UNIT I

DATA MINING

Introduction – Data – Types of Data – Data Mining Functionalities – Interestingness of Patterns – Classification of Data Mining Systems – Data Mining Task Primitives – Integration of a Data Mining System with a Data Warehouse – Major Issues in Data Mining –Data Preprocessing.

Data

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
- Examples: eye color of a person, temperature, etc.
- Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
- Object is also known as record, point, case, sample, entity, or instanceAttributes

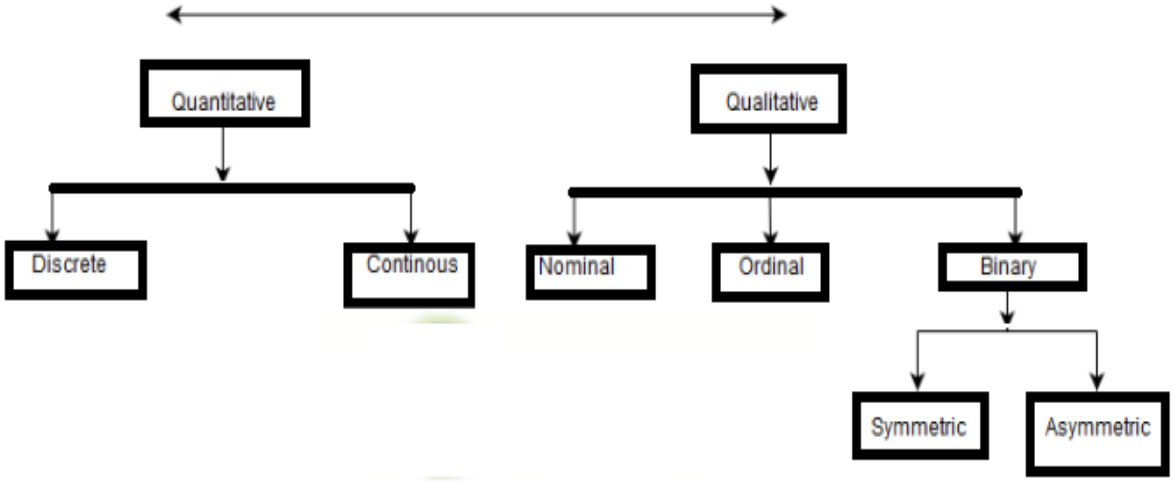
Objects Attribute Values

- Attribute values are numbers or symbols assigned to an attribute
- Distinction between attributes and attribute values
- Same attribute can be mapped to different attribute values
- Example: height can be measured in feet or meters
- Different attributes can be mapped to the same set of values
- Example: Attribute values for ID and age are integers
- But properties of attribute values can be different
- ID has no limit but age has a maximum and minimum value

Types of Attributes

This is the First step of Data-preprocessing. We differentiate between different types of attributes and then preprocess the data. So here is the description of attribute types.

1. Qualitative (Nominal (N), Ordinal (O), Binary(B)).
2. Quantitative (Numeric, Discrete, Continuous)



Qualitative Attributes:

1. Nominal Attributes – related to names: The values of a Nominal attribute are names of things, some kind of symbols. Values of Nominal attributes represents some category or state and that's why nominal attribute also referred as **categorical attributes** and there is no order (rank, position) among values of the nominal attribute. Example :

Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

Binary Attributes: Binary data has only 2 values/states. For Example yes or no, affected or unaffected, true or false

.

- **Symmetric:** Both values are equally important (Gender).

Attribute	Values
Gender	Male , Female

- **Asymmetric:** Both values are not equally important (Result).

Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail

Ordinal Attributes : The Ordinal Attributes contains values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don't indicate how important it is.

Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

Quantitative Attributes:

1. Numeric: A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of 2 types, **interval**, and **ratio**.

- An **interval-scaled** attribute has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point, or we can call zero points. Data can be added and subtracted at an interval scale but can not be multiplied or divided. Consider an example of temperature in degrees Centigrade. If a day's temperature of one day is twice of the other day we cannot say that one day is twice as hot as another day.
- A **ratio-scaled** attribute is a numeric attribute with a fix zero-point. If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value. The values are ordered, and we can also compute the difference between values, and the mean, median, mode, Quantile-range, and Five number summary can be given.

2. Discrete : Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite set of values.

Example:

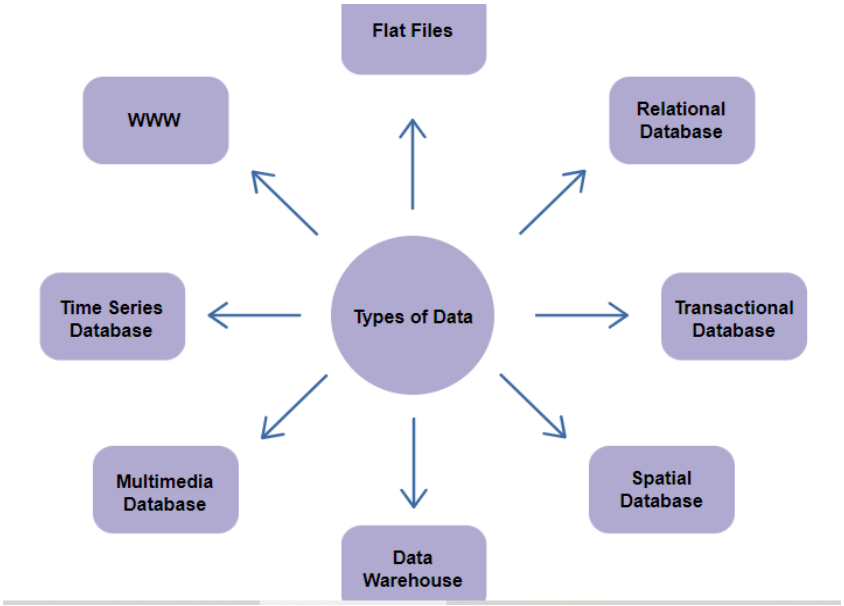
Attribute	Value
Profession	Teacher, Business man, Peon
ZIP Code	301701, 110040

3. Continuous: Continuous data have an infinite no of states. Continuous data is of float type. There can be many values between 2 and 3.

Example :

Attribute	Value
Height	5.4, 6.2 ...etc
weight	50.33 .....etc

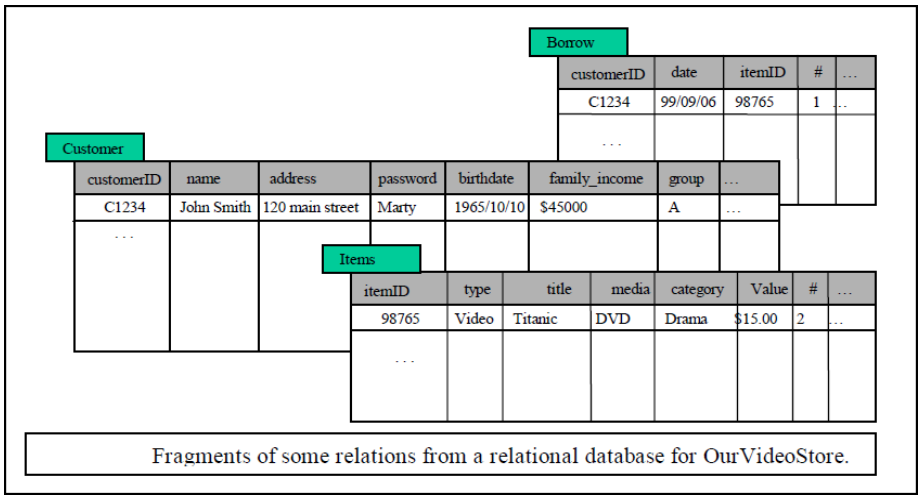
Types of Data



Flat files: Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a

structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific measurements, etc.

Relational Databases: a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key. In following figure it presents some relations Customer, Items, and Borrow representing business activity in a video store. These relations are just a subset of what could be a database for the video store and is given as an example.



Fragments of some relations from a relational database for OurVideoStore.

The most commonly used query language for relational database is SQL, which allows retrieval and manipulation of the data stored in the tables, as well as the calculation of aggregate functions such as average, sum, min, max and count. For instance, an SQL query to select the videos grouped by category would be:

SELECT count(*) FROM Items WHERE type=video GROUP BY category.

Data mining algorithms using relational databases can be more versatile than data mining algorithms specifically written for flat files, since they can take advantage of the structure inherent to relational databases. While data mining can benefit from SQL for data selection, transformation and consolidation, it goes beyond what SQL could provide, such as predicting, comparing, detecting deviations, etc.

Transactional databases

In general, a transactional database consists of a flat file where each record represents a transaction. A transaction typically includes a unique transaction identity number (trans ID), and a list of the items making up the transaction (such as items purchased in a store) as shown below:

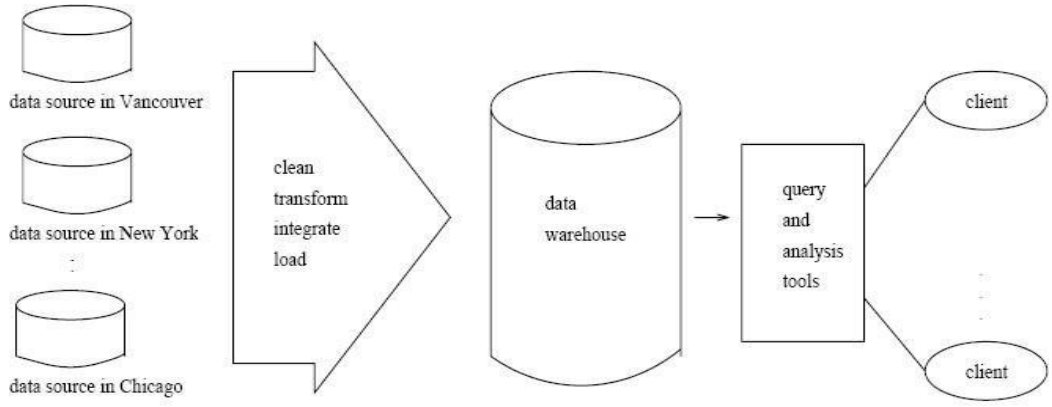
SALES

Trans-ID	List of item_ID's
T100 ……..	I1,I3,I8 ……

A spatial database contains spatial-related data, which may be represented in the form of raster or vector data. Raster data consists of n-dimensional bit maps or pixel maps, and vector data are represented by lines, points, polygons or other kinds of processed primitives, Some examples of spatial databases include geographical (map) databases, VLSI chip designs, and medical and satellite images databases.

Data warehouses

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleansing, data transformation, data integration, data loading, and periodic data refreshing. The figure shows the basic architecture of a data warehouse.

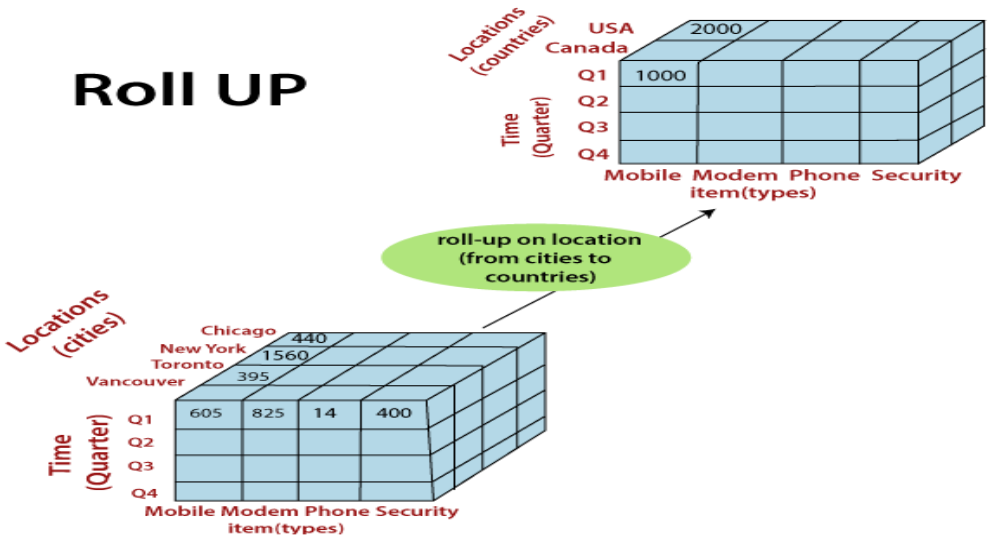


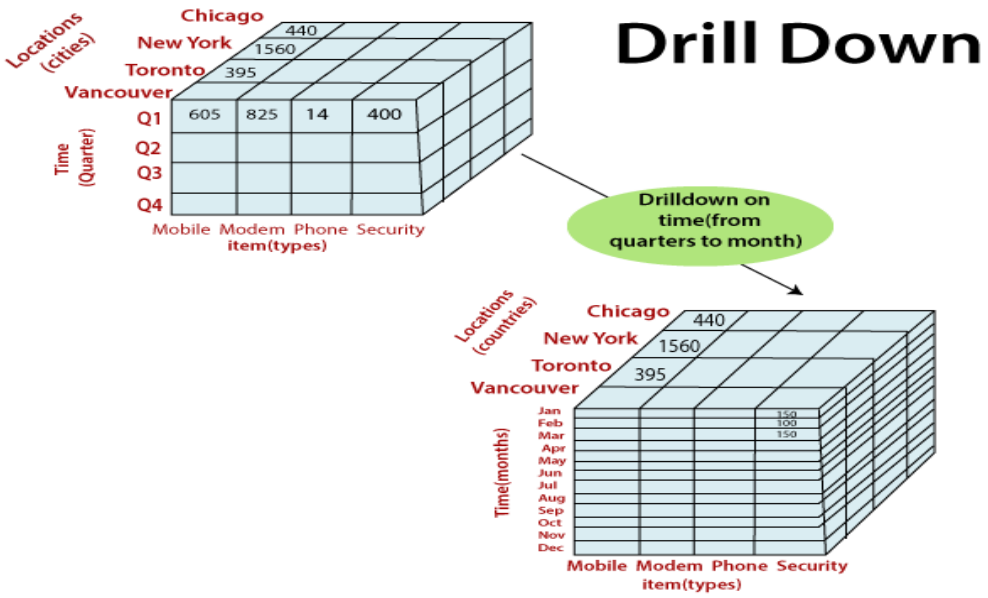
Architecture of a typical data warehouse.

In order to facilitate decision making, the data in a data warehouse are organized around major subjects, such as customer, item, supplier, and activity. The data are stored to provide information from a historical perspective and are typically summarized.

A data warehouse is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or sales amount. The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube. It provides a multidimensional view of data and allows the precomputation and fast accessing of summarized data.

The data cube structure that stores the primitive or lowest level of information is calleda base cuboid. Its corresponding higher level multidimensional (cube) structures are called (non-base) cuboids. A base cuboid together with all of its corresponding higher level cuboids form a data cube. By providing multidimensional data views and the precomputation of summarized data, data warehouse systems are well suited for On- Line Analytical Processing, or OLAP. OLAP operations make use of background knowledge regarding the domain of the data being studied in order to allow the presentation of data at different levels of abstraction. Such operations accommodate different user viewpoints. Examples of OLAP operations include drill-down and roll-up, which allow the user to view the data at differing degrees of summarization, as illustrated in above figure.





A multimedia database stores images, audio, and video data, and is used in applications such as picture content-based retrieval, voice-mail systems, video-on- demand systems, the World Wide Web, and speech-based user interfaces.

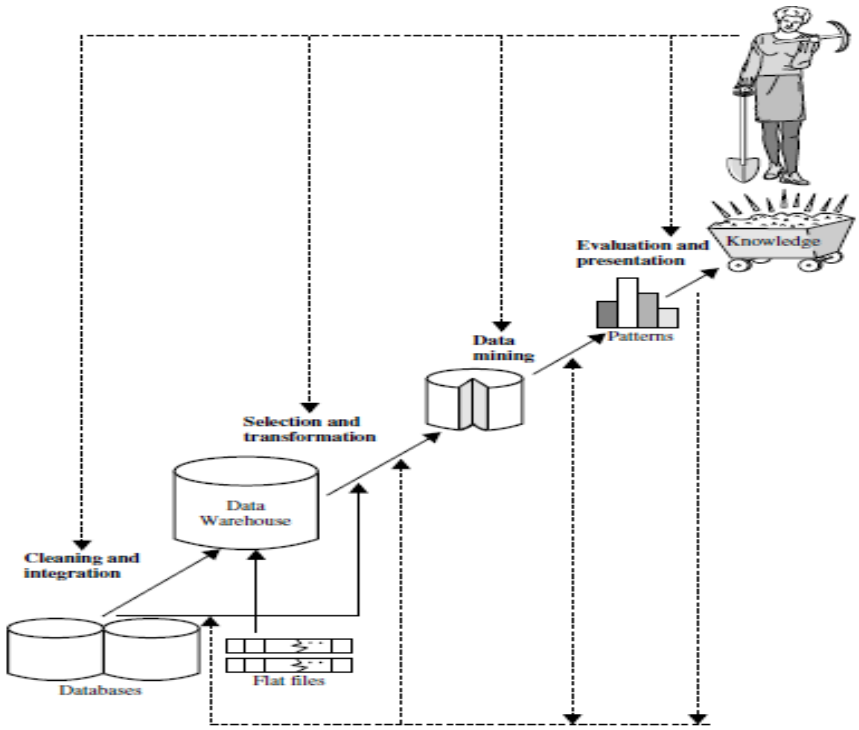
Time-Series Databases: Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.

The World-Wide Web provides rich, world-wide, on-line information services, where data objects are linked together to facilitate interactive access. Some examples of distributed information services associated with the World-Wide Web include America Online, Yahoo!, AltaVista, and Prodigy.

Data mining

Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called ***Knowledge Discovery in Database (KDD)***. The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

Many people treat data mining as a synonym for another popularly used term,**knowledge discovery from data**, or **KDD**, while others view data mining as merely an essential step in the process of knowledge discovery. The knowledge discovery process is shown in Figure as an iterative sequence of the following steps:



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 (to identify the truly interesting patterns representing knowledge based on *interestingness measures*

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

The preceding view shows data mining as one step in the knowledge discovery process,albeit an essential one because it uncovers hidden patterns for evaluation. However,in industry, in media, and in the research milieu, the term *data mining* is often used to refer to the entire knowledge discovery process (perhaps because the term is shorter than *knowledge discovery from data*). Therefore, we adopt a broad view of data mining functionality: **Data mining** is the *process* of discovering interesting patterns and knowledge from *large* amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically

Architecture of a typical data mining system.

The architecture of a typical data mining system may have the following major components

1. **Database, data warehouse, or other information repository.** This is one or a set of databases, data warehouses, spread sheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

2. **Database or data warehouse server.** The database or data warehouse server is responsible or fetching the relevant data, based on the user's data mining request.

3. **Knowledge base.** This is the domain knowledge that is used to guide the search, or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included.

4. **Data mining engine**. This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association analysis, classification, evolution and deviation analysis.

5. **Pattern evaluation module**. This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search towards interesting patterns. It may access interestingness thresholds stored in the knowledge base. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used.

6. **Graphical user interface.** This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task,

providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.

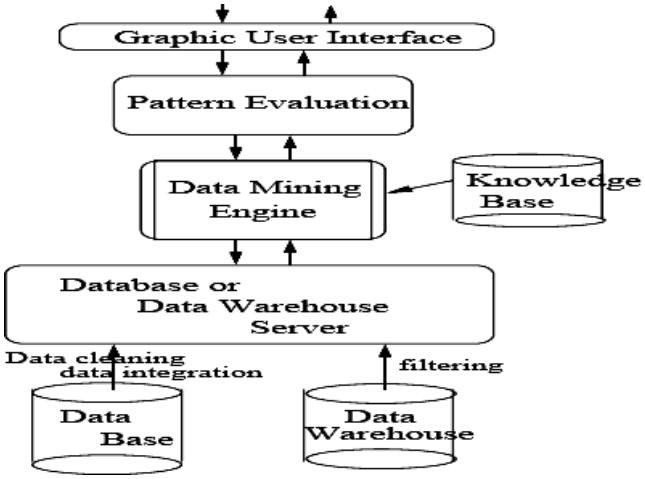


Figure: Architecture of a typical data mining system

Data Mining Functionalities

Data mining functions are used to define the trends or correlations contained in data mining activities. In comparison, data mining activities can be divided into two categories:

- **Descriptive Data Mining:** It includes certain knowledge to understand what is happening within the data without a previous idea. The common data features are highlighted in the data set. For example, count, average etc.
- **Predictive Data Mining:** It helps developers to provide unlabeled definitions of attributes. With previously available or historical data, data mining can be used to make predictions about critical business metrics based on data's linearity. For example, predicting the volume of business next quarter based on performance in the previous quarters over several years or judging from the findings of a patient's medical examinations that is he suffering from any particular disease.

Data mining functionalities are used to represent the type of patterns that have to be discovered in data mining tasks. Data mining is extensively used in many areas or sectors. It is used to predict and characterize data. But the ultimate objective in **Data Mining Functionalities** is to observe the various trends in data mining. There are several data mining functionalities that the organized and scientific methods offer, such as:

1. Class/Concept Descriptions

A class or concept implies there is a data set or set of features that define the class or a concept. A class can be a category of items on a shop floor, and a concept could be the abstract idea on which data may be categorized like products to be put on clearance sale and non-sale products. There are two concepts here, one that helps with grouping and the other that helps in differentiating.

- **Data Characterization:** This refers to the summary of general characteristics or features of the class, resulting in specific rules that define a target class. A data analysis technique called Attribute-oriented Induction is employed on the data set for achieving characterization.
- **Data Discrimination:** Discrimination is used to separate distinct data sets based on the disparity in attribute values. It compares features of a class with features of one or more contrasting classes.g., bar charts, curves and pie charts.

2. Mining Frequent Patterns

One of the functions of data mining is finding data patterns. Frequent patterns are things that are discovered to be most common in data. Various types of frequency can be found in the dataset.

- **Frequent item set:**This term refers to a group of items that are commonly found together, such as milk and sugar.
- **Frequent substructure:** It refers to the various types of data structures that can be combined with an item set or subsequences, such as trees and graphs.
- **Frequent Subsequence:** A regular pattern series, such as buying a phone followed by a cover.

3. Association Analysis

It analyses the set of items that generally occur together in a transactional dataset. It is also known as Market Basket Analysis for its wide use in retail sales. Two parameters are used for determining the association rules:

Support : It provides which identifies the common item set in the database.

Confidence : is the conditional probability that an item occurs when another item occurs in a transaction.

4. Classification

Classification is a data mining technique that categorizes items in a collection based on some predefined properties. It uses methods like if-then, decision trees or neural networks to predict a class or essentially classify a collection of items. A training set containing items whose properties are known is used to train the system to predict the category of items from an unknown collection of items.

5. Prediction

It predicts some unavailable data values or spending trends. An object can be anticipated based on the attribute values of the object and attribute values of the classes. It can be a prediction of missing numerical values or increase or decrease trends in time-related information. There are primarily two types of predictions in data mining: numeric and class predictions.

- ***Numeric predictions*** are made by creating a linear regression model that is based on historical data. Prediction of numeric values helps businesses ramp up for a future event that might impact the business positively or negatively.
- ***Class predictions*** are used to fill in missing class information for products using a training data set where the class for products is known.

6. Cluster Analysis

In image processing, pattern recognition and bioinformatics, clustering is a popular data mining functionality. It is similar to classification, but the classes are not predefined. Data attributes represent the classes. Similar data are grouped together, with the difference being that a class label is not known. Clustering algorithms group data based on similar features and dissimilarities.

7. Outlier Analysis

Outlier analysis is important to understand the quality of data. If there are too many outliers, you cannot trust the data or draw patterns. An outlier analysis determines if there is something out of turn in the data and whether it indicates a situation that a business needs to consider and take measures to mitigate. An outlier analysis of the data that cannot be grouped into any classes by the algorithms is pulled up.

8. Evolution and Deviation Analysis

Evolution Analysis pertains to the study of data sets that change over time. Evolution analysis models are designed to capture evolutionary trends in data helping to characterize, classify, cluster or discriminate time-related data.

9. Correlation Analysis

Correlation is a mathematical technique for determining whether and how strongly two attributes is related to one another. It refers to the various types of data structures, such as trees and graphs, that can be combined with an item set or subsequence. It determines how well two numerically measured continuous variables are linked. Researchers can use this type of analysis to see if there are any possible correlations between variables in their study.

Interestingness Patterns

A pattern is **interesting** if it is (1) *easily understood* by humans, (2) *valid* on new or test data with some degree of *certainty*, (3) potentially *useful*, and (4) *novel*. A pattern is also interesting if it validates a hypothesis that the user *sought to confirm*. An interesting pattern represents **knowledge**.

objective measures of pattern interestingness are based on the structure of discovered patterns and the statistics underlying them. An objective measure for association rules of the form $X \Rightarrow Y$ is rule **support**, representing the percentage of transactions from a transaction database that the given rule satisfies. This is taken to be the probability $P.(X \text{ U } Y)$ where X UY indicates that a transaction contains both X and Y, that is, the union of itemsets X and Y.

Another objective measure for association rules is **confidence**, which assesses the degree of certainty of the detected association. This is taken to be the conditional probability $P.(Y \mid X)$, that is, the probability

that a transaction containing X also contains Y.

$$support(X \Rightarrow Y) = P(X \cup Y),$$
$$confidence(X \Rightarrow Y) = P(Y|X).$$

In general, each interestingness measure is associated with a threshold, which may be controlled by the user. For example, rules that do not satisfy a confidence threshold of, say, 50% can be considered uninteresting. Rules below the threshold likely reflect noise,exceptions, or minority cases and are probably of less value.

Other objective interestingness measures include *accuracy* and *coverage* for classification(IF-THEN) rules. In general terms, accuracy tells us the percentage of data that are correctly classified by a rule. Coverage is similar to support, in that it tells us the percentage of data to which a rule applies.

.

Subjective interestingness measures are based on user beliefs in the data. These measures find patterns interesting if the patterns are **unexpected** (contradicting a user's belief) or offer strategic information on which the user can act. In the latter case, such patterns are referred to as **actionable**. For example, patterns like "a large earthquake often follows a cluster of small quakes" may be highly actionable if users can act on the information to save lives. Patterns that are **expected** can be interesting if they confirm a hypothesis that the user wishes to validate or they resemble a user's hunch.

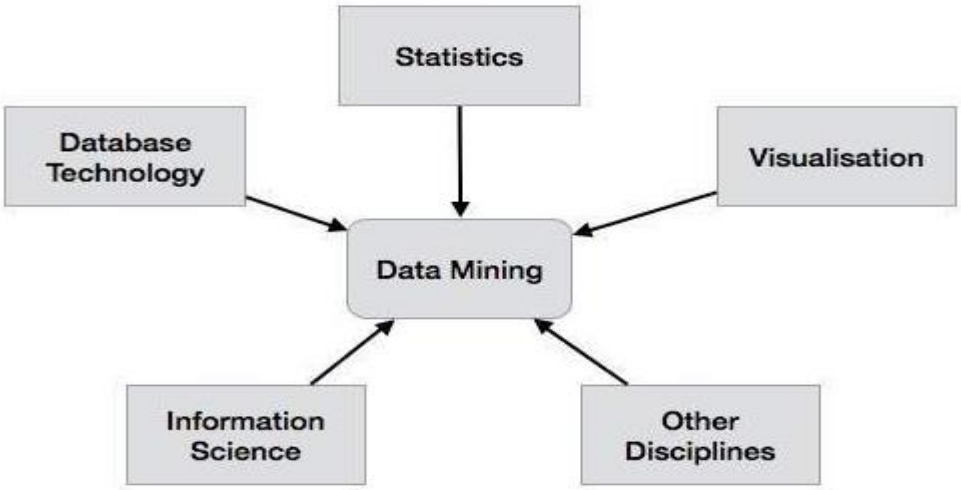
It is often unrealistic and inefficient for data mining systems to generate all possible patterns. Instead, user provided constraints and interestingness measures should be used to focus the search. For some mining tasks, such as association, this is often sufficient to ensure the completeness of the algorithm. Association rule mining is an example where the use of constraints and interestingness measures can ensure the completeness of mining.

It is highly desirable for data mining systems to generate only interesting patterns. This would be efficient for users and data mining systems because neither would have to search through the patterns generated to identify the truly interesting ones. Progress has been made in this direction; however, such optimization remains a challenging issue in data mining. Measures of pattern interestingness are essential for the efficient discovery of patterns by target users. Such measures can be used after the data mining step to rank the discovered patterns according to their interestingness, filtering out the uninteresting ones.More important, such measures can be used to guide and constrain the discovery process, improving the search efficiency by pruning away subsets of the pattern space that do not satisfy prespecified interestingness constraints.

Classification of Datamining Systems

A data mining system can be classified according to the following criteria −

- Database Technology
- Statistics
- Machine Learning
- Information Science
- Visualization
- Other Disciplines



Apart from these, a data mining system can also be classified based on the kind of
(a) databases mined, (b) knowledge mined, (c) techniques utilized, and (d) applications adapted.

Classification Based on the Databases Mined

We can classify a data mining system according to the kind of databases mined. Database system can be classified according to different criteria such as data models, types of data, etc. And the data mining system can be classified accordingly.For example, if we classify a database according to the data model, then we may have a relational, transactional, object-relational, or data warehouse mining system.

Classification Based on the kind of Knowledge Mined

We can classify a data mining system according to the kind of knowledge mined. It means the data mining system is classified on the basis of functionalities such as −

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Outlier Analysis

- Evolution Analysis

Classification Based on the Techniques Utilized

We can classify a data mining system according to the kind of techniques used. We can describe these techniques according to the degree of user interaction involved or the methods of analysis employed.

Classification Based on the Applications Adapted

We can classify a data mining system according to the applications adapted. These applications are as follows −

- Finance
- Telecommunications
- DNA
- Stock Markets
- E-mail

DataMining Task Primitives

A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during discovery to direct the mining process or examine the findings from different angles or depths. The data mining primitives specify the following,

1. Set of task-relevant data to be mined.
2. Kind of knowledge to be mined.
3. Background knowledge to be used in the discovery process.
4. Interestingness measures and thresholds for pattern evaluation.
5. Representation for visualizing the discovered patterns.

1. The set of task-relevant data to be mined

This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest (the relevant attributes or dimensions).In a relational database, the set of task-relevant data can be collected via a relational query involving operations like selection, projection, join, and aggregation.The data

collection process results in a new data relational called the ***initial data relation***. The initial data relation can be ordered or grouped according to the conditions specified in the query. This data retrieval can be thought of as a subtask of the data mining task. This initial relation may or may not correspond to physical relation in the database. Since virtual relations are called Views in the field of databases, the set of task-relevant data for data mining is called a minable view.

2. The kind of knowledge to be mined

This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

3. The background knowledge to be used in the discovery process

This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and evaluating the patterns found. Concept hierarchies are a popular form of background knowledge, which allows data to be mined at multiple levels of abstraction. Concept hierarchy defines a sequence of mappings from low-level concepts to higher-level, more general concepts.

- **Rolling Up - Generalization of data:** Allow to view data at more meaningful and explicit abstractions and makes it easier to understand. It compresses the data, and it would require fewer input/output operations.
- **Drilling Down - Specialization of data:** Concept values replaced by lower-level concepts. Based on different user viewpoints, there may be more than one concept hierarchy for a given attribute or dimension.

An example of a concept hierarchy for the attribute (or dimension) age is shown below. User beliefs regarding relationships in the data are another form of background knowledge.

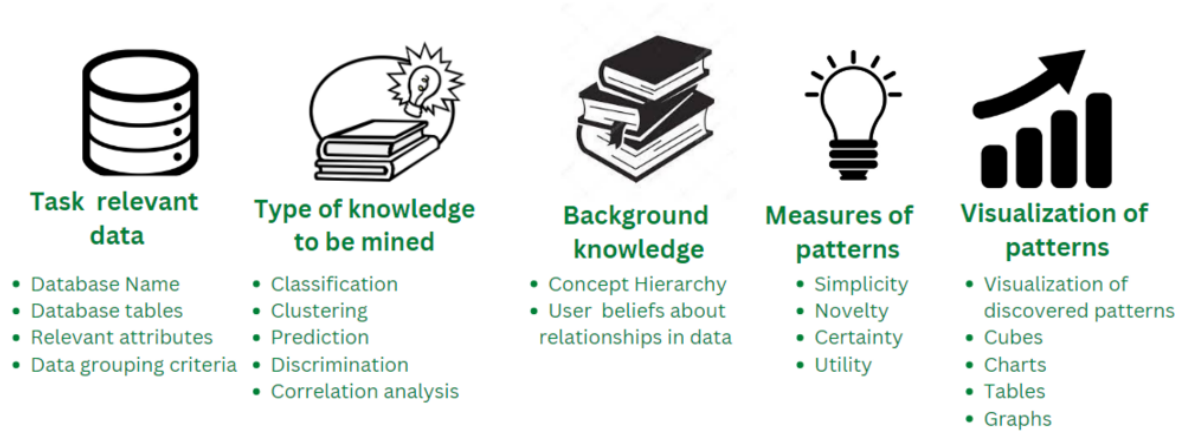
4. The interestingness measures and thresholds for pattern evaluation

Different kinds of knowledge may have different interesting measures. They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. For example, interesting measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.

- **Simplicity:** A factor contributing to the interestingness of a pattern is the pattern's overall simplicity for human comprehension. For example, the more complex the structure of a rule is, the more difficult it is to interpret, and hence, the less interesting it is likely to be. Objective measures of pattern simplicity can be viewed as functions of the pattern structure, defined in terms of the pattern size in bits or the number of attributes or operators appearing in the pattern.
- **Certainty (Confidence):** Each discovered pattern should have a measure of certainty associated with it that assesses the validity or "trustworthiness" of the pattern. A certainty measure for association rules of the form "A =>B" where A and B are sets of items is confidence. Confidence is a certainty measure. Given a set of task-relevant data tuples, the confidence of "A=>B" is defined as
  Confidence (A=>B) = # tuples containing both A and B /# tuples containing A
- **Utility (Support):** The potential usefulness of a pattern is a factor defining its interestingness. It can be estimated by a utility function, such as support. The support of an association pattern refers to the percentage of task-relevant data tuples (or transactions) for which the pattern is true.
  Utility (support): usefulness of a pattern
  Support (A=>B) = # tuples containing both A and B / total #of tuples
- **Novelty:** Novel patterns are those that contribute new information or increased performance to the given pattern set. For example -> A data exception. Another strategy for detecting novelty is to remove redundant patterns.

5. The expected representation for visualizing the discovered patterns

This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, cross tabs, charts, graphs, decision trees, cubes, or other visual representations.Users must be able to specify the forms of presentation to be used for displaying the discovered patterns. Some representation forms may be better suited than others for particular kinds of knowledge.For example, generalized relations and their corresponding cross tabs or pie/bar charts are good for presenting characteristic descriptions, whereas decision trees are common for classification.



Integration of Data Mining System with a Data warehouse

If a data mining system is not integrated with a database or a data warehouse system, then there will be no system to communicate with. This scheme is known as the non-coupling scheme. In this scheme, the main focus is on data mining design and on developing efficient and effective algorithms for mining the available data sets.

The list of Integration Schemes is as follows −

- **No Coupling** − In this scheme, the data mining system does not utilize any of the database or data warehouse functions. It fetches the data from a particular source and processes that data using some data mining algorithms. The data mining result is stored in another file.
- **Loose Coupling** − In this scheme, the data mining system may use some of the functions of database and data warehouse system. It fetches the data from the data respiratory managed by these systems and performs data mining on that data. It then stores the mining result either in a file or in a designated place in a database or in a data warehouse.
- **Semi−tight Coupling** − In this scheme, the data mining system is linked with a database or a data warehouse system and in addition to that, efficient implementations of a few data mining primitives can be provided in the database.
- **Tight coupling** − In this coupling scheme, the data mining system is smoothly integrated into the database or data warehouse system. The data mining subsystem is treated as one functional component of an information system.

Major Issues in DataMining

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues.

- **Mining Methodology**

In addition, mining methodologies should consider issues such as data uncertainty, noise, and incompleteness. Some mining methods explore how user specified measures can be used to assess the interestingness of discovered patterns as well as guide the discovery process.

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.

Boosting the power of discovery in a networked environment: Most data objects reside in a linked or interconnected environment, whether it be the Web, database relations, files, or

documents. Semantic links across multiple data objects can be used to advantage in data mining. Knowledge derived in one set of objects can be used to boost the discovery of knowledge in a "related" or semantically linked set of objects.

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 pre-processing, outlier detection and removal, and uncertainty reasoning are examples of techniques that need to be integrated with the data mining process.

Pattern evaluation and pattern- or constraint-guided mining: Not all the patterns generated by data mining processes are interesting. What makes a pattern interesting may vary from user to user. Therefore, techniques are needed to assess the interestingness of discovered patterns based on subjective measures. These estimate the value of patterns with respect to a given user class, based on user beliefs or expectations. Moreover, by using interestingness measures or user-specified constraints to *guide* the discovery process, we may generate more interesting patterns and reduce the search space.

- **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.

Ad hoc data mining and data mining query languages: Query languages (e.g., SQL) have played an important role in flexible searching because they allow users to pose ad hoc queries. Similarly, high-level data mining query languages or other high-level flexible user interfaces will give users the freedom to define ad hoc data mining tasks. This should facilitate specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered patterns. Optimization of the processing of such flexible mining requests is another promising area of study.

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 tomultiply, 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.

Cloud computing* and *cluster computing, which use computers in a distributed and collaborative way to tackle very large-scale computational tasks, are also active research themes in parallel data mining. In addition, the high cost of some data mining processes and the incremental nature of input promote **incremental** data mining, which incorporates new data updates without having to mine the entire data "from scratch." Such methods perform knowledge modification incrementally to amend and strengthen what was previously discovered.

- **Diversity of Database Types**

The wide diversity of database types brings about challenges to data mining. These include

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. Mining such gigantic, interconnected information networks may help disclose many more patterns and knowledge in heterogeneous data sets than can be discovered from a small set of isolated data repositories. Web mining, multisource data mining, and information network mining have become challenging and fast-evolving data mining fields.

- **Data Mining and Society**

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 use data 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.

Data Preprocessing

Data pre-processing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.Preprocessing of data is mainly to check the data quality. The quality can be checked by the following:

- **Accuracy**: To check whether the data entered is correct or not.
- **Completeness**: To check whether the data is available or not recorded.
- **Consistency:** To check whether the same data is kept in all the places that do or do not match.
- **Timeliness**: The data should be updated correctly.
- **Believability**: The data should be trustable.
- **Interpretability**: The understandability of the data.

There are 4 major tasks in data preprocessing – Data cleaning, Data integration, Data reduction, and Data transformation.

Data Cleaning

Data cleaning is the process of removing incorrect data, incomplete data, and inaccurate data from the datasets, and it also replaces the missing values. Here are some techniques for data cleaning:

Handling Missing Values

- Standard values like “Not Available” or “NA” can be used to replace the missing values.
- Missing values can also be filled manually, but it is not recommended when that dataset is big.
- The attribute’s mean value can be used to replace the missing value when the data is normally distributed wherein in the case of non-normal distribution median value of the attribute can be used.
- While using regression or decision tree algorithms, the missing value can be replaced by the most probable value.

Handling Noisy Data

Noisy generally means random error or containing unnecessary data points. Handling noisy data is one of the most important steps as it leads to the optimization of the model we are using Here are some of the methods to handle noisy data.

- **Binning:** This method is to smooth or handle noisy data. First, the data is sorted then, and then the sorted values are separated and stored in the form of bins. There are three methods for smoothing data in the bin. **Smoothing by bin mean method**: In this method, the values in the bin are replaced by the mean value of the bin; **Smoothing by bin median**: In this method, the values in the bin are replaced by the median value; **Smoothing by bin boundary**: In this method, the using minimum and maximum values of the bin values are taken, and the closest boundary value replaces the values.
- **Regression:** This is used to smooth the data and will help to handle data when unnecessary data is present. For the analysis, purpose regression helps to decide the variable which is suitable for our analysis.

- **Clustering:** This is used for finding the outliers and also in grouping the data. Clustering is generally used in unsupervised learning.

  Data Integration

  The process of combining multiple sources into a single dataset. The Data integration process is one of the main components of data management. There are some problems to be considered during data integration.

  • ***Schema integration***: Integrates metadata(a set of data that describes other data) from different sources.

  • ***Entity identification problem:*** Identifying entities from multiple databases. For example, the system or the user should know the student *id of one database and student*name of another database belonging to the same entity.

  • ***Detecting and resolving data value concepts****:* The data taken from different databases while merging may differ. The attribute values from one database may differ from another database. For example, the date format may differ, like “MM/DD/YYYY” or “DD/MM/YYYY”.

  Data Reduction

  This process helps in the reduction of the volume of the data, which makes the analysis easier yet produces the same or almost the same result. This reduction also helps to reduce storage space. Some of the data reduction techniques are dimensionality reduction, numerosity reduction, and data compression.

- ***Dimensionality reduction:*** This process is necessary for real-world applications as the data size is big. In this process, the reduction of random variables or attributes is done so that the dimensionality of the data set can be reduced. Combining and merging the attributes of the data without losing its original characteristics. This also helps in the reduction of storage space, and computation time is reduced. When the data is highly dimensional, a problem called the “Curse of Dimensionality” occurs.

- ***Numerosity Reduction****:* In this method, the representation of the data is made smaller by reducing the volume. There will not be any loss of data in this reduction.
- ***Data compression****:* The compressed form of data is called data compression. This compression can be lossless or lossy. When there is no loss of information during compression, it is called lossless compression. Whereas lossy compression reduces information, but it removes only the unnecessary information.

Data Transformation

The change made in the format or the structure of the data is called data transformation. This step can be simple or complex based on the requirements. There are some methods for data transformation.

- ***Smoothing:*** With the help of algorithms, we can remove noise from the dataset, which helps in knowing the important features of the dataset. By smoothing, we can find even a simple change that helps in prediction.
- ***Aggregation:*** In this method, the data is stored and presented in the form of a summary. The data set, which is from multiple sources, is integrated into with data analysis description. This is an important step since the accuracy of the data depends on the quantity and quality of the data. When the quality and the quantity of the data are good, the results are more relevant.
- ***Discretization:*** The continuous data here is split into intervals. Discretization reduces the data size. For example, rather than specifying the class time, we can set an interval like (3 pm-5 pm, or 6 pm-8 pm).
- ***Normalization:*** It is the method of scaling the data so that it can be represented in a smaller range. Example ranging from -1.0 to 1.0.

Note : Refer to Unit-I PPT for Preprocessing Examples