Data Mining Notes

- It is about extracting hidden information(useful information) from huge amount of data.
- Data mining is looking for hidden, valid and potentially useful patterns in huge datasets.
- It is all about discovering unsuspected or previous unknown relationships amongst the data. It is inter-disciplinary i.e. uses machine learning, statistics, database, AI, etc.

Basics tasks of data mining

1. **Classification**: It maps data onto pre-defined groups or classes. It is also called as supervised learning. The classes are decided based on the characteristics of the data already belonging to the class.
2. **Pattern Recognition**: It is a type of classification where a given pattern is classified into one of the several classes based on its similarity with pre-defined patterns.
3. **Regression**: It maps data to real valued prediction variable, this function assumes that the target data fits into some known function and tries to find out best function that models the given data. Error analysis is used to determine which function is best.
4. **Prediction**: In many data mining applications future data is predicted based on current or past data, for example, prediction of heavy rains, etc.
5. **Clustering**: It is similar to classification however, the classes are not pre-defined but they are defined by data. It is also called as unsupervised learning.
6. **Association Rules**: It tries to find out relationships between data. It is also called as link analysis or affinity analysis.
7. **Segmentation**: It partitions into disjoint groups. It is a special type of clustering.

- Note: There are many data mining tasks however, few of them are mentioned above.

Steps involved in finding out useful information

1. **Business problems(problem statement)**: for example, to predict whether a credit card transaction is fraudulent or not.
2. **Data Collection**: Following ways can be used to collect data
    a. Primary research: Check within the organization(past project documents)
    b. Secondary research: No documents available(using internet, using similar king of case studies, white papers, etc.).
    c. Survey or experimentation: Getting the data by preparing the questionnaire.
3. **Understand the Datatypes**: The data is broadly classified into
    a. **Continuous**: This data is represented using decimal point which makes sense.
        i. **Interval Scale Data**: Interval scales frequently record continuous data but not always i.e. Credit and SAT scores are integers. On these scales the order of values and the interval or distance between any two points is meaningful. e.g. The twenty degree difference between 10 and 30 Celsius is equivalent to the difference between 50 and 70 degrees, however, these variables don't have zero measurement that indicates the lace of the characteristics. e.g. 0 Celsius represents a temperature rather than a lack of the temperature.
        ii. **Ratio Scale Data**: This typically uses continuous data for this level of measurement interval is still meaningful, however, additionally this scale has zero measurement representing a lack of attribute. e.g. 0 kgs indicates a lack of weight. You can add, subtract, multiply and divide values on a ratio scale, however, you can only add and subtract values on an interval scale.
    b. **Discrete**: It cannot be represented in decimal point. This is also categorized as:
        i. **Nominal**: We use different categories of data(only names) where order doesn't matter.
        ii. **Ordinal**: We use categorical data, here order matters.

▼ Q. Identify datatypes for the following(Continuous or Discrete):

Particulars	Datatypes
Number of beating from wife	Discrete

Particulars	Datatypes
Results of a rolling dice	Discrete
Weight of a person	Continuous
Weight of gold	Continuous
Distance between two places	Continuous
Length of leaf	Continuous
Number of kids	Discrete
Number of tickets in Indian railways	Discrete
Highschool class ranking	Discrete
Gender	Discrete

▼ Q. Identify datatypes for the following data(Nominal or Ordinal):

Particulars	Datatypes
Hair color	Nominal
Socioeconomic status	Ordinal
Type of living accommodation	Ordinal
Level of agreement	Ordinal
Religious preference	Nominal
Gender	Nominal

▼ Q. Identify following type of data(Interval or Ratio):

Particulars	Datatypes
Celsius temperature	Interval
Weight	Ratio
Fahrenheit temperature	Interval
Sales figure	Interval
Time of the day	Interval
Barometer pressure	Interval
Examination scores	Ratio
Height	Ratio
IQ	Interval

Implementation of algorithm

- Implementation of algorithm i.e. preparing model for the data i.e. selecting appropriate algorithm based on the datatypes and implementing it.
- Displaying the result s or the o/p using visualization.

Getting to Know Your Data

- To know your data you should know the answers for the following questions:

1. What are the types of the attributes or fields that make up your data?
2. What kind of values does each attribute have?
3. What are the types of attribute values(Continuous or Discrete)?
4. What does the data look like?
5. How are the values distributed?
6. Are their better ways we can visualize the data.

- Measure of central tendency is used to find out where our center of data is lying.
- **Mean**: The mean can be influenced by the extreme values known as outliers.
- **Median**: It is not influenced by outliers. It is used to find the middle value of your dataset.
- **Mode**: It is used to find the frequently occurring value in your dataset. you can't use mode for continuous data. Mode can be unimodal, bimodal and multimodal.
- Sometimes each value xi in a set maybe associated with a weight wi, for i = 1 to n. The weights reflect the importance or occurrence(frequency) attached to the respective values, this is called as **Weighted Mean**.
- Trimmed Mean removes a proportion of the largest and smallest observations and then takes the average of the numbers that remain in the data set

---

Data Pre-processing

- Why do we need to preprocess the data?
    - Real world data is generally incomplete noisy and inconsistence.
    - Incomplete data can occur for a number of reasons:
        1. Attribute of interest may not always be available. e.g. Customer information for sales transaction data.

2. Other data may not be included simply because it was not considered important at a time of entry.
3. Relevant data may not be recorded due to misunderstanding or failure of an equipment.
4. Data that was inconsistent with other recorded data may have been deleted.
5. Missing data particularly for tuples with missing values for some attribute may be needed to be inferred.

- Noisy data:
  1. The data collection can be faulty.
  2. There may have been human error or computer errors while doing data entry.
  3. Inconsistencies in naming conventions or data codes used.

Steps for Data Preprocessing

1. **Data Cleaning**: Data cleaning routine attends to fill in missing values smooth out noise by identifying outliers and correcting inconsistent data.
   a. **Dealing with missing values** i.e. how to fill in missing values for certain attributes.
      i. **Ignore the tuple**: This is usually done when the class label is missing. This method is not very effective unless the tuple contains several attributes with missing values.
      ii. **Fill in the missing value manually**: It is time consuming and may not be feasible given a large dataset with many missing values.
      iii. **Use a global constant to fill in the missing value**: Replace all the missing attribute values by the same constant.
      iv. **Use the attribute mean to fill in the missing value**:
      v. **Use the most probable value to fill in the missing value**: Maybe determined with regression, inference based tools using Bayesian classification, etc.
   b. **Dealing with noisy data**: Noise is a random error or variance in a measured variable. To remove noise we use:

i. **Smoothing techniques(Binning)**: Binning methods smooth a sorted data values by considering values around it. The sorted values are distributed into a number of buckets or bins.

1. **Smoothing by bin means**: Here each value in a bin is replaced by the mean value of the bin.

   ▼ e.g. Consider sorted data for price in dollars: 4, 8, 15, 21, 21, 24, 25, 28, 34 (sort the data if it isn't). Partitioned the given data into equal frequency bins.

   Bin 1 : 4 8 15

   Bin 2 : 21 21 24

   Bin 3 : 25 28 34

   Bin 1 : 9 9 9

   Bin 2 : 22 22 22

   Bin 3 : 29 29 29

2. **Smoothing by bin medians**: Each bin value is replaced by bin medians.

   ▼ e.g. Consider sorted data for price in dollars: 4, 8, 15, 21, 21, 24, 25, 28, 34 (sort the data if it isn't). Partitioned the given data into equal frequency bins.

   Bin 1 : 4 8 15

   Bin 2 : 21 21 24

   Bin 3 : 25 28 34

   Bin 1 : 8 8 8

   Bin 2 : 21 21 21

   Bin 3 : 28 28 28

3. **Smoothing by bin values(boundaries)**: The minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is than replaced by the closest boundary value. NOTE: The larger the width the greater is the smoothing effect.

   ▼ e.g. Consider sorted data for price in dollars: 4, 8, 15, 21, 21, 24, 25, 28, 34 (sort the data if it isn't). Partitioned the given data into equal frequency bins.

Bin 1 : 4 8 15

Bin 2 : 21 21 24

Bin 3 : 25 28 34

Bin 1 : 4 4 15

Bin 2 : 21 21 24

Bin 3 : 25 25 34

ii. **Regression**: Data can be smoothed by fitting it to a function with regression.

iii. **Clustering**: Outliers maybe determined by using clustering where similar values are organized into groups or clusters.

2. **Data Integration**: In data integration the data is combined from multiple sources into a coherent data store as in data warehouse, these sources may include multiple databases, data-cubes, flat files.

   Issues during data integration:

   1. **Entity identification problem**
   2. **Redundancy**
   3. **Detection and resolution of data value conflicts**

3. **Data Reduction**:

   a. **Strategies for data reduction:**

      i. **Data cube aggregation**

      ii. **Attribute subset selection**

      iii. **Dimension reduction**

      iv. **Data compression**

      v. **Numerosity reduction**

      vi. **Data discretization**

4. **Data Transformation**: Data is transformed or consolidated into forms appropriate for mining.

   a. Following strategies are used for data transformation:

i. **Smoothing**: It works to remove noise from the data. Techniques include Binning, Regression and Clustering.

ii. **Attribute Construction(Feature Construction)**: New attributes are constructed to help the mining process.

iii. **Aggregation**: Summary or aggregation operations are applied to the data.

iv. **Normalization**: Attribute data are scaled so as to fall within a smaller range such as -1.0 to 1.0 or 0.0 to 1.0

v. **Discretization**: The raw values of the numeric attribute ex. age are replaced by interval labels. ex. 0 to 10, 11 to 20, etc. or conceptual labels ex. Youth, Adult, Senior, etc. The labels in turn recursively organized into higher level concepts resulting in a concept hierarchy for the numeric attribute.

vi. **Concept hierarchy generation for nominal data**: Attribute such as street can be generalized to higher level concepts like city or country.

Association Rule Mining

- Association Rule: If I purchase item A how likely & frequently I will purchase item B
- **Itemset**: A set of items is referred to as an item set.
- **K-itemset**: An itemset that consists of k number of items, where k > 0, is called as k-itemset. For example, in the set {computer, anti-virus software} as k = 2 it is called as 2-itemset.
- **Support Count or Count or Frequency of itemset**: The occurrence frequency of an itemset is the number of transactions that contains the itemset.
- **Frequent itemset**: If the frequency of an itemset is greater than or equal to the minimum support threshold then the item or itemset is called as frequent.
- **Support**: Frequency of items bought over all transactions. $\frac{A>B}{Support} * p(A \cup B)$
- **Association Rule**: something that I missed. Associations rule mining is useful in:
  - Marketing and Sales promotion
  - Supermarket self management

- Analyzing datasets
- Cross selling, etc.

Apriori Algorithm

- **Apriori property**: All non-empty subsets of a frequent itemset must also be frequent.
- **Confidence**: How likely item Y or B is purchased when item X or A is purchased which is expressed as {X ⇒ Y} or {A ⇒ B}. This is measured by the proportion of transactions with item X in which item Y also appears.
- Generating Association Rules from the generated frequent itemsets (it is the second part of Apriori algorithm)
  - Strong association rules need to be generated from frequent itemsets. The rules are said to be strong if they satisfy both minimum support and minimum confidence. The confidence is given as: $Confidence(A \Rightarrow B) = \frac{SupportCount(A \cup B)}{SupportCount(A)}$
  - The conditional probability is expressed in terms of itemset support count where Support Count of AUB is number of transactions containing the itemsets AUB and Support Count of A is the number of transactions containing the itemset A.

Classifications

- Classification is a technique where we categorize the data into a given number of classes. The main goal of classification is to identify the category or class to which a new data will fall under. Given a database $D = \{t_1, t_2, ..., t_n\}$ and a set of classes $C = c_1, c_2, ..., c_n$, the classification problem is to define a mapping $f : D \rightarrow C$ where, each t_i is assigned to one of the classes.

Construction of Decision Tree

- A decision tree is a popular classification method that results in a flowchart like tree structure, where each node denotes a test on an attribute and each branch represents an outcome of the test. The tree leaves represents the classes.
- Decision tree is a model that is both predictive and descriptive. It is used to display the relationships found in the training data. The training process that generates a tree is called as induction.

Induction Tree Algorithm

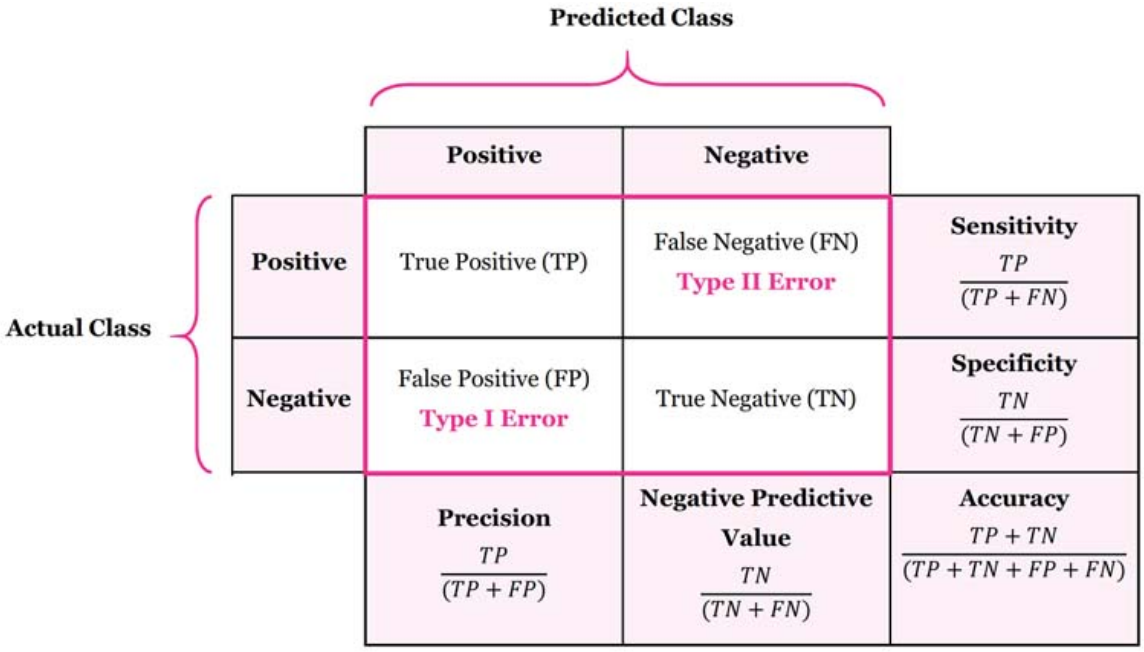
1. Let the set of training data be S, if some of the attributes are continuous value they should be discretized. e.g. Age values under 18, 18 to 40, 41 to 65 can be written as youth, adult, senior citizen, etc.
2. If all the instances S are in the same class than stop.
3. Split the next node by selecting an attribute A from the independent attributes that best divides or splits the objects in the nodes into subsets and create a decision tree node.
4. Split the node according to the values of A.
5. Stop if either of the following conditions met otherwise continue with step 3:
    a. If this partition divides the data into the subsets that belong to a single class and no other node needs splitting.
    b. If there are no remaining attributes on which the sample maybe further divided.
6. To find the split attributes, all the attributes that have not yet been used needs to be given a goodness value that shows the description power of the attribute. The attributes than maybe ranked according to the and the highest ram attribute selected.
7. Following evaluation are used to split the attributes:
    a. Rules based on information theory (information gain).
    b. Rules based on Gini index.

Confusion Matrix

- Following terms are used when we deal with performance measures.
    1. **True Positive(TP)**: These refers to the positive tuples that were correctly labeled by the classifier.
        a. Let TP be the number of True Positives. Interpretation: You predicted positive and its true.
    2. **True Negative(TN)**: These are negative tuples that were correctly labeled by the classifier.
        a. Let TN be the number of True Negatives. Interpretation: You predicted negative and its true.

3. **False Positives(FP)**: These are negative tuples that were incorrectly labeled as positive. e.g. Tuples of class buys_computer = no for which the classifier predicted buys_computer = yes.
    a. Let FP be the number of False Positives. Interpretation: You predicted positive and it is false.
    b. It is also called as Type I Error.
4. **False Negative(FN)**: These are the positive tuples that were mislabeled as negatives. e.g. Tuples of class buys_computer = yes for which the classifier predicted buys_computer = no.
    a. Let FN be the number of false negatives. Interpretation: You predicted negative and it is false.
    b. It is also called as Type II Error.

- **Confusion Matrix** is a performance measurement for classification problem, where the output can be two or more classes. It is a table with four different combinations of actual and predicted values.
- Give n classes, the confusion matrix is a table of size m x n. An entry $CM_{i,j}$ in the first m rows and n columns indicates the number of tuples of class i that were labeled by the classifier as class j.



- **Accuracy** of an classifier on a given test set is the percentage of test set tuples that have correctly classified by the classifier.

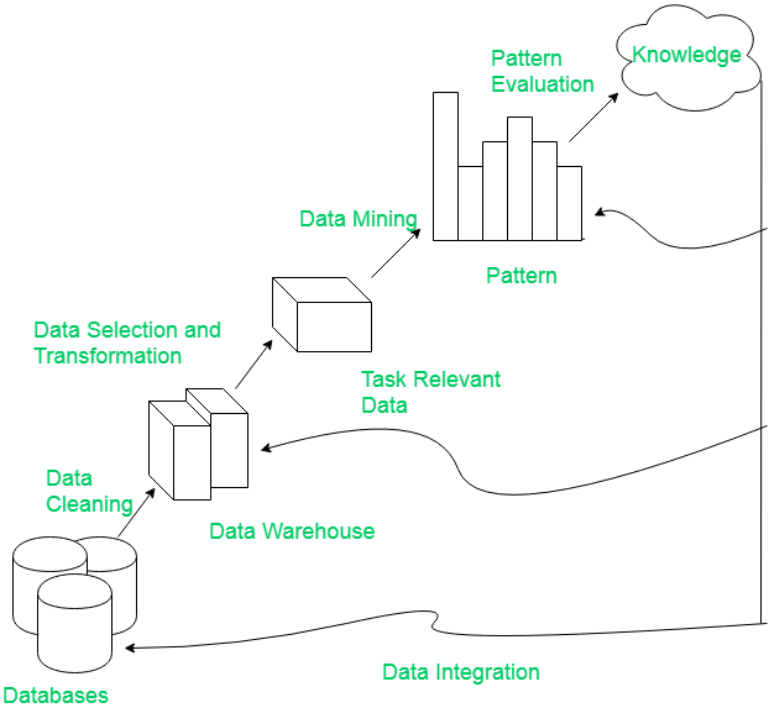
    - In a pattern recognition terminology the accuracy is also refers to as the overall recognition rate of the classifier(how well the classifier recognizes tuples of various classes).
- **Error Rate or Misclassification build** of a classifier is equal to $1 - A(M)$, $A(M)$ is the accuracy of the matrix, which is equal to $\frac{FP+FN}{P+N}$.
    - If we were to use training set instead of the testing set to estimate the error rate of the model than this is called as re-substitution error.
- **Precision** is the measure of exactness(i.e. what percentage of tuples labeled as positives are actually positive)(predicted yes).
- **Recall** is a measure of completeness(what percentage of positive tuples are labeled as such)(actual yes). It's formula is same as Sensitivity.
- **Sensitivity** is also refer to as true positive(recognition) rate. The proportion of positive tuples that are correctly identified. Formula give above.
- **Specificity** is the measure for true negative rate(proportion of negative tuples that have correctly identified). Formula given above.

Questions

▼ Q1) Define data mining and explain with an example ?

1. Data Mining is the process of extracting meaningful and valuable information to identify patterns, trends that allow businesses to improve their performances. It is the process of extracting information from huge sets of data.
2. Data miners can then use those findings to make decisions or predict an outcome. Data Mining is a key part of the data analytics.
3. Eg:-
    a. It is used to explore increasingly large databases and to improve market segmentation.
    b. Data mining in marketing also predicts which users are likely to unsubscribe from a service.
    c. Banking sectors.

▼ Q2) Draw a knowledge discovery diagram and explain the steps in it ?



1)Database:- It consists of unstructured data

2)Data Cleaning:- It is used to remove noise and inconsistent data from collection.

3)Data Integration:- Useful data can be combined from multiple data sources

4)Data Selection:- Data relevant to the analysis are retrieved from the data source

5)Data Warehouse:- After cleaning and integration data is sent to the Data Warehouse. It is a technology that consists of structured data and it sends this data for data mining.

6)Data Transformation:- Data is transformed to an appropriate form required for mining procedure

7)Data Mining:- It is the process of extracting information from a large set of data.

8)Evaluation and Presentation:- It is used to generate frequent patterns based on the data.

9)Knowledge:- The knowledge of the code is presented using visualization tools and understood by the data miner team.

▼ Q3) Explain what types of data , data mining can be performed ?

Data Mining is the process of extracting meaningful and valuable information to identify patterns, trends that allow businesses to improve their performances. It is the process of extracting information from huge sets of data

Types of data that can be mined are as follows:-

1)Relational Databases:- A relational database is a set of tables , each of which Is authorized with unique name. Each table includes a set of attributes (columns) and tuples (rows)

2)Data Warehouses:- It is a technology which converts unstructured data to structured data. It consists of useful data combined from different data sources. A data warehouse is a repository for long-term storage of data from multiple sources, organized so as to facilitate management decision making.

3)Flat files:- These are simple data files in text or binary format with the structure known by the data mining algorithm to be applied

4)Transaction Database:- A transaction database is a set of records representing transactions each with a time stamp.

5)World Wide Web(WWW):- A very large number of authors continuously publishes and updates the data on WWW.

▼ Q4) Applications of Data Mining ?

1)Business Transactions:- Based on business transactions data mining identifies the frequent patterns and can help businesses improve their businesses

2)Scientific analysis:- Big data is generated daily from scientific laboratories which can be easily processed by applying data mining algorithm

3)Market Basket analysis:- It is a data mining technique used by retailers to increase businesses by better understanding customer purchasing patterns

4)Insurance and Healthcare:- It can help users to detect fraud and abuse, insurance agreements. Helps pharmacy to improve drugs based on the data analyzed

5)Education Sector:- It helps to analyze large data collected from users learning patterns

▼ Q5) What kind of patterns can be mined?

The patterns that can be mined are:-

1)Frequent Item sets:- An item set that appears together or frequently in the dataset is called a Frequent Item set. E.g. Bread and Butter are bought together.

2)Frequent subsequences(Sequential Patterns):- The patterns that occur sequentially in dataset are called as Frequent subsequences(Sequential Patterns). E.g. A customer first purchase a laptop, followed by a digital camera, then a memory card.

3)Frequent Structural Patterns:- It refers to different structures such as graphs, trees or lattices, etc. If a structure occurs frequently then it is called Frequent Structural Patterns.

▼ Q6) How can data mining be used for the benefit of society ?

1)Customer – relationship management:- By using this data mining technique companies provide customized and preferred services for customers, which provides a pleasant experience while using the service

2)Personal Search Engine:- With the help of data mining algorithms frequent spam accounts can be identified and they are automatically shifted to spam

3)Mining in health sector:- It can help users to detect fraud and abuse, insurance agreements Helps pharmacy to improve drugs based on the data analyzed

4)E – Shopping:- It helps in announcing offers and discounts to keep the customers active and increase sales.

5)Data Privacy:- Helps to keep user information safe from third party

▼ Q7) Limitations in Data Mining ?

1)Data mining tools are complex and requires training to use

2)The results are not 100 % accurate

3)Privacy issues

4)Lack of standards

5)It requires large databases

6)It is expensive

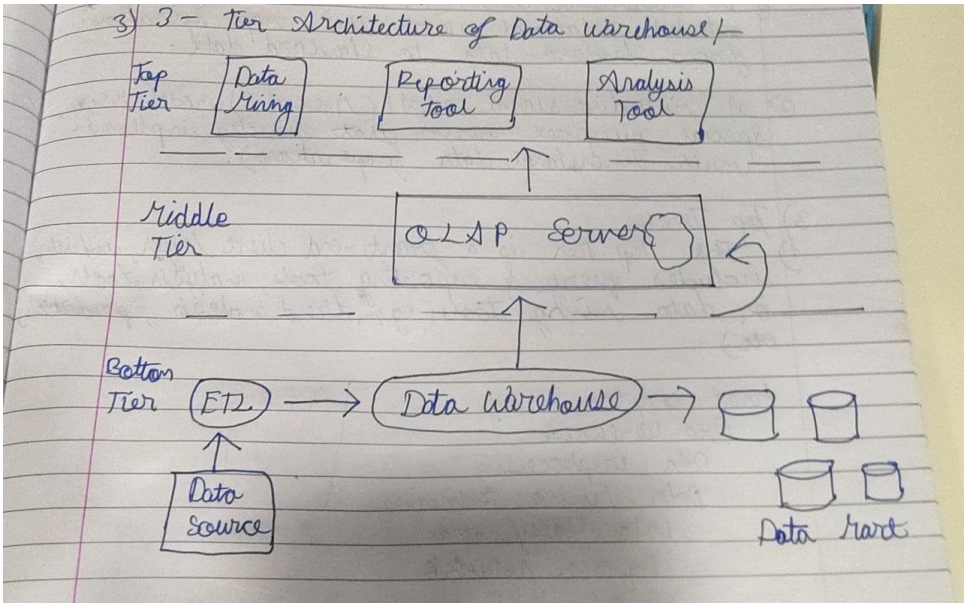
▼ Q8) What Is an Attribute?

An attribute is a data field, representing a characteristic or feature of a data object.

▼ Q9) Explain Data Warehouse with a diagram ?

A data warehouse is considered as a repository to store data for longer duration. A data warehouse is where data is collected for data mining process. It is a technology which converts unstructured data to structured data. E.g. A data

warehouse may contain customer information from an organization containing mailing lists, comment cards, sales records ,etc.



1. Bottom Tier:
    a. It acts a data storage unit to store data collected from different data resources
    b. It is a warehouse database server. Eg:- RDBMS
    c. In this tier data is extracted from operational and external resources using an application program (called gateways).
    d. ETL which stands for extract, transform, and load is a data integration process to combine data collected from different data resources into a single data store.
2. Middle Tier:
    a. It is an OLAP(Online analytical processing)server that is typically implemented using either:
        i. A relational OLAP (ROLAP) i.e an extended RDBMS that maps operation from standard data to standard data.
        ii. A multidimensional OLAP (MOLAP) i.e a special purpose server that directly implements multi-dimensional data and operations
3. Top Tier:
    a. It is a front-end client layer

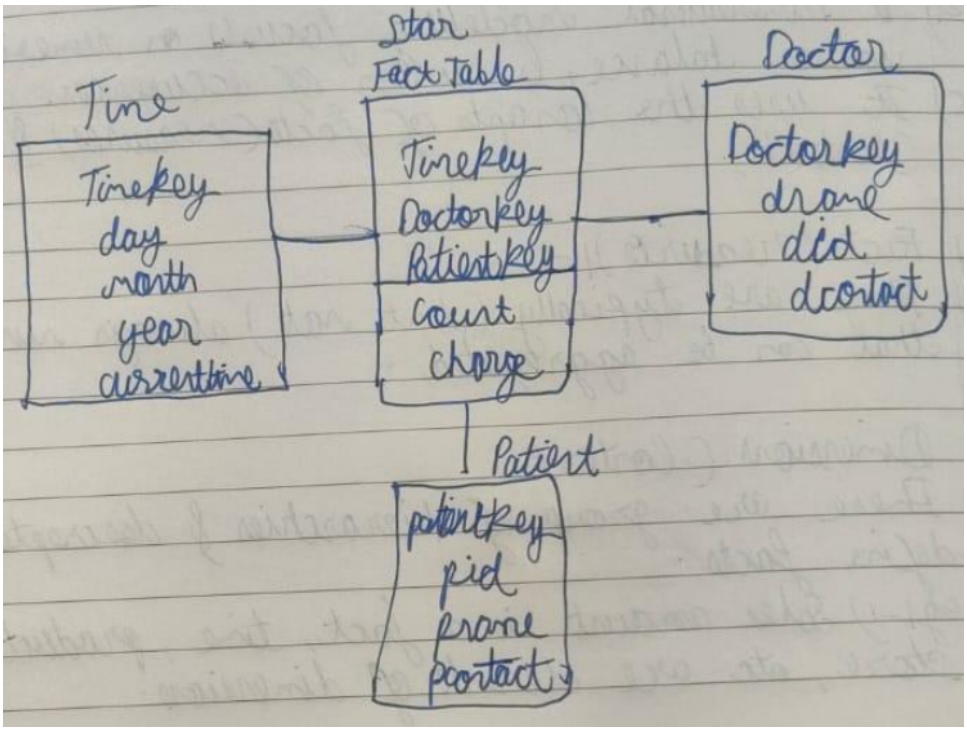
b. It consists of:

i. Data Mining:- It is a process of extracting meaningful and valuable information to identify patterns and trends from large dataset to help businesses improve their performance.

ii. Data Mining tools and Analysis tools

iii. E.g. trend analysis, prediction, etc.

▼ Q10) Suppose a Data warehouse consists of three dimensions : time, doctor, patient and there are 2 measures that are count and charge where charge is the fee that a doctor takes. Draw star schema to represent fact and dimension tables ?



▼ Q11) Data Cubes ?

1. It stores multidimensional aggregated info.
2. Each cell holds an aggregate data value corresponding to a data point in multi-dimensional space.
3. Data cubes provide fast access to pre-computed summarized data.
4. Thereby, benefiting online analytical processing as well as dimensional modeling.

▼ Q12) Multi-Dimensional Data Modelling Schemas ?

1) Star Schema:- It is the simplest data warehouse schema , it is called a star schema because the diagram resembles a star, consisting of a fact table at the center and the points of the star as dimension tables. The Fact table in the star schema is in 3NF form, whereas dimensional tables are denormalized.

2) SnowFlake Schema:- It is the more complex variation of Star schema as its dimension tables are normalized. Only dimension tables can be split. The redundancy and space is reduced because of the normalized table.

3) Fact constellation schema:- It consists of multiple fact tables that share many dimension tables, it is also called the galaxy Schema. The disadvantage of the Fact constellation schema is a more complicated design because many variants for particular kinds of aggregation must be considered and selected. Also the dimension tables are large.

▼ Q13) What is Clustering ?

The process of dividing the dataset into a meaningful group is called Clustering. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

Types of cluster analysis methods:-

1) Partitioning methods:- These methods partition the objects into k clusters and each partition forms one cluster. E.g. K means clustering

2) Hierarchical based methods:- The clusters formed in this method forms a tree type structure based on the hierarchy.

- These methods either start with 1 cluster and then split into smaller and smaller clusters (top down approach) or start with each object in an individual cluster and then try to merge similar clusters into larger and larger clusters(agglomerative or bottom up approach).

3) Density based methods:- These methods consider the clusters as the dense region having some similarity and different from the lower dense region of the space. These methods have good accuracy and ability to merge to clusters.

- Grid based methods:-
    1. In these method the data space is formulated into a finite number of cells that form a grid like structure.
    2. Eg:- STING (Statistical Information Grid).