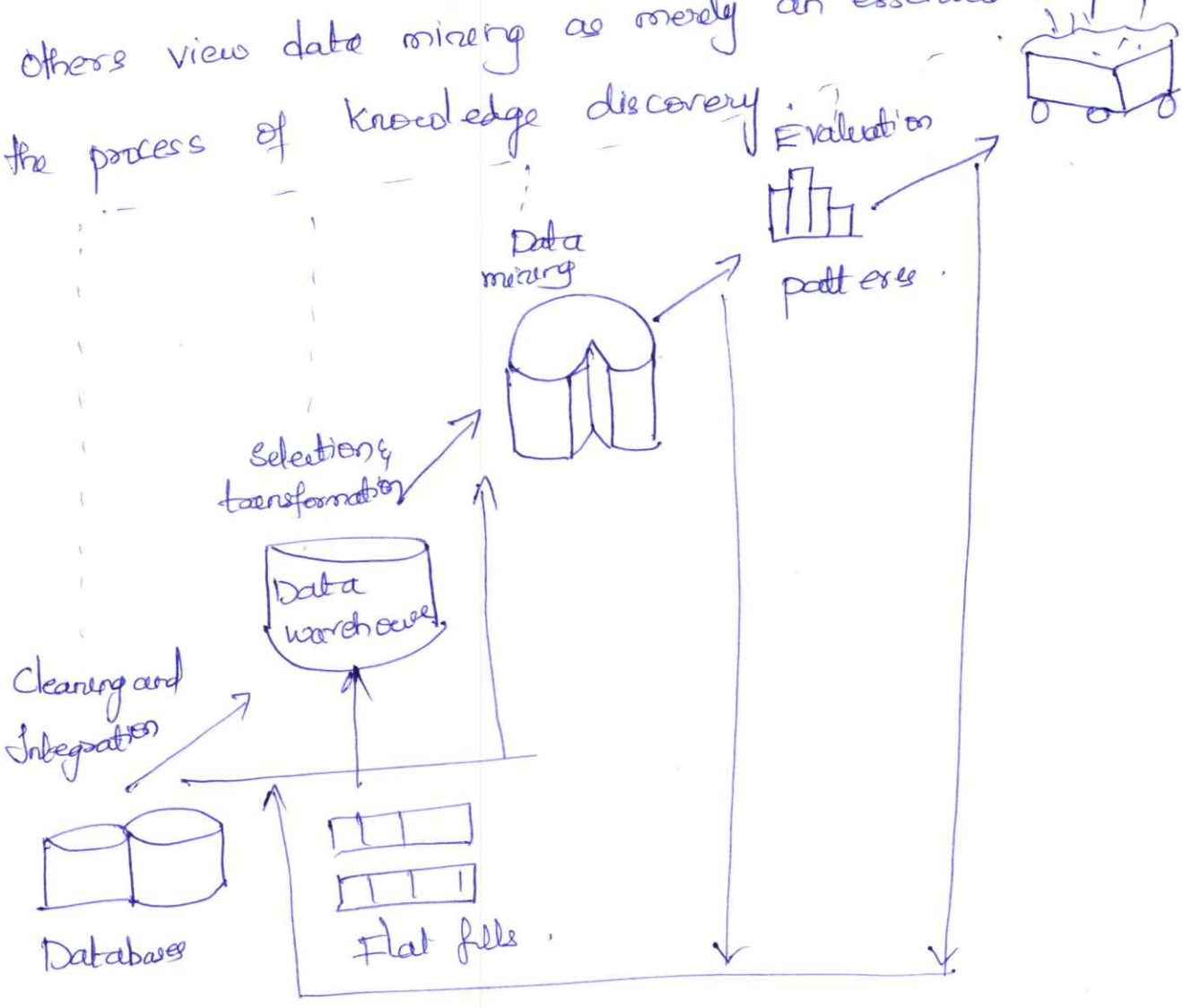


Fig: Data mining as a step in the process of knowledge discovery.

1. Data cleaning : (to remove noise and inconsistent data)

2. Data integration : where multiple data sources may be combined.

3. Data selection : where data relevant to the analysis task are retrieved from the database.

4. Data transformation : where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.

5. Data mining : an essential process where intelligent methods are applied to extract data patterns.

6. Pattern evaluation : an essential process where intelligent methods are applied to extract data patterns.

7. Knowledge Presentation : where visualization and knowledge representation techniques are used to present mined knowledge to user.

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.

4. Describe scalable methods for mining Frequent patterns.

The basic algorithm for finding frequent itemsets is the Apriori Algorithm.

Apriori is based on the fact that use prior knowledge of frequent itemset properties. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1) itemsets.

First the set of frequent 1-itemset is found by scanning the database to accumulate the count for each item and collecting those items that satisfy minimum support. The resulting set is denoted by L_1

Next L_1 is used to find L_2, the set of frequent 2-itemsets, which is used to find L_3 and so on.

Apriori property: All nonempty subsets of a frequent itemset must also be frequent.

A two-step process is followed

1. The join step: To find L_k, a set of candidate k-itemset is generated by joining L_{k-1} with itself.
2. Prune step: C_k is a superset of L_k, all candidates having a count no less than the minimum support count are frequent.

Generating association rule.

$$\text{Confidence}(A \Rightarrow B) = P(B/A) = \frac{\text{Support-Count}(A \cup B)}{\text{Support-count}(A)}$$

2. FP-growth :-

Concept : Compresses the dataset into a compact FP-tree (Frequent Pattern Tree) to avoid generating large candidate sets, making it faster than Apriori.

Process : It Builds the tree by scanning the data, then mines patterns directly from the tree, recursively partitioning the problem.

Scalability : It is highly efficient for large data sets, especially those with many transactions and forms the base.

5 a) Write the applications and advantages of Market Basket Analysis.

A data mining technique that is used to uncover purchase patterns in any retail setting is known as Market Basket Analysis. Market basket analysis mainly works with the Association Rule if → then.

- if means Antecedent: An antecedent is an item found within the data.
- Then means Consequent: A Consequent is an item found in combination with the antecedent.

Applications of Market Basket Analysis:

1. Retail: It is used in the retail sector to examine consumes buying patterns. and inform decisions about product placement, inventory management and pricing tactics.
2. E-Commerce: It can help online merchants better understand the customer buying habits and make data-driven decisions about product recommendations.

3. finance : It can be used to evaluate investor behaviours and forecast the types of investment items that investors will likly buy in future.

4. Telecommunication : To evaluate consumer behaviour and make data driven decisions about which goods and services to provide the telecommunications busines.

5. Manufacturing : To evaluate consumes behaviour and make data-driven decisions about which products to produe and which materials to employ in the production process.

Advantages of Market Basket Analysis.

1. Enhanced customer Understanding.
2. Improved Inventory Management.
3. Better Pricing Strategies.
4. Sales Growth.

5(b) Discuss the FP growth Algorithm with an example.

Frequent pattern Growth algorithm is an efficient method in data mining used to discover frequent itemsets in a dataset without generating candidate sets, a key advantage over algorithms like Apriori.

Example.

TID	List of items_IDs
T100	I_1, I_2, I_5
T200	I_2, I_4
T300	I_2, I_3
T400	$I_1\ I_2\ I_4$
T500	I_1, I_3
T600	$I_2\ I_3$
T700	$I_1\ I_3$
T800	$I_1\ I_2\ I_3\ I_5$
T900	$I_1\ I_2\ I_3$.

First create the root of the tree, labeled with "null".

Scan a database D. The items in each transaction are processed in L order.

i) The scan of the first transaction. "T100 : I_1, I_2, I_5", Contains three items I_2, I_1, I_5 in Lorded., and Construct $\langle I_2 : 1 \rangle$, $\langle I_1 : 1 \rangle$ and $\langle I_5 : 1 \rangle$.

Second transaction, T200 : $\langle I_2, I_4 \rangle$. we increment the count of I_2 node by 1 and create a new node $\langle I_4 : 1 \rangle$. which is linked as a child to $\langle I_2 : 2 \rangle$.

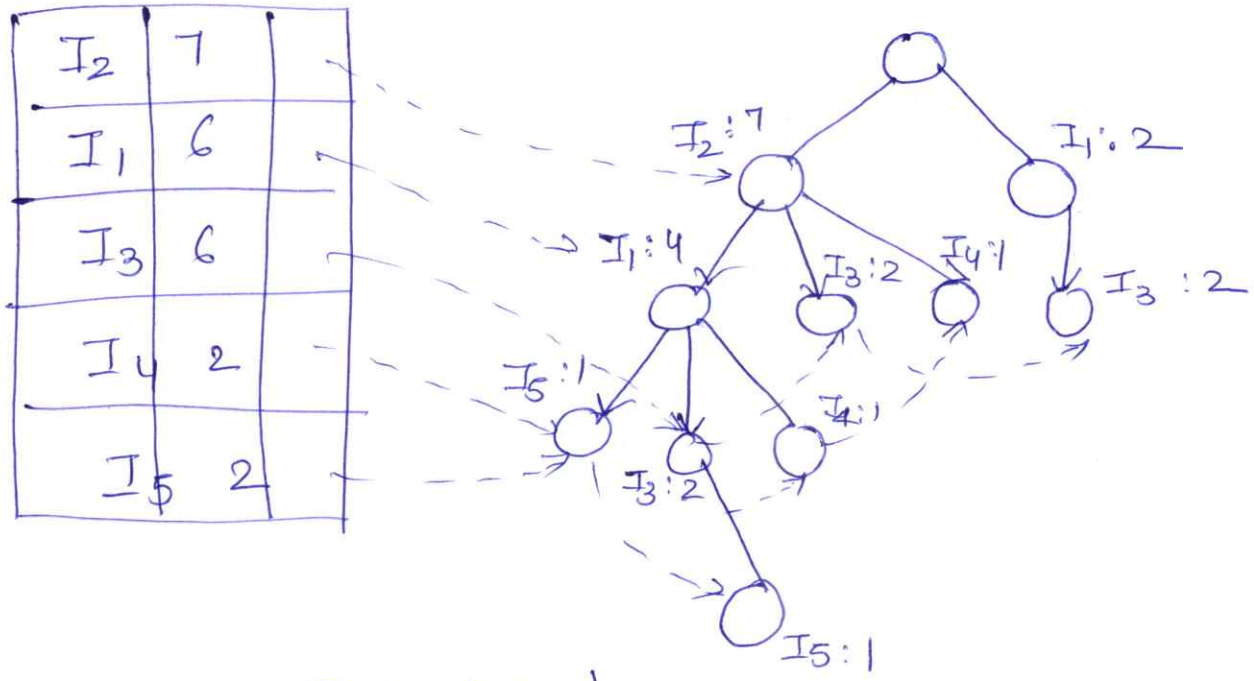


Fig: FP-tree.

Item	Conditional Pattern Base	Conditional FP-tree	Frequent Pattern Generated
I_5	$\{I_2, I_1 : 1\}$ $\{I_2\ I_1, I_3 : 1\}$	$\{I_2 : 2, I_1 : 2\}$	$\{I_2, I_5 : 2\}$ $\{I_1, I_5 : 2\}$ $\{I_2\ I_1 : 2\}$
I_4	$\{I_2, I_1 : 1\}$ $\{I_2 : 1\}$	$\{I_2 : 2\}$	$\{I_2, I_4 : 2\}$
I_3	$\{I_2, I_1 : 2\}$, $\{I_2 : 2\}$, $\{I_1 : 2\}$	$\langle I_2 : 4\ I_1 : 2 \rangle$ $\langle I_1 : 2 \rangle$	$\{I_2\ I_3 : 4\}$, $\{I_1, I_3 : 4\}$
I_1	$\{I_2 : 4\}$	$\langle I_2 : 4 \rangle$	$\{I_2, I_1 : 4\}$.

6. What are the metrics for evaluating classifies Performance? Explain.

Ans:- The classifies revaluation measures presented in the table.

Measure	Formula.
Accuracy, recognition rate	$\frac{TP + TN}{P+N}$
error rate, misclassification rate	$\frac{FP + FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F1, F-score harmonic mean of precision and recall.	$\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$
F_β, where β is a non-negative real number	$\frac{(1+\beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$

Accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

$$\text{Accuracy} = \frac{TP + TN}{P + N}.$$

$$\text{Error rate} = \frac{FP + FN}{P + N}.$$

True positives (TP): It refers to the positive tuples that were correctly labeled by the classifier.

True negatives (TN): These are the negative tuples that were correctly labeled by the classifier.

False positives (FP): These are the negative tuples that were incorrectly labeled as positive.

False negatives: These are the positive tuples that were mislabeled as negative.

7 a) Describe K-Nearest-Neighbour classifier.

K nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These K training tuples are the k-nearest neighbors of the unknown tuple.

"closeness" is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say $X = (x_{11}, x_{12} \cdots x_{1n})$ and $X_2 = (x_{21}, x_{22}, \cdots x_{2n})$ is.

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

KNN algorithm. Computing 3 nearest-neighbor for tuple $t_4 = (3, 7)$

TID	a_1	a_2	o/p	Distance	Neighbor Rank.
I_1	7	7	0	4	3
I_2	7	4	0	5	4
I_3	3	4	1	3	1
I_4	1	4	1	3.6	2.

Steps:

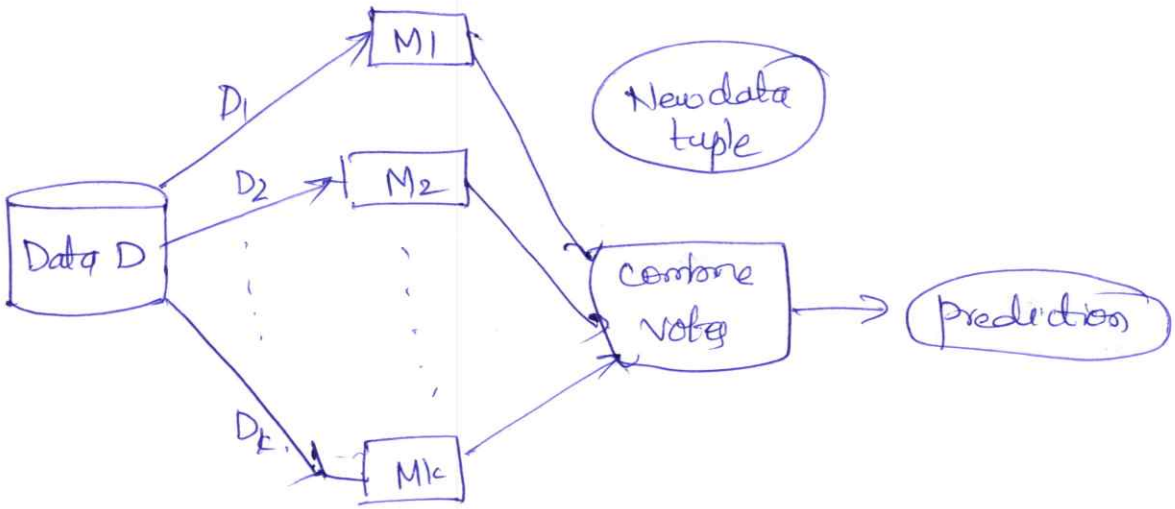
1. Select the number of neighbors (K)
2. Calculate distance : Euclidean distance
3. Find the k nearest neighbors.
4. Gather categories.
5. Assign the class.

7 (b) Illustrate ensemble methods and its applications.

An ensemble for classification is a composite model, made up of a combination of classifier.

Bagging, boosting and random forests are examples of ensemble methods.

An ensemble combines a series of k learned models, $M_1, M_2 \ldots M_k$, with the aim of creating an improved composite classification model M^*.



1. Bagging : Models are trained independently on different random subsets of the training data. Their results are then combined usually by averaging or voting. This helps reduce variance and prevents overfitting.

Boosting is an ensemble technique that combines multiple weak learners to create a strong learner. Weak models are trained in series such that each next model tries to correct errors of the previous model until the entire training dataset is predicted correctly.

8. Explain about Hierarchical Methods.

A hierarchical clustering method works by grouping data objects into a hierarchy ~~is use~~ or "tree" of clusters. Representing data objects in the form of a hierarchy is useful for data summarization and visualization.

1. Agglomerative versus Divisive Hierarchical clustering (AGNES)

An agglomerative hierarchical clustering method uses a bottom-up strategy. It typically starts by letting each object form its own cluster and iteratively merges clusters into larger and larger clusters until all the objects are in a single cluster or certain termination conditions are satisfied.

A divisive hierarchical clustering method employs a top-down strategy. It starts by placing all objects in one cluster, which is the hierarchy's root. It then divides the root cluster into several smaller subclusters and recursively partitions those clusters into smaller ones. The partitioning process continues until each cluster at the lowest level is coherent enough.

2. BIRCH : Multiphase Hierarchical clustering :

Balanced Iterative Reducing and clustering using Hierarchies (BIRCH) is designed for clustering a large amount of numeric data by integrating hierarchical clustering and other clustering methods such as iterative partitioning. It overcomes the two difficulties in agglomerative clustering methods :

1. scalability.
2. Inability to undo what was done in the previous step.

BIRCH uses the notions of clustering feature to summarize a cluster and clustering feature tree to prepresent a cluster hierarchy.

3. Chameleon : Multiphase Hierarchical clustering using Dynamic modeling :

Chameleon is a hierarchical clustering algorithm that uses dynamic modeling to determine the similarity between pairs of clusters. The similarity is assessed based on (1) how well connected objects are within a cluster and (2) the proximity of clusters. The two clusters are merged if their interconnectivity is high and they are close together.

9 a) Discuss about basic clustering Methods.

clustering is the process of partitioning a set of data objects into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as clustering.

There are many clustering algorithms.

i) Partitioning methods :- Given a set of n objects, a partitioning method constructs k partitions of the data where each partition represents a cluster and $k \leq n$. that is, it divides the data into k groups such that each group must contain at least one object.

* K-Means : A Centroid-Based Technique.

Suppose a dataset, D, contains n objects in Euclidean space. Partitioning methods distribute the objects in D into k clusters, $C_1, \ldots C_k$, that is, $C_i \subset D$ and $C_i \cap C_j \neq \emptyset$ for $(1 \leq i, j \leq k)$.

* The difference between an object $p \in C_i$ and c_i. is measured by $dist(p, c_i)$, where $dist(x, y)$ is the Euclidean distance between two point x and y.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(P, C_i)^2 .$$

② k-Medoids Algorithm.

Suppose that the data mang task is to cluster points into three clusters.

The distance function is Euclidean distance.

9b) Explain density based method of DBSCAN.

Desity based clustering is a clustering method based on a set of density distribution functions. Density estimation is the estimation of an unobservable underlying probability density function based on a set of observed data.

Apply DBSCAN algorithm to the given data points create clusters with min points = 4 and epsilon (ε) = 1.9

Data points.

P_1: (3, 7) P_2: (4, 6) P_3: (5, 5) P_4: (6, 4) P_5: (7, 3)

P_6: (6, 2) P_7: (7, 2) P_8: (8, 4) P_9: (3, 3), P_{10}: (2, 6)

P_{11} (3, 5) P_{12}: (2, 4)

use Euclidean distance. between two points

, is $\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}
P_1	0											
P_2	1.41	0										
P_3	2.83	1.41	0									
P_4	4.24	2.83	1.41	0								
P_5	5.66	4.24	2.83	1.41	0							
P_6	5.83	4.47	3.16	2.00	1.41	0						
P_7	6.40	5.00	3.61	2.24	1.00	1.00	0					
P_8	5.83	4.47	3.16	2.00	1.41	2.83	2.24	0				
P_9	4.00	3.16	2.83	3.16	4.00	3.16	4.12	5.10	0			
P_{10}	1.41	2.00	3.16	4.47	5.83	5.66	6.40	6.32	3.16	0		
P_{11}	2.00	1.41	2.00	3.16	4.47	4.24	5.00	5.10	2.00	1.41	0	
P_{12}	3.16	2.83	3.16	4.00	5.10	4.47	5.39	6.00	1.41	2.00	1.4	0

P_1 : P_2, P_{10}

P_2 : P_1, P_3, P_{11}

P_3 : P_2, P_4

P_4 : P_3, P_5

P_5 : P_4, P_6, P_7, P_8

P_6 : P_5, P_7

P_7 : P_5, P_6

P_8 : P_5

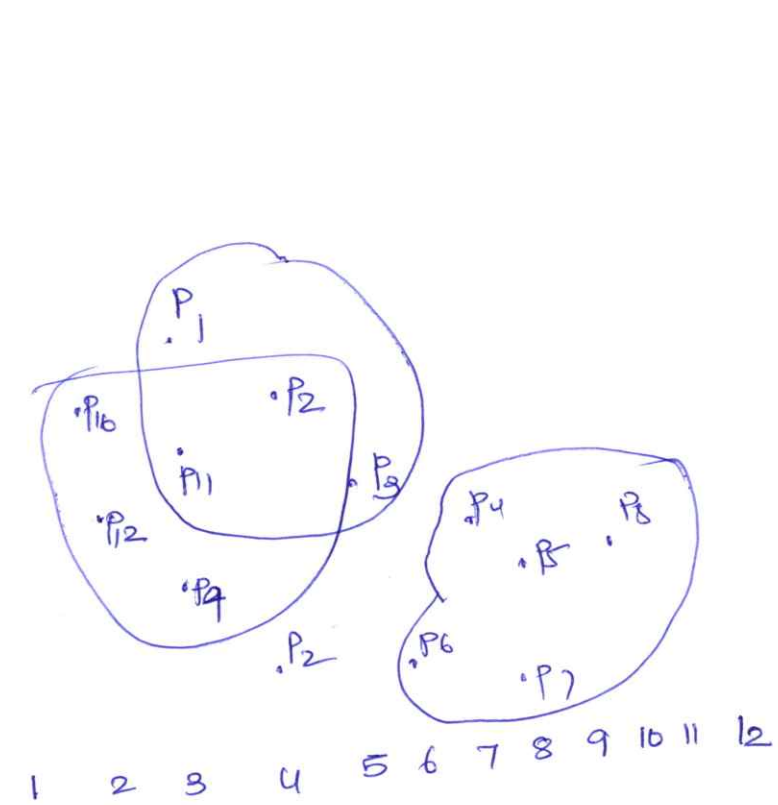
P_9 : P_{12}

P_{10} : P_1, P_{11}

P_{11} : P_2, P_{10}, P_{12}

P_{12} : P_9, P_{11}

Graph :



10. Discuss about Spatial Rules and Classification Algorithm.

Spatial rules are core components of spatial data mining which aims to uncover patterns and relationships in data that are tied to a geographic location.

Spatial Rules:

Spatial rules describe inherent relationships, patterns and constraints within the spatial data that are not explicitly stored in a database

Key types of spatial rules include:

1. Spatial Association Rules: These rules indicate the co-occurrence or spatial proximity of different features or events. A classic example is a rule that states "if a house is near the coast, then its price range is likely high". to define relationships.

2. Spatial Characteristic Rules: These provide a general description of a specific set of spatially related data.

3. Spatial Discriminant Rules: These highlight the distinguishing features of one class of spatial data compared to other classes.

Spatial classification Algorithms.

1. Decision Trees : These are used for spatial classification where the splitting and pruning strategies are adapted to consider spatial attributes or relationships.

2. k-Nearest Neighbour : A common algorithm in spatial analysis where the class of an object is determined by its nearest neighbors, inherently using spatial proximity.

3. Random Forest : An ensemble method that combines multiple decision trees and is robust for spatial analysis tasks like land cover classification using remote sensing.

4. Support Vector Machines : used for classification tasks in spatial data, particularly effective in handling high-dimensional data such as image recognition in remote sensing.

11 Explain Temporal Mining Modeling and events.

Temporal mining in data mining discovers patterns trends and relationships in time ordered data by analyzing sequence of events, using models such as Markov Models or sequential patterns to understand how things change over time for forecasting, anomaly detection and better decision making in real time systems.

* Temporal Data : Data points linked to specific times including time-series or event sequences.

* Events : Discrete occurences within the data often time-stamped, that change a system's state.

* Temporal Patterns : Recurring sequences or relationships between events, like customers buying x then y.

* Temporal Modeling : Building models to represent these time-dependent processes, predicting future states or behaviors.

1. Data transformation: Raw time-stamped data is converted into a usable format, often sequences or time series.

2. Similarity Measures: Defining how similar two time sequences or patterns are crucial for finding approximate matches, not just exact ones.

3. Mining Operations like sequential pattern mining, Trend Analysis, Anomaly detection & forecasting and Association Rule Mining.

4. Modeling: Using techniques like Markov chains or other methods to formalize these patterns for prediction.

Scheme of Evaluation.

III B Tech I sem End Examination

Datamining.

Part - A

1. a) Definition — 2M
   b) Data objects — ½ M
      Data attribute — ½ M
   c) Definition — 1M
   d) Definition — 1M
   e) definition — 1M
   f) Uses — Any 3 — 1M
   g) List — Any 2 — 1M
   h) 3 types — 1M
   i) Definition — 1M
   j) Definition — 1M.

Part - B.

2. Definition — 2M.
   Steps — 8M

3. Diagram — 4M
   Description — 6M.

4. Description — 10 M

5. a) Any 5 applications & uses — 5M

b) FP growth — Defnition — 2M
Numerical Example — 3M

6. Description — 10M

7 a) Description — 5M.

b) Description — 5M

8. List & explain 3 types of methods — 10 M

9 a) Description — 5M

b) Description and Example — 2M & 3M

10. Description — 10 M

11. Description — 10M.