Internal Assessment Test 2 – Sep 2022

Sub:	Data Mining with Business Intelligence							Sub Code:	20MCA252
Date:	01/09/22	Duration:	90 min's	Max Marks:	50	Sem:	II	Branch:	MCA

Question-1

What is KDD? Explain KDD process in detail. OR

KDD (Knowledge Discovery from Data) Process

- KDD stands for knowledge discoveries from database. There are some pre-processing operations which are required to make pure data in data warehouse before use that data for Data Mining processes.
- A view data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery as a process is depicted in Figure 2 and consists of an iterative sequence of the following steps:
  - ✓ **Data cleaning:** To remove noise and inconsistent data.
  - ✓ **Data integration:** where multiple data sources may be combined.
  - ✓ **Data selection:** where data relevant to the analysis task are retrieved from the database.
  - ✓ **Data transformation**: where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance.
  - ✓ **Data mining**: An essential process where intelligent methods are applied in order to extract data patterns.
  - ✓ **Pattern evaluation**: To identify the truly interesting patterns representing knowledge based on some interestingness measures.
  - ✓ **Knowledge presentation**: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

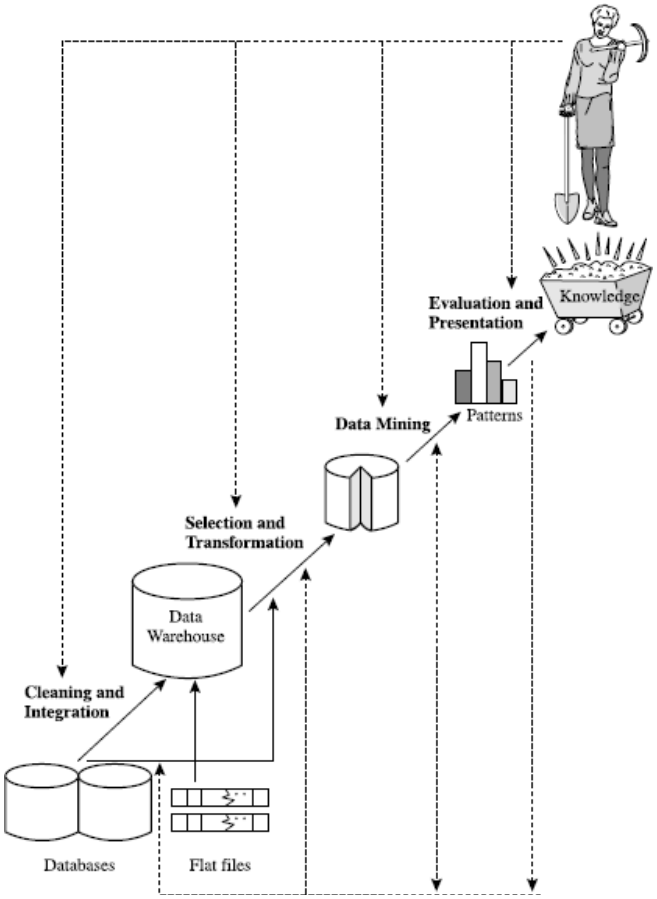


Fig. 2 Data mining as a step in the process of knowledge discovery

- KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.
- Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.
- Objective of Pre-processing on data is to remove noise from data or to remove redundant data.
- There are mainly 4 types of Pre-processing Activities included in KDD Process that is shown in fig. as Data cleaning, Data integration, Data transformation, Data reduction.

Question-2
What are the various features used to classify Data mining? Explain.

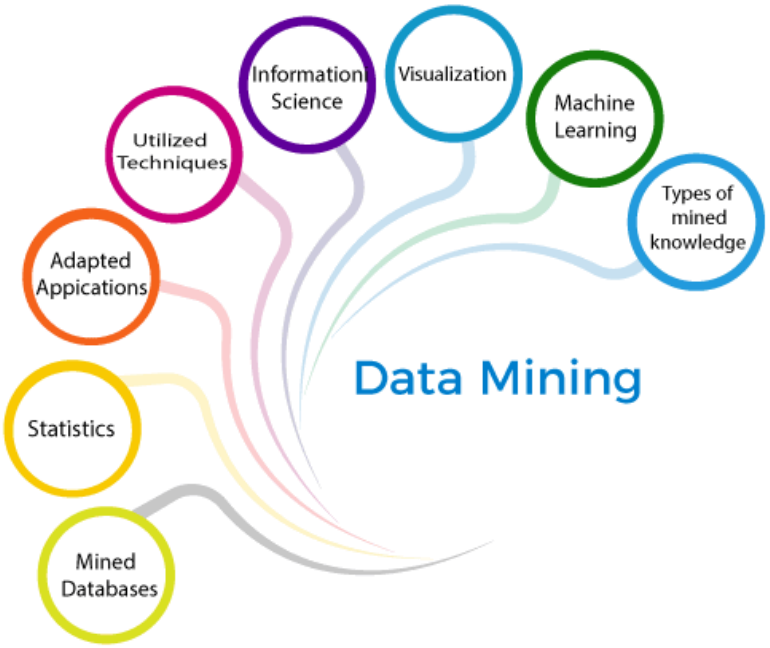
Classification of Data Mining Systems

Data mining refers to the process of extracting important data from raw data. It analyses the data patterns in huge sets of data with the help of several software. Ever since the development of data mining, it is being incorporated by researchers in the research and development field.

With Data mining, businesses are found to gain more profit. It has not only helped in understanding customer demand but also in developing effective strategies to enforce overall business turnover. It has helped in determining business objectives for making clear decisions.

Data collection and data warehousing, and computer processing are some of the strongest pillars of data mining. Data mining utilizes the concept of mathematical algorithms to segment the data and assess the possibility of occurrence of future events.

To understand the system and meet the desired requirements, data mining can be classified into the following systems:



- Classification based on the mined Databases
- Classification based on the type of mined knowledge
- Classification based on statistics
- Classification based on Machine Learning
- Classification based on visualization

- Classification based on Information Science
- Classification based on utilized techniques
- Classification based on adapted applications

Classification Based on the mined Databases

A data mining system can be classified based on the types of databases that have been mined. A database system can be further segmented based on distinct principles, such as data models, types of data, etc., which further assist in classifying a data mining system.

For example, if we want to classify a database based on the data model, we need to select either relational, transactional, object-relational or data warehouse mining systems.

Classification Based on the type of Knowledge Mined

A data mining system categorized based on the kind of knowledge mind may have the following functionalities:

1. Characterization
2. Discrimination
3. Association and Correlation Analysis
4. Classification
5. Prediction
6. Outlier Analysis
7. Evolution Analysis

Classification Based on the Techniques Utilized

A data mining system can also be classified based on the type of techniques that are being incorporated. These techniques can be assessed based on the involvement of user interaction involved or the methods of analysis employed.

Classification Based on the Applications Adapted

Data mining systems classified based on adapted applications adapted are as follows:

1. Finance
2. Telecommunications
3. DNA
4. Stock Markets
5. E-mail

Examples of Classification Task

Following is some of the main examples of classification tasks:

- Classification helps in determining tumor cells as benign or malignant.
- Classification of credit card transactions as fraudulent or legitimate.
- Classification of secondary structures of protein as alpha-helix, beta-sheet, or random coil.
- Classification of news stories into distinct categories such as finance, weather, entertainment, sports, etc.

Question-3
Explain the issues in Data Mining and why Data mining is difficult to implement. OR

- Data Mining is a dynamic and fast-expanding field with great strengths. Major issues in data mining research, partitioning them into five groups: Mining methodology, User interaction, Efficiency and scalability, Diversity of data types, and Data mining & Society.
- Many of these issues have been addressed in recent data mining research and development to a certain extent and are now considered data mining requirements; others are still at the research stage. The issues continue to stimulate further investigation and improvement in data mining.
- **Mining Methodology**: This involves the investigation of new kinds of knowledge, mining in multidimensional space, integrating methods from other disciplines, and the consideration of semantic ties among data objects.
- In addition, mining methodologies should consider issues such as data uncertainty, noise, and incompleteness.

- **Mining various and new kinds of knowledge:** Data mining covers a wide spectrum of data analysis and knowledge discovery tasks, from data characterization and discrimination to association and correlation analysis, classification, regression, clustering, outlier analysis, sequence analysis, and trend and evolution analysis.
- These tasks may use the same database in different ways and require the development of numerous data mining techniques. Due to the diversity of applications, new mining tasks continue to emerge, making data mining a dynamic and fast-growing field.
- For example, for effective knowledge discovery in information networks, integrated clustering and ranking may lead to the discovery of high-quality clusters and object ranks in large networks.
- **Mining knowledge in multidimensional space:** When searching for knowledge in large data sets, we can explore the data in multidimensional space. That is, we can search for interesting patterns among combinations of dimensions (attributes) at varying levels of abstraction. Such mining is known as (exploratory) multidimensional data mining.
- In many cases, data can be aggregated or viewed as a multidimensional data cube. Mining

  knowledge in cube space can substantially enhance the power and flexibility of data mining.
- **Data mining—an interdisciplinary effort:** The power of data mining can be substantially enhanced by integrating new methods from multiple disciplines. For example, to mine data with natural language text, it makes sense to fuse data mining methods with methods of information retrieval and natural language processing.

- As another example, consider the mining of software bugs in large programs. This form of mining, known as bug mining, benefits from the incorporation of software engineering knowledge into the data mining process.
- **Handling uncertainty, noise, or incompleteness of data:** Data often contain noise, errors, exceptions, or uncertainty, or are incomplete. Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns.
- Data cleaning, data preprocessing, outlier detection and removal, and uncertainty reasoning are examples of techniques that need to be integrated with the data mining process.

- **User Interaction**: The user plays an important role in the data mining process. Interesting areas of research include how to interact with a data mining system, how to incorporate a user's background

  knowledge in mining, and how to visualize and comprehend data mining results.

  - **Interactive mining:** The data mining process should be highly interactive. Thus, it is important to build flexible user interfaces and an exploratory mining environment, facilitating the user's interaction with the system.
  - A user may like to first sample a set of data, explore general characteristics of the data, and estimate potential mining results. Interactive mining should allow users to dynamically change the focus of a search, to refine mining requests based on returned results, and to drill, dice, and pivot through the data and knowledge space interactively, dynamically exploring "cube space" while mining.
  - **Incorporation of background knowledge:** Background knowledge, constraints, rules, and other information regarding the domain under study should be incorporated into the knowledge discovery process. Such knowledge can be used for pattern evaluation as well as to guide the search toward interesting patterns.
  - **Presentation and visualization of data mining results:** How can a data mining system present data mining results, vividly and flexibly, so that the discovered knowledge can be easily understood and directly usable by humans? This is especially crucial if the data mining process is interactive.
  - It requires the system to adopt expressive knowledge representations, user-friendly interfaces, and visualization techniques.

- **Efficiency and Scalability**: Efficiency and scalability are always considered when comparing data mining algorithms. As data amounts continue to multiply, these two factors are especially critical.

- **Efficiency and scalability of data mining algorithms**: Data mining algorithms must be efficient and scalable in order to effectively extract information from huge amounts of data in many data repositories or in dynamic data streams.
- In other words, the running time of a data mining algorithm must be predictable, short, and acceptable by applications. Efficiency, scalability, performance, optimization, and the ability to execute in real time are key criteria that drive the development of many new data mining algorithms.
- **Parallel, distributed, and incremental mining algorithms**: The humongous size of many data sets, the wide distribution of data, and the computational complexity of some data mining methods are factors that motivate the development of parallel and distributed data- intensive mining algorithms. Such algorithms first partition the data into "pieces."
- Each piece is processed, in parallel, by searching for patterns. The parallel processes may interact with one another. The patterns from each partition are eventually merged.

- **Diversity of Database Types**: The wide diversity of database types brings about challenges to data mining. These includes are as below.
  - **Handling complex types of data:** Diverse applications generate a wide spectrum of new data types, from structured data such as relational and data warehouse data to semi- structured and unstructured data; from stable data repositories to dynamic data streams; from simple data objects to temporal data, biological sequences, sensor data, spatial data, hypertext data, multimedia data, software program code, Web data, and social network data.
  - It is unrealistic to expect one data mining system to mine all kinds of data, given the diversity of data types and the different goals of data mining. Domain- or application-dedicated data mining systems are being constructed for in depth mining of specific kinds of data.
  - The construction of effective and efficient data mining tools for diverse applications remains a challenging and active area of research.
  - **Mining dynamic, networked, and global data repositories:** Multiple sources of data are connected by the Internet and various kinds of networks, forming gigantic,

distributed, and heterogeneous global information systems and networks.

- The discovery of knowledge from different sources of structured, semi-structured, or unstructured yet interconnected data with diverse data semantics poses great challenges to data mining.

- **Data Mining and Society**: How does data mining impact society? What steps can data mining take to preserve the privacy of individuals? Do we use data mining in our daily lives without even knowing that we do? These questions raise the following issues:
  - **Social impacts of data mining:** With data mining penetrating our everyday lives, it is important to study the impact of data mining on society. How can we used at a mining technology to benefit society? How can we guard against its misuse?
  - The improper disclosure or use of data and the potential violation of individual privacy and data protection rights are areas of concern that need to be addressed.
  - **Privacy-preserving data mining:** Data mining will help scientific discovery, business management, economy recovery, and security protection (e.g., the real-time discovery of intruders and cyberattacks).
  - However, it poses the risk of disclosing an individual's personal information. Studies on privacy-preserving data publishing and data mining are ongoing. The philosophy is to observe data sensitivity and preserve people's privacy while performing successful data mining.
  - **Invisible data mining:** We cannot expect everyone in society to learn and master data mining techniques. More and more systems should have data mining functions built within so that people can perform data mining or use data mining results simply by mouse clicking, without any knowledge of data mining algorithms.
  - Intelligent search engines and Internet-based stores perform such invisible data mining by incorporating data mining into their components to improve their functionality and performance. This is done often unbeknownst to the user.
  - For example, when purchasing items online, users may be unaware that the store is likely collecting data on the buying patterns of its customers, which may be used to recommend other items for purchase in the future.

Question-4
What are the various Data preprocessing techniques?

- Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources.
- Low-quality data will lead to low-quality mining results. How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results? How can the data be preprocessed so as to improve the efficiency and ease of the mining process?
- Data have quality if they satisfy the requirements of the intended use. There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability.
- **Example**
  - Imagine that you are a manager at **AllElectronics** and have been charged with analyzing the company's data with respect to your branch's sales.
    - You immediately set out to perform this task. You carefully inspect the company's database and data warehouse, identifying and selecting the attributes or dimensions (e.g., item, price, and units sold) to be included in your analysis.
    - Alas! You notice that several of the attributes for various tuples have no recorded value. For your analysis, you would like to include information as to whether each item purchased was advertised as on sale, yet you discover that this information has not been recorded.
    - Furthermore, users of your database system have reported errors, unusual values, and inconsistencies in the data recorded for some transactions.
    - In other words, the data you wish to analyze by data mining techniques are incomplete (lacking attribute values or certain attributes of interest, or containing only aggregate data); inaccurate or noisy (containing errors, or values that deviate from the expected); and inconsistent (e.g., containing discrepancies in the department codes used to categorize items).
  - Above example illustrates three of the elements defining data quality: **accuracy, completeness,** and **consistency**.
- Inaccurate, incomplete, and inconsistent data are commonplace properties of large real-world databases and data warehouses.

- There are many possible reasons for inaccurate data (i.e., having incorrect attribute values). The data collection instruments used may be faulty.

- There may have been human or computer errors occurring at data entry. Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information (e.g., by choosing the default value "January 1" displayed for birthday). This is known as disguised missing data. Errors in data transmission can also occur.
- There may be technology limitations such as limited buffer size for coordinating synchronized data transfer and consumption. Incorrect data may also result from inconsistencies in naming conventions or data codes, or inconsistent formats for input fields (e.g., date).
- Incomplete data can occur for a number of reasons. Attributes of interest may not always be available, such as customer information for sales transaction data.
- Other data may not be included simply because they were not considered important at the time of entry. Relevant data may not be recorded due to a misunderstanding or because of equipment malfunctions. Data that were inconsistent with other recorded data may have been deleted.
- Furthermore, the recording of the data history or modifications may have been overlooked. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.
- **Data Preprocessing Methods/Techniques:**
    - **Data Cleaning** routines work to "clean" the data by filling in missing values, smoothing noisy

      data, identifying or removing outliers, and resolving inconsistencies.
    - **Data Integration** which combines data from multiple sources into a coherent data store, as in data warehousing.
    - **Data Transformation**, the data are transformed or consolidated into forms appropriate for mining
    - **Data Reduction** obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results.

Question-5

What are the various types of attributes? Explain Mean, Median, Mode, Variance and Standard Deviation in brief. OR

- **Mean:** The sample mean is the **average** and is computed as the sum of all the observed outcomes from the sample divided by the total number of events. We use x as the symbol for the sample mean. In math terms,
- $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x$$

  where n is the sample size and the x correspond to the observed valued.
- Let's look to Find out Mean.
- Suppose you randomly sampled six acres in the Desolation Wilderness for a non-indigenous weed and came up with the following counts of this weed in this region: 34, 43, 81, 106, 106 and 115
- We compute the sample mean by adding and dividing by the number of samples,

  $$\frac{34 + 43 + 81 + 106 + 106 + 115}{6}$$
- We can say that the sample mean of non-indigenous weed is 80.83.
- The mode of a set of data is the number with the highest frequency. In the above example 106 is the mode, since it occurs twice and the rest of the outcomes occur only once.
- The population mean is the average of the entire population and is usually impossible to compute. We use the Greek letter μ for the population mean.
- **Median:** One problem with using the mean, is that it often does not depict the typical outcome. If there is one outcome that is very far from the rest of the data, then the mean will be strongly affected by this outcome. Such an outcome is called and **outlier**.
- An alternative measure is the median; the median is the **middle score**. If we have an even number of events, we take the average of the two middles. The median is better for describing the typical value. It is often used for income and home prices.
- Let's Look to Find out Median.
- Suppose you randomly selected **10** house prices in the South Lake area. You are interested in the typical house price. In **$100,000** the prices were: 2.7, 2.9, 3.1, 3.4, 3.7, 4.1, 4.3, 4.7, 4.7,

40.8.

- If we computed the mean, we would say that the average house price is **744,000**. Although this number is true, it does not reflect the price for available housing in South Lake Tahoe.
- A closer look at the data shows that the house valued at **40.8 x $100,000 = $4.08** million skews the data. Instead, we use the median. Since there is an even number of outcomes, we take the average of the middle two is 3.9.

$$\frac{3.7 + 4.1}{2} = 3.9$$

- The median house price is $390,000. This better reflects what house shoppers should expect to spend.
- **Mode:** The mode is another measure of central tendency. The mode for a set of data is the value that occurs most frequently in the set.
- Therefore, it can be determined for qualitative and quantitative attributes. It is possible for the greatest frequency to correspond to several different values, which results in more than one mode. Data sets with one, two, or three modes are respectively called **unimodal, bimodal, and trimodal**.
- In general, a dataset with two or more modes is **multimodal**. At the other extreme, if each data value occurs only once, then there is no mode.
- Let's Look for find Mode.
- In Above Example We Consider 4.7 As Mode.
- **Variance & Standard Deviation:** The mean, mode and median do a nice job in telling where the center of the data set is, but often we are interested in more.
- For example, a pharmaceutical engineer develops a new drug that regulates iron in the blood. Suppose she finds out that the average sugar content after taking the medication is the optimal level. This does not mean that the drug is effective. There is a possibility that half of the patients have dangerously low sugar content while the other half have dangerously high content.
- Instead of the drug being an effective regulator, it is a deadly poison. What the pharmacist needs is a measure of how far the data is spread apart. This is what the variance and standard deviation do. First we show the formulas for these measurements. Then we will go through the steps on how to use the formulas.

- We define the variance to be

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$$

- and the standard deviation to be

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2}$$

Variance and Standard Deviation: Step by Step

- Calculate the mean**, x**.
- Write a table that subtracts the mean from each observed value.
- Square each of the differences.
- Add this column.
- Divide by **n -1** where **n** is the number of items in the sample this is the **variance**.
- To get the **standard deviation** we take the square root of the variance.
- Let's Look to Find out variance & standard deviation

- The owner of the Indian restaurant is interested in how much people spend at the restaurant. He examines **10** randomly selected receipts for parties of four and writes down the following data.
- 44, 50, 38, 96, 42, 47, 40, 39, 46, 50
- He calculated the mean by adding and dividing by 10 to get Average(Mean) = 49.2.
- Below is the table for getting the standard deviation:

x	x - 49.2	(x - 49.2)2
44	-5.2	27.04
50	0.8	0.64
38	11.2	125.44
96	46.8	2190.24
42	-7.2	51.84
47	-2.2	4.84

40	-9.2	84.64
39	-10.2	104.04
46	-3.2	10.24
50	0.8	0.64
Total		2600.4

- Now 2600.4/10 – 1 = 288.7
- Hence the variance is **289** and the standard deviation is the square root of **289 = 17**.
- Since the standard deviation can be thought of measuring how far the data values lie from the mean, we take the mean and move one standard deviation in either direction. The mean for this example was about 49.2 and the standard deviation was 17.
- We have: 49.2 - 17 = 32.2 and 49.2 + 17 = 66.2
- What this means is that most of the patrons probably spend between $32.20 and $66.20.
- The **sample standard deviation** will be denoted by s and the **population standard deviation** will be denoted by the Greek letter σ.
- The sample variance will be denoted by **s2** and the population variance will be denoted by **σ2**.
- The variance and standard deviation describe how spread out the data is. If the data all lies close to the mean, then the standard deviation will be small, while if the data is spread out over a large range of values, s will be large. Having outliers will increase the standard deviation.

Question-6

What is Data Cleaning? Discuss various ways of handling missing values and noisy data during data cleaning.

- Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.
- **Missing Values:** Imagine that you need to analyze AllElectronics sales and customer data. You note that many tuples have no recorded value for several attributes such as customer income. How can you go about filling in the missing values for this attribute? Let's look at the following methods.
    - **Ignore the tuple:** This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

- By ignoring the tuple, we do not make use of the remaining attributes values in the tuple. Such data could have been useful to the task at hand.
- **Fill in the missing value manually:** In general, this approach is time consuming and may not be feasible given a large data set with many missing values.
- **Use a global constant to fill in the missing value:** Replace all missing attribute values by the same constant such as a label like "Unknown" or 1. If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of "Unknown." Hence, although this method is simple, it is not foolproof.
- **Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value:** For normal (symmetric) data distributions, the mean can be used, while skewed data distribution should employ the median.
- For example, suppose that the data distribution regarding the income of AllElectronics customers is symmetric and that the mean income is $56,000. Use this value to replace the missing value for income.
- **Use the attribute mean or median for all samples belonging to the same class as the given tuple:** For example, if classifying customers according to credit risk, we may replace the missing value with the mean income value for customers in the same credit risk category as that of the given tuple. If the data distribution for a given class is skewed, the median value is a better choice.
- **Use the most probable value to fill in the missing value:** This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

- **Noisy Data:** Noise is a random error or variance in a measured variable. Given a numeric attribute such as say, price, how can we "smooth" out the data to remove the noise? Let's look at the following data smoothing techniques.
  - **Binning:** Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it. The sorted values are distributed into a number of "buckets," or bins. Because binning methods consult the neighborhood of values, they perform local smoothing.

- Figure 1 illustrates some binning techniques. In this example, the data for price are first sorted and then partitioned into equal-frequency bins of size 3 (i.e., each bin contains three values).
- In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.
- For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9. Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing. Alternatively, bins may be equal width, where the interval range of values in each bin is constant.
- **Sorted data for price (in dollars):** 4, 8, 15, 21, 21, 24, 25, 28, 34

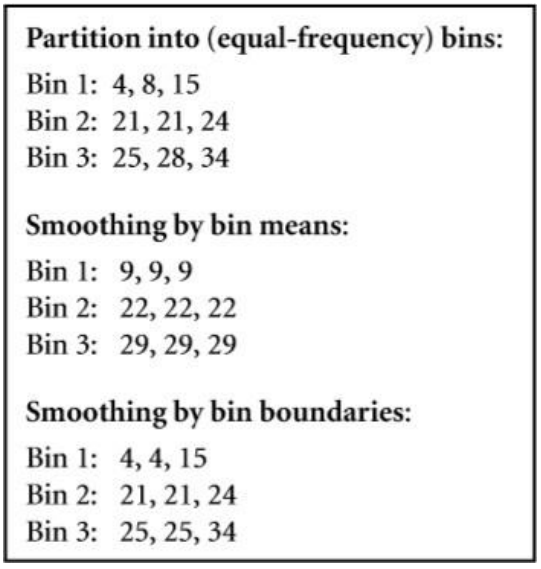


Fig. 1: Binning methods for data smoothing

- **Regression:** Data smoothing can also be done by regression, a technique that conforms data values to a function. Linear regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other.

  Multiple linear regression is an extension of linear regression, where more than two attributes are

  involved and the data are fit to a multidimensional surface.

- **Outlier analysis**: Outliers may be detected by clustering, for example, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers.

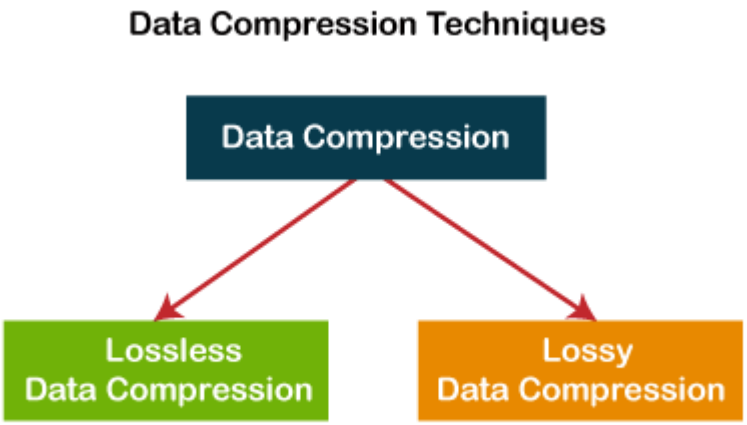
Question-7
What are the various Data Compression techniques? OR

Data Compression is also referred to as **bit-rate reduction** or **source coding**. This technique is used to reduce the size of large files.

The advantage of data compression is that it helps us save our disk space and time in the data transmission.

There are mainly two types of data compression techniques -

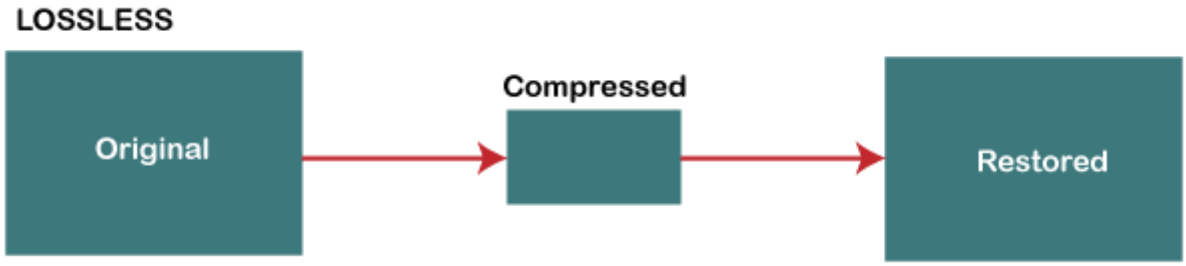
1. Lossless Data Compression
2. Lossy Data Compression



Lossless data compression is used to compress the files **without losing an original file's quality and data**. Simply, we can say that in lossless data compression, file size is reduced, but the quality of data remains the same.

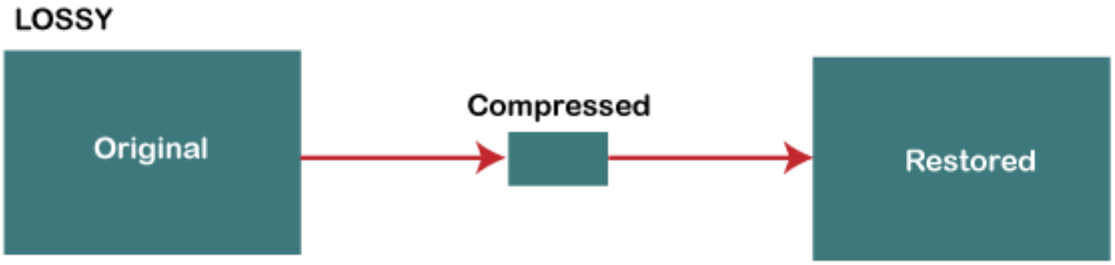
The main advantage of lossless data compression is that we can restore the original data in its original form after the decompression.

Lossless data compression mainly used in the sensitive documents, confidential information, and PNG, RAW, GIF, BMP file formats.



Lossy data compression is used to compress larger files into smaller files. In this compression technique, some specific amount of **data and quality are removed (loss) from the original file**. It takes less memory space from the original file due to the loss of original data and quality. This technique is generally useful for us when the quality of data is not our first priority.

Lossy data compression is most widely used in JPEG images, MPEG video, and MP3 audio formats.



Some important Lossy data compression techniques are -

1. Transform coding
2. Discrete Cosine Transform (DCT)
3. Discrete Wavelet Transform (DWT)

Question-8
Explain the Data Transformation techniques in detail with examples.

- In data transformation, the data are transformed or consolidation to forms appropriate for mining. Strategies for data transformation include the following:
- **Smoothing,** which works to remove noise from the data. Techniques include binning, regression, and clustering.
- **Attribute construction (or feature construction)**, where new attributes are constructed and added from the given set of attributes to help the mining process.
- **Aggregation**, where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This

step is typically used in constructing a data cube for data analysis at multiple abstraction levels.

- **Normalization**, where the 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.

Example: Data Transformation -2, 32, 100, 59, 48⟶

- **Discretization,** where the raw values of a numeric attribute (e.g. Age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior). The labels, in turn, can be recursively organized into higher-level concepts, resulting in a concept hierarchy for the numeric attribute. Figure 2 shows a concept hierarchy for the attribute price. More than one concept hierarchy can be defined for the same attribute to accommodate the needs of various users.

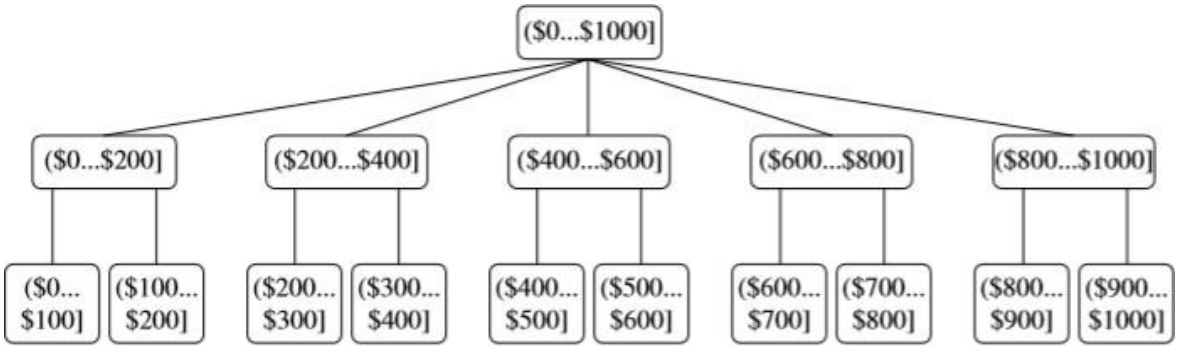


Fig. 2 A concept hierarchy for the attribute price, where an interval ($X... $Y] denotes the range from $X (exclusive) to $Y (inclusive).

- **Concept hierarchy generation for nominal data**, where attributes such as street can be generalized to higher-level concepts, like city or country. Many hierarchies for nominal attributes are implicit within the database schema and can be automatically defined at the schema definition level.

Question-9
For given transaction data, generate frequent item set and identify valid association rules with minimum support as 60% and minimum confidence as 75%

TID	Items
1	Bread, Milk
2	Bread, Chocolate, Pepsi, Eggs
3	Milk, Chocolate, Pepsi, Coke
4	Bread, Milk, Chocolate, Pepsi
5	Bread, Milk, Chocolate, Coke

The support for the following Itemset is 60%
Bread, Milk 3/5
Bread, Diaper 3/5
Milk, Diaper 3/5
Beer, Diaper 3/5

The confidence is 75% for
Bread, Milk
vs bread is 3/4
vs milk is 3/4
Bread, Diaper 3/5
vs bread is 3/4
vs diaper is 3/4
Milk, Diaper 3/5
vs milk is 3/4
vs diaper is 3/4

But not for

Beer, Diaper 3/5
vs diaper is ¾

The valid association rules are
Bread → Milk
Bread → Diaper
Milk→ Diaper

OR
Question-10
Explain Apriori Algorithm with an example and also explain methods to Improve Apriori's Efficiency.

- **Purpose**: The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.
- **Key Concepts**:
  - **Frequent Itemsets**: The sets of item which has minimum support (denoted by L_i for ith-Itemset).

• **Apriori Property**: Any subset of frequent itemset must be frequent.

• **Join Operation**: To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} itself.

- Find the frequent itemsets: the sets of items that have minimum support – A subset of a frequent itemset must also be a frequent itemset **(Apriori Property)**
- i.e., if {AB} is a frequent itemset, both {A} and {B} should be a frequent itemset – Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)
- Use the frequent itemsets to generate association rules.

The Apriori Algorithm : Pseudo code

- **Join Step**: C k is generated by joining Lk-1with itself
- **Prune Step**: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

• Pseudo-code:

C_k: Candidate itemset of size

k L_k: frequent itemset of

size k L_1= {frequent

items};

for (k = 1; L_k != ∅; k++) **do begin**

C_{k+1} = candidates generated from L_k;

for each transaction t in database do

Increment the count of all candidates in

C_{k+1} That are contained in t

L_{k+1} = candidates in C_{k+1} with min_support

end

return ∪k L_k;

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Example

- Consider a database, **D**, consisting of 9 transactions.
- Suppose min. support count required is **2** (i.e. min_sup = 2/9 = 22 %)
- Let minimum confidence required is **70%**.
- We have to first find out the frequent itemset using Apriori algorithm.
- Then, Association rules will be generated using min. support & min. confidence.

Step 1: Generating 1-itemset Frequent Pattern

Scan D for count of each candidate

Itemset	Sup.Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

C_1

Compare candidate support count with minimum support count

Itemset	Sup.Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

L_1

- **The set of frequent 1-itemsets, L1**, consists of the candidate 1- itemsets satisfying minimum support.
- In the first iteration of the algorithm, each item is a member of the set of candidate.

Step 2: Generating 2-itemset Frequent Pattern

Generate C_2 candidates from L_1

Itemset
{I1, I2}
{I1, I3}
{I1, I4}
{I1, I5}
{I2, I3}
{I2, I4}
{I2, I5}
{I3, I4}
{I3, I5}
{I4, I5}

C_2

Scan D for count of each candidate

Itemset	Sup. Count
{I1, I2}	4
{I1, I3}	4
{I1, I4}	1
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2
{I3, I4}	0
{I3, I5}	1
{I4, I5}	0

C_2

Compare candidate support count with minimum support count

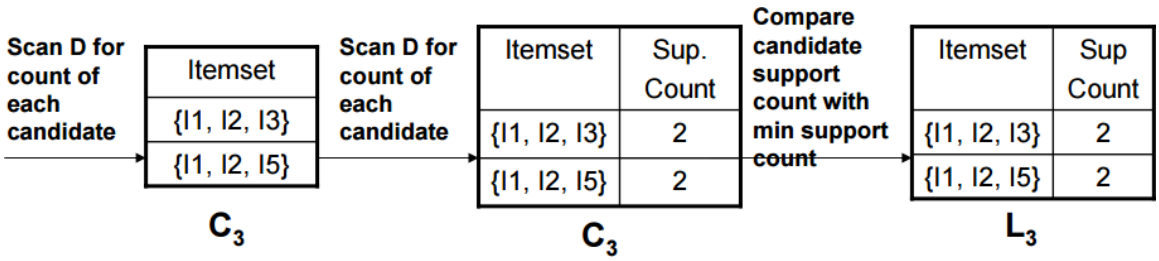
Itemset	Sup Count
{I1, I2}	4
{I1, I3}	4
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

L_2

- To discover the set of frequent 2-itemsets, L_2, the algorithm uses L_1 Join L_1 to generate a candidate set of 2-itemsets, C_2.
- Next, the transactions in D are scanned and the support count for each candidate itemset in C_2 is accumulated (as shown in the middle table).
- The set of frequent 2-itemsets, L_2, is then determined, consisting of those candidate 2-itemsets in C_2 having minimum support.

- **Note**: We haven't used Apriori Property yet.

Step 3: Generating 3-itemset Frequent Pattern



- The generation of the set of candidate 3-itemsets, C_3 , involves use of the Apriori Property.
- In order to find C_3, we compute L_2 Join L_2.

o $C_3 = L_2$ join L_2 = {{I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5}}.

- Now, Join step is complete and Prune step will be used to reduce the size of C_3. Prune step helps to avoid heavy computation due to large C_k.
- Based on the **Apriori property** that all subsets of a frequent itemset must also be frequent, we can determine that four latter candidates cannot possibly be frequent. How ?
- For example, lets take **{I1, I2, I3}**. The 2-item subsets of it are {I1, I2}, {I1, I3} & {I2, I3}. Since all 2- item subsets of {I1, I2, I3} are members of L_2, We will keep {I1, I2, I3} in C_3.
- Lets take another example of **{I2, I3, I5}** which shows how the pruning is performed. The 2-item subsets are {I2, I3}, {I2, I5} & {I3,I5}.
- But, {I3, I5} is not a member of L_2 and hence it is not frequent **violating Apriori Property**. Thus We will have to remove {I2, I3, I5} from C_3.
- Therefore, C_3 = {{I1, I2, I3}, {I1, I2, I5}} after checking for all members of result of Join operation for Pruning.
- Now, the transactions in D are scanned in order to determine **L_3, consisting of those candidates 3- itemsets in C_3 having minimum support.**

Step 4: Generating 4-itemset Frequent Pattern

- The algorithm uses L_3 Join L_3 to generate a candidate set of 4-itemsets, **C_4**. Although the join results in {{I1, I2, I3, I5}}, this itemset is pruned since its subset {{I2, I3, I5}}

is not frequent.

- Thus, **C4 = φ**, and algorithm terminates, **having found all of the frequent items. This completes our Apriori Algorithm.** What's Next?
- These frequent itemsets will be used to generate **strong association rules** (where strong association rules satisfy both minimum support & minimum confidence).

- **Step 5: Generating Association Rules from Frequent Itemsets**

Procedure:

- For each frequent itemset "l", generate all nonempty subsets of l.
- For every nonempty subset s of l, output the rule "s -> (l-s)" if support_count(l) / support_count(s) >= min_conf where min_conf is minimum confidence threshold.

Back to Example:

- We had L = {{I1}, {I2}, {I3}, {I4}, {I5}, {I1, I2}, {I1, I3}, {I1, I5}, {I2, I3}, {I2, I4}, {I2, I5}, {I1, I2, I3}, {I1, I2, I5}}.
- Let's take l = {I1, I2, I5}. – It's all nonempty subsets are {I1, I2}, {I1, I5}, {I2, I5}, {I1}, {I2}, {I5}.
- Let **minimum confidence threshold** is, say 70%.
- The resulting association rules are shown below, each listed with its confidence.
- R1: I1 ^ I2 -> I5 Confidence = sc{I1, I2, I5}/sc{I1,I2} = 2/4 = 50% (R1 is Rejected)
- R2: I1 ^ I5 -> I2 Confidence = sc{I1, I2, I5}/sc{I1,I5} = 2/2 = 100% (**R2 is Selected**)
- R3: I2 ^ I5 -> I1 Confidence = sc{I1, I2, I5}/sc{I2,I5} = 2/2 = 100% (**R3 is Selected**)
- R4: I1 -> I2 ^ I5 Confidence = sc{I1, I2, I5}/sc{I1} = 2/6 = 33% (R4 is Rejected)
- R5: I2 -> I1 ^ I5 Confidence = sc{I1, I2, I5}/{I2} = 2/7 = 29% (R5 is Rejected)
- R6: I5 -> I1 ^ I2 Confidence = sc{I1, I2, I5}/ {I5} = 2/2 = 100% (**R6 is Selected**)
- In this way, we have found **three strong association rules**.