Data Mining

Content to be covered

- **Introduction to Data Mining**:
  - What is data mining,
  - Challenges
  - Data Mining Tasks,
  - Phases of data mining
  - Benefits of data mining,
  - What Kinds of Data Can Be Mined,
  - What Kinds of Patterns Can Be Mined,
  - Major Issues in Data Mining,

What is data mining?

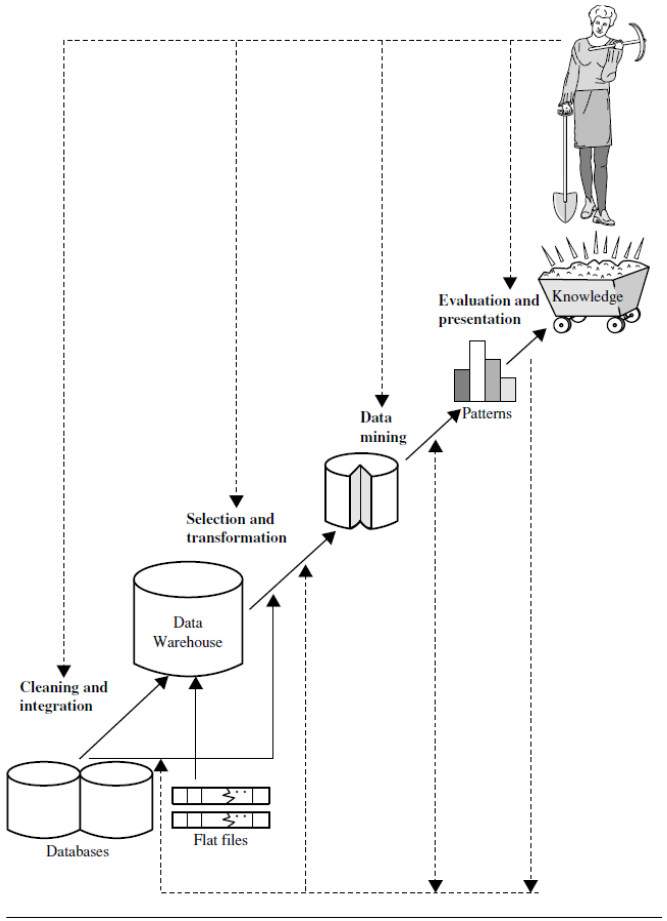
Mining is a process that finds a small set of precious nuggets from a great deal of raw materials

Data mining is

- knowledge mining from data,
- knowledge extraction,
- data/pattern analysis,
- data archaeology,
- data dredging

Many treat data mining as a synonym for knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery

Knowledge Discovery Process



Knowledge Discovery Process

- **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 and consolidated into forms appropriate for mining by performing summary or aggregation operations)
- **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
- **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on *interestingness measures*)
- **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Data Mining

- **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.

What Kinds of Data Can Be Mined?

- As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application.
- The most basic forms of data for mining applications are
  - database data
  - data warehouse data
  - transactional data
- Data mining can also be applied to other forms of data (e.g., data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW)

Database Data

- A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
- A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows). Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values. An ER data model represents the database as a set of entities and their relationships.

customer	(*cust_ID, name, address, age, occupation, annual_income, credit_information, category, . . .*)
item	(*item_ID, brand, category, type, price, place_made, supplier, cost, . . .*)
employee	(*empl_ID, name, category, group, salary, commission, . . .*)
branch	(*branch_ID, name, address, . . .*)
purchases	(*trans_ID, cust_ID, empl_ID, date, time, method_paid, amount*)
items_sold	(*trans_ID, item_ID, qty*)
works_at	(*empl_ID, branch_ID*)

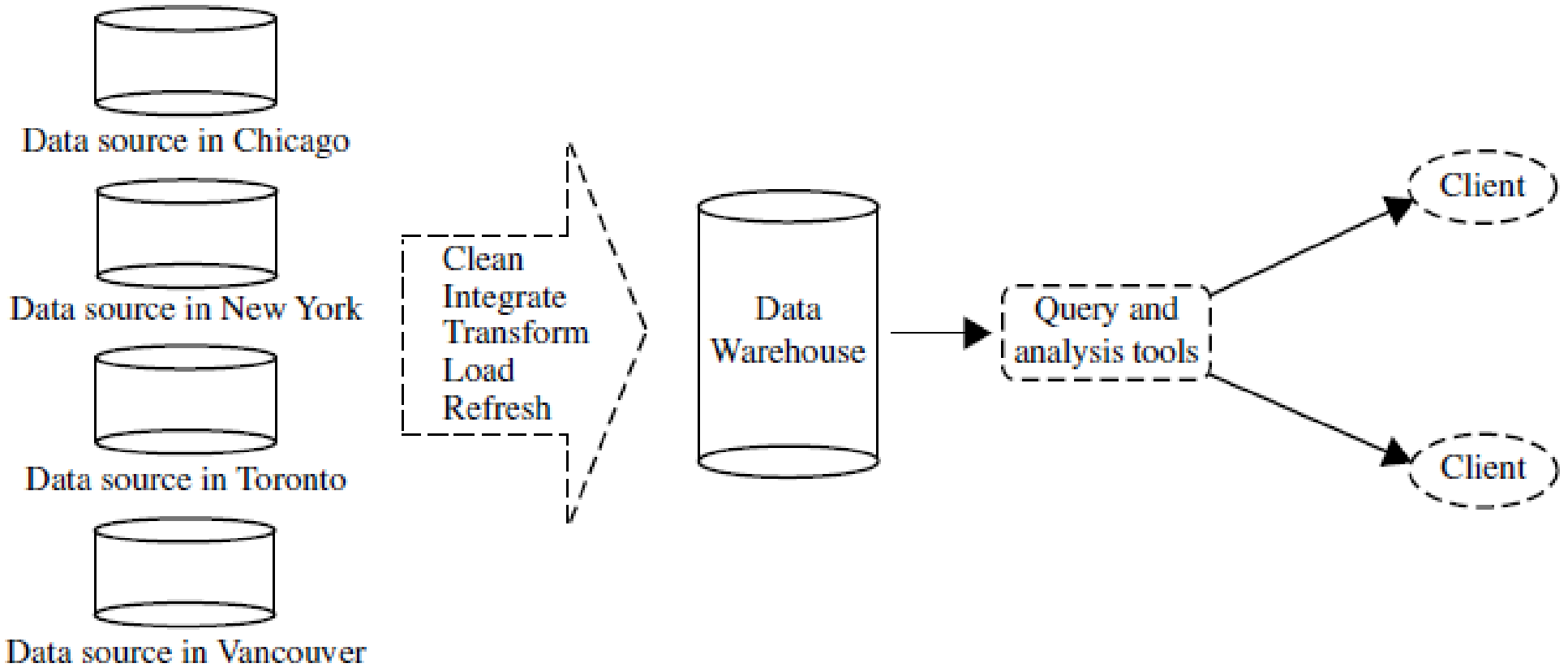
Database Data

- Relational data can be accessed by database queries written in a relational query language (e.g., SQL) or with the assistance of graphical user interfaces. A given query is transformed into a set of relational operations, such as join, selection, and projection, and is then optimized for efficient processing.
- A query allows retrieval of specified subsets of the data. Through the use of relational queries, you can ask things like, “Show me a list of all items that were sold in the last quarter.”
- Relational languages also use aggregate functions such as sum, avg (average), count, max (maximum), and min (minimum).
- When mining relational databases, we can go further by searching for trends or data patterns. Data mining systems may also detect deviations. Such deviations can then be further investigated.
- Relational databases are a major data form in the study of data mining.

Data Warehouses

- A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
- To facilitate decision making, the data in a data warehouse are organized around major subjects (e.g., customer, item, supplier, and activity). The data are stored to provide information from a historical perspective, such as in the past 6 to 12 months, and are typically summarized.
- A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which 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 sum (sales_amount). A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data.

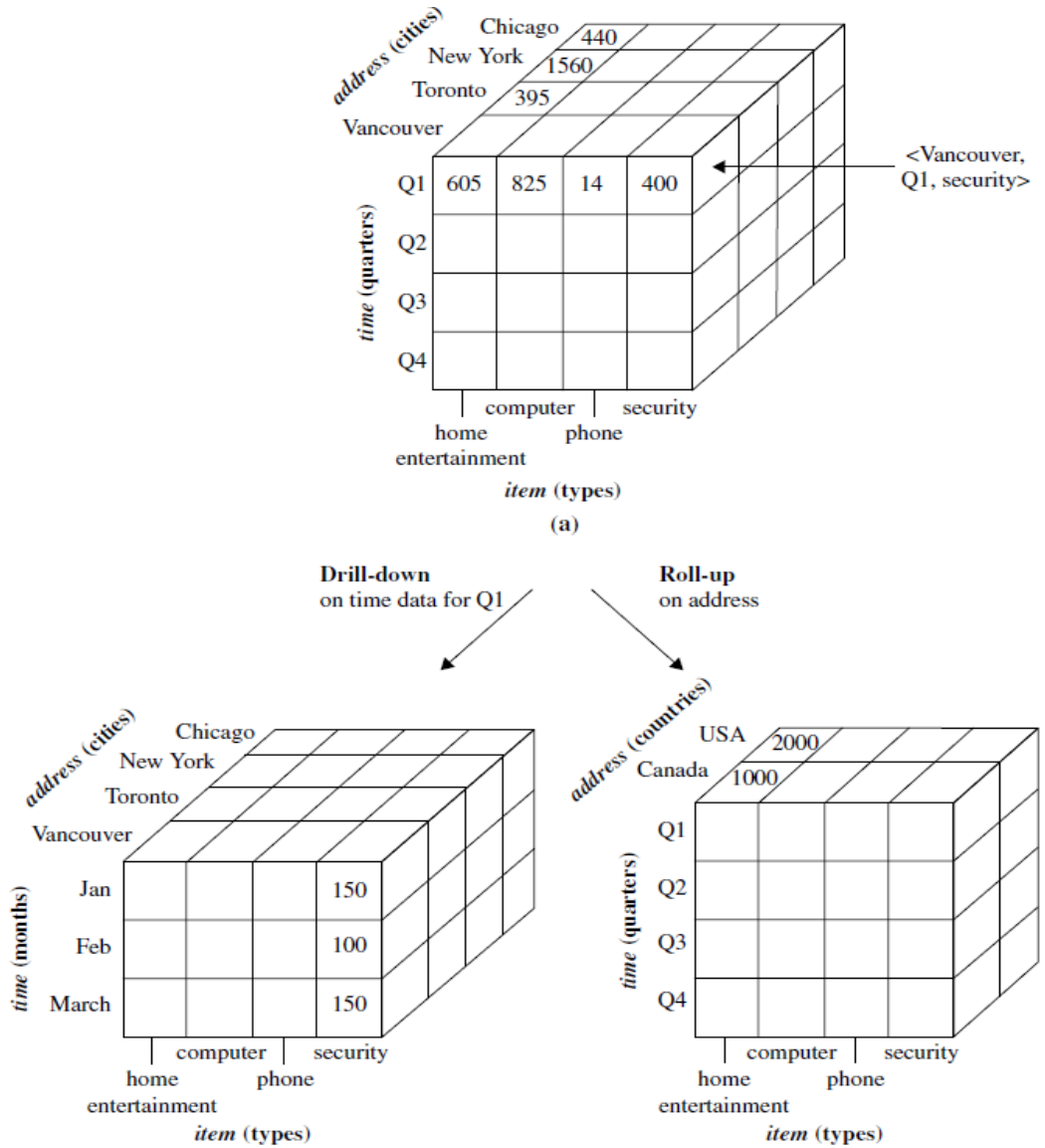
Data Warehouse Framework



Data Warehouses

- By providing multidimensional data views and the precomputation of summarized data, data warehouse systems can provide inherent support for OLAP. Online analytical processing operations make use of background knowledge regarding the domain of the data being studied 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
- For Example, we can drill down on sales data summarized by quarter to see data summarized by month. Similarly, we can roll up on sales data summarized by city to view data summarized by country.

Data Warehouses



Transactional Data

- In general, each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page.
- A transaction typically includes a unique transaction identity number (trans ID) and a list of the items making up the transaction, such as the items purchased in the transaction.
- A transactional database may have additional tables, which contain other information related to the transactions, such as item description, information about the salesperson or the branch, and so on.

trans_ID	*list_of_item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
…	…

Other Kinds of Data

- Time-related or sequence data (e.g., historical records, stock exchange data, and time-series and biological sequence data),
- Data streams (e.g., video surveillance and sensor data, which are continuously transmitted),
- Spatial data (e.g., maps),
- Engineering design data (e.g., the design of buildings, system components, or integrated circuits),
- Hypertext and multimedia data (including text, image, video, and audio data),
- Graph and networked data (e.g., social and information networks),
- and the Web (a huge, widely distributed information repository made available by the Internet).
- These applications bring about new challenges, like how to handle data carrying special structures (e.g., sequences, trees, graphs, and networks) and specific semantics (such as ordering, image, audio and video contents, and connectivity), and how to mine patterns that carry rich structures and semantics.

Other Kinds of Data

- Regarding temporal data, for instance, we can mine banking data for changing trends, which may aid in the scheduling of bank tellers according to the volume of customer traffic.
- Stock exchange data can be mined to uncover trends that could help you plan investment strategies.
- We could mine computer network data streams to detect intrusions based on the anomaly of message flows, which may be discovered by clustering, dynamic construction of stream models or by comparing the current frequent patterns with those at a previous time.
- With spatial data, we may look for patterns that describe changes in metropolitan poverty rates based on city distances from major highways. The relationships among a set of spatial objects can be examined in order to discover which subsets of objects are spatially autocorrelated or associated.
- By mining text data, such as literature on data mining from the past ten years, we can identify the evolution of hot topics in the field.
- By mining user comments on products, we can assess customer sentiments and understand how well a product is embraced by a market.
- From multimedia data, we can mine images to identify objects and classify them by assigning semantic labels or tags. By mining video data of a hockey game, we can detect video sequences corresponding to goals.
- *Web mining* can help us learn about the distribution of information on the WWW in general, characterize and classify web pages, and uncover web dynamics and the association and other relationships among different web pages, users, communities, and web-based activities.

Other Kinds of Data

- It is important to keep in mind that, in many applications, multiple types of data are present. For example, in web mining, there often exist text data and multimedia data (e.g., pictures and videos) on web pages, graph data like web graphs, and map data on some web sites.
- Mining multiple data sources of complex data often leads to fruitful findings due to the mutual enhancement and consolidation of such multiple sources.
- On the other hand, it is also challenging because of the difficulties in data cleaning and data integration, as well as the complex interactions among the multiple sources of such data.
- While such data require sophisticated facilities for efficient storage, retrieval, and updating, they also provide fertile ground and raise challenging research and implementation issues for data mining.

What Kinds of Patterns Can Be Mined?

Data mining functionalities.

- Characterization and discrimination
- The mining of frequent patterns, associations, and correlations
- Classification and regression
- clustering analysis
- outlier analysis.
- Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. Such tasks can be classified into two categories: descriptive and predictive.
- Descriptive mining tasks characterize properties of the data in a target data set. Predictive mining tasks perform induction on the current data in order to make predictions.

Class/Concept Description: Characterization and Discrimination

- Data entries can be associated with classes or concepts. Eg. classes of items for sale include computers and printers, and concepts of customers include bigSpenders and budgetSpenders.
- It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived using
  - (1) data characterization, by summarizing the data of the class under study (often called the target class) in general terms,
  - (2) data discrimination, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes),
  - (3) both data characterization and discrimination.

Class/Concept Description: Characterization and Discrimination

- Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query.

Methods for effective data summarization and characterization.

- Simple data summaries based on statistical measures and plots
- The data cube-based OLAP roll-up operation can be used to perform user-controlled data summarization along a specified dimension.
- Data warehousing.
- An attribute-oriented induction technique can be used to perform data generalization and characterization

The output of data characterization can be presented in various forms. Eg. pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables, including crosstabs. The resulting descriptions can also be presented as generalized relations or in rule form (called characteristic rules).

Class/Concept Description: Characterization and Discrimination

- Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.
- The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries.
- For example, a user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period.
- The methods used for data discrimination are similar to those used for data characterization.
- “How are discrimination descriptions output?” The forms of output presentation are similar to those for characteristic descriptions, although discrimination descriptions should include comparative measures that help to distinguish between the target and contrasting classes. Discrimination descriptions expressed in the form of rules are referred to as discriminant rules.

Mining Frequent Patterns, Associations, and Correlations

- Frequent patterns, are patterns that occur frequently in data. There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures.
- A frequent itemset typically refers to a set of items that often appear together in a transactional data set. Eg. milk and bread, which are frequently bought together
- A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern.
- A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences.
- If a substructure occurs frequently, it is called a (frequent) structured pattern. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

Buys(X, “computer”) => buys(X, “software”) [support = 1%, confidence = 50%],

Age(X, “20..29”)^income(X, “40K..49K”) => buys(X, “laptop”) [support = 2%, confidence = 60%].

Classification and Regression for Predictive Analysis

- Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The model are derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown.
- “How is the derived model presented?” The derived model may be represented in various forms, such as classification rules (i.e., IF-THEN rules), decision trees, mathematical formulae, or neural networks.
- A decision tree is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can easily be converted to classification rules.
- A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units. There are many other methods for constructing classification models, such as naïve Bayesian classification, support vector machines, and k-nearest-neighbor classification.

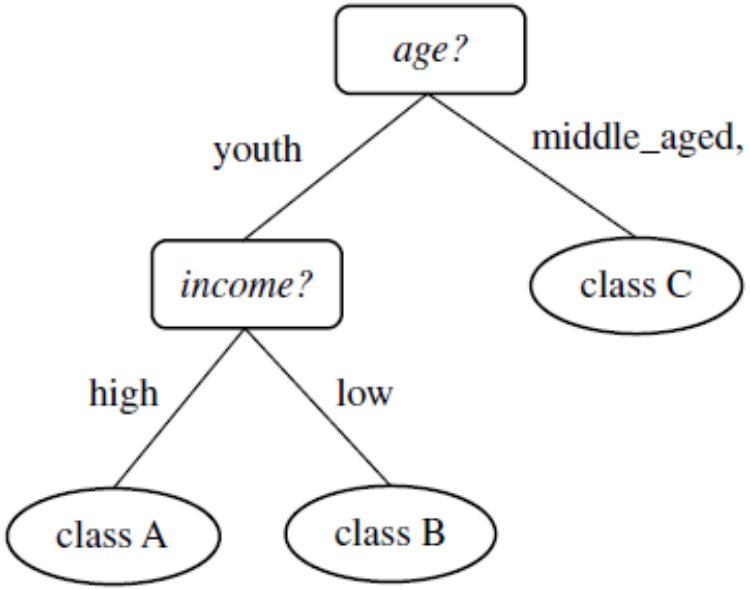
Classification and Regression for Predictive Analysis

- Classification predicts categorical (discrete, unordered) labels.
- Regression models are continuous-valued functions. That is, regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels.
- The term prediction refers to both numeric prediction and class label prediction.
- Regression also encompasses the identification of distribution trends based on the available data.
- Classification and regression may need to be preceded by relevance analysis, which attempts to identify attributes that are significantly relevant to the classification and regression process. Such attributes will be selected for the classification and regression process. Other attributes, which are irrelevant, can then be excluded from consideration.
- Eg. classify a large set of items in the store, based on three kinds of responses to a sales campaign: good response, mild response and no response. You want to derive a model for each of these three classes based on the descriptive features of the items, such as price, brand, place made, type, and category. The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set.

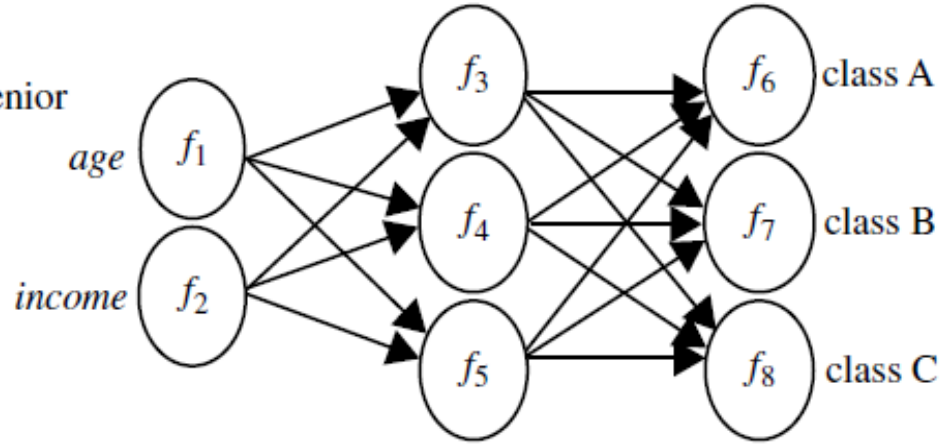
Classification and Regression for Predictive Analysis

age(X, "youth") AND income(X, "high") ⟶ *class(X, "A")*
age(X, "youth") AND income(X, "low") ⟶ *class(X, "B")*
age(X, "middle_aged") ⟶ *class(X, "C")*
age(X, "senior") ⟶ *class(X, "C")*

(a)



(b)

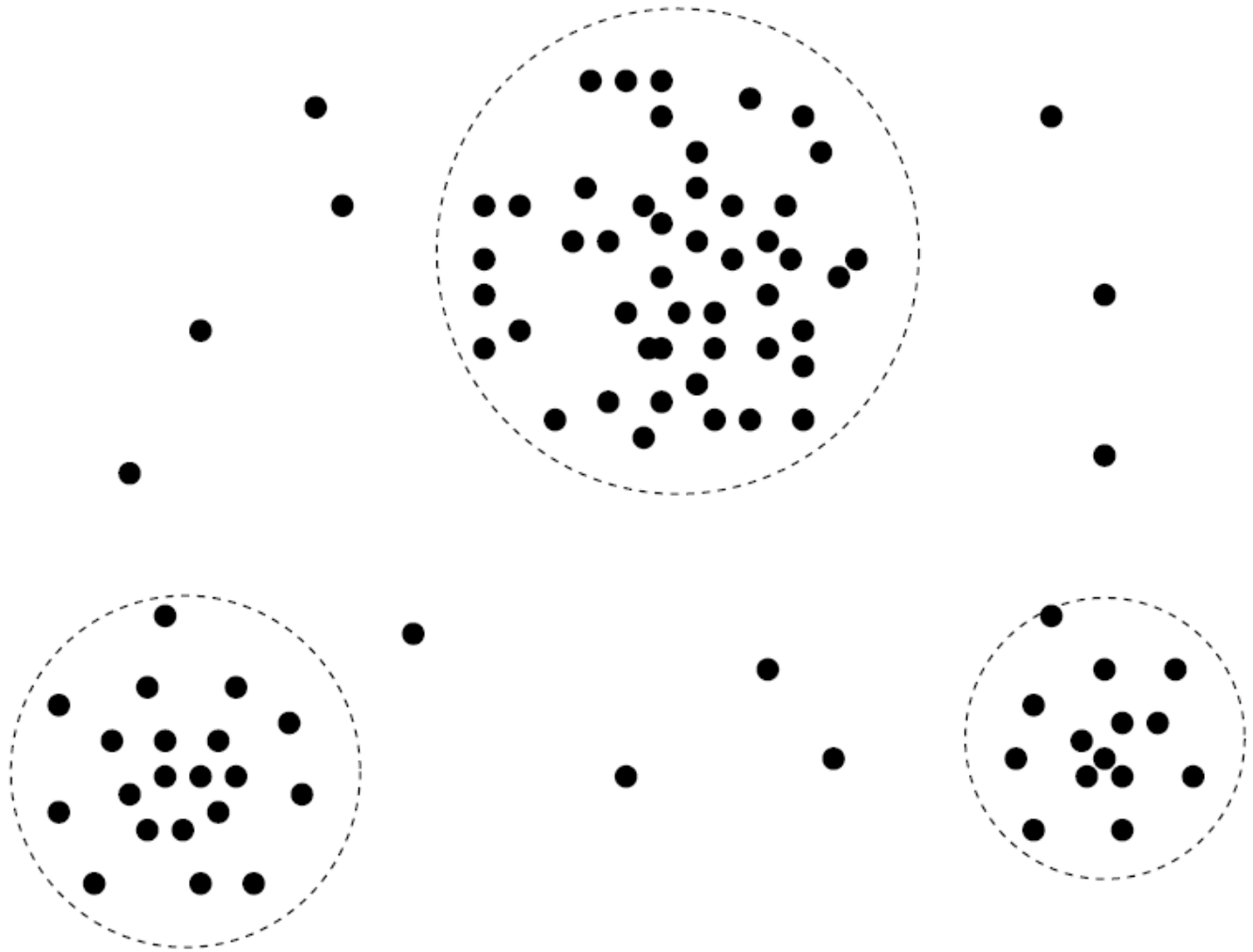


(c)

Cluster Analysis

- Unlike classification and regression, which analyze class-labeled (training) data sets, clustering analyzes data objects without consulting class labels.
- In many cases, class labeled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data.
- The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters.
- Each cluster so formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.
- Cluster analysis can be performed customer data to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.

Cluster Analysis



Outlier Analysis

- A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are outliers.
- Many data mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection) the rare events can be more interesting than the more regularly occurring ones.
- The analysis of outlier data is referred to as outlier analysis or anomaly mining.
- Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are remote from any other cluster are considered outliers.
- Rather than using statistical or distance measures, density-based methods may identify outliers in a local region, although they look normal from a global statistical distribution view.
- Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account.

Are All Patterns Interesting?

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,

(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

- An objective measure for association rules of the form X=>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(XUY), where XUY 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|X), that is, the probability that a transaction containing X also contains Y.
- 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.

Objective measures of pattern interestingness

- 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. Regarding understandability, we may use simple objective measures that assess the complexity or length in bits of the patterns mined.
- Although objective measures help identify interesting patterns, they are often insufficient unless combined with subjective measures that reflect a particular user's needs and interests.
- For example, patterns describing the characteristics of customers who shop frequently should be interesting to the marketing manager, but may be of little interest to other analysts studying the same database for patterns on employee performance.
- Many patterns that are interesting by objective standards may represent common sense and, therefore, are actually uninteresting.

Subjective interestingness measures

- 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.
- The second question—"Can a data mining system generate all of the interesting patterns?"— refers to the completeness of a data mining algorithm. 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.

Subjective interestingness measures

- Finally, the third question—“Can a data mining system generate only interesting patterns?”— is an optimization problem in data 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.
- 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. 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.

Benefits of Data Mining

- “We are living in the information age” is a popular saying; however, we are actually living in the data age.
- Terabytes or petabytes (1 million gigabytes) of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business, society, science and engineering, medicine, and almost every other aspect of daily life.
- Businesses worldwide generate gigantic data sets, including sales transactions, stock trading records, product descriptions, sales promotions, company profiles and performance, and customer feedback.
- Scientific and engineering practices generate high orders of petabytes of data in a continuous manner, from remote sensing, process measuring, scientific experiments, system performance, engineering observations, and environment surveillance.
- Global backbone telecommunication networks carry tens of petabytes of data traffic every day.
- The medical and health industry generates tremendous amounts of data from medical records, patient monitoring, and medical imaging.
- Billions of Web searches supported by search engines process tens of petabytes of data daily.
- Communities and social media have become increasingly important data sources, producing digital pictures and videos, blogs, Web communities, and various kinds of social networks. The list of sources that generate huge amounts of data is endless.
- Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining.

Benefits of Data Mining

- Data mining can be viewed as a result of the natural evolution of information technology. The database and data management industry evolved in the development of several critical functionalities: data collection and database creation, data management (including data storage and retrieval and database transaction processing), and advanced data analysis (involving data warehousing and data mining).
- Huge volumes of data have been accumulated beyond databases and data warehouses. During the 1990s, the World Wide Web and web-based databases (e.g., XML databases) began to appear. Internet-based global information bases, such as the WWW and various kinds of interconnected, heterogeneous databases, have emerged and play a vital role in the information industry. The effective and efficient analysis of data from such different forms of data by integration of information retrieval, data mining, and information network analysis technologies is a challenging task.
- In summary, the abundance of data, coupled with the need for powerful data analysis tools, has been described as a data rich but information poor situation. The fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools. Consequently, important decisions are often made based not on the information-rich data stored in data repositories but rather on a decision maker's intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. The manual knowledge input procedure is prone to biases and errors and is extremely costly and time consuming. The widening gap between data and information calls for the systematic development of data mining tools that can turn data tombs into "golden nuggets" of knowledge.

Major Issues in Data Mining

- Mining methodology,
- User interaction,
- Efficiency and scalability,
- Diversity of data types, and
- Data mining and society.

Mining methodology

- Mining various and new kinds of knowledge
- Mining knowledge in multidimensional space
- Data mining—an interdisciplinary effort
- Boosting the power of discovery in a networked environment
- Handling uncertainty, noise, or incompleteness of data
- Pattern evaluation and pattern- or constraint-guided mining

User Interaction

- Interactive mining
  - 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
- Incorporation of background knowledge
  - Background knowledge, constraints, rules, and other information regarding the domain under study should be incorporated
- Ad hoc data mining and data mining query languages
- Presentation and visualization of data mining results

Efficiency and Scalability

- Efficiency and scalability of data mining algorithms
  - The running time of a data mining algorithm must be predictable, short, and acceptable. 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

Diversity of Database Types

- Handling complex types of data
  - 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.
- Mining dynamic, networked, and global data repositories

Data Mining and Society

- Social impacts of data mining:
  - 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 poses the risk of disclosing an individual's personal information. The philosophy is to observe data sensitivity and preserve people's privacy while performing successful data mining.
- Invisible data mining