Lecture 1 - Introduction to Data Mining and Knowledge Discovery

Adnan Ferdous Ashrafi

Stamford University Bangladesh

Table of Contents

WHAT IS DATA MINING?

DATA MINING

"Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques." [1]

Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data is invariably present in substantial quantities.

WHAT IS DATA MINING?

Definition by Hand et al. [2]

"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner"

Definition by Evangelos Simoudis in Cabena et al. [3]

"Data mining is an interdisciplinary field bringing togther techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases"

Knowledge discovery in databases

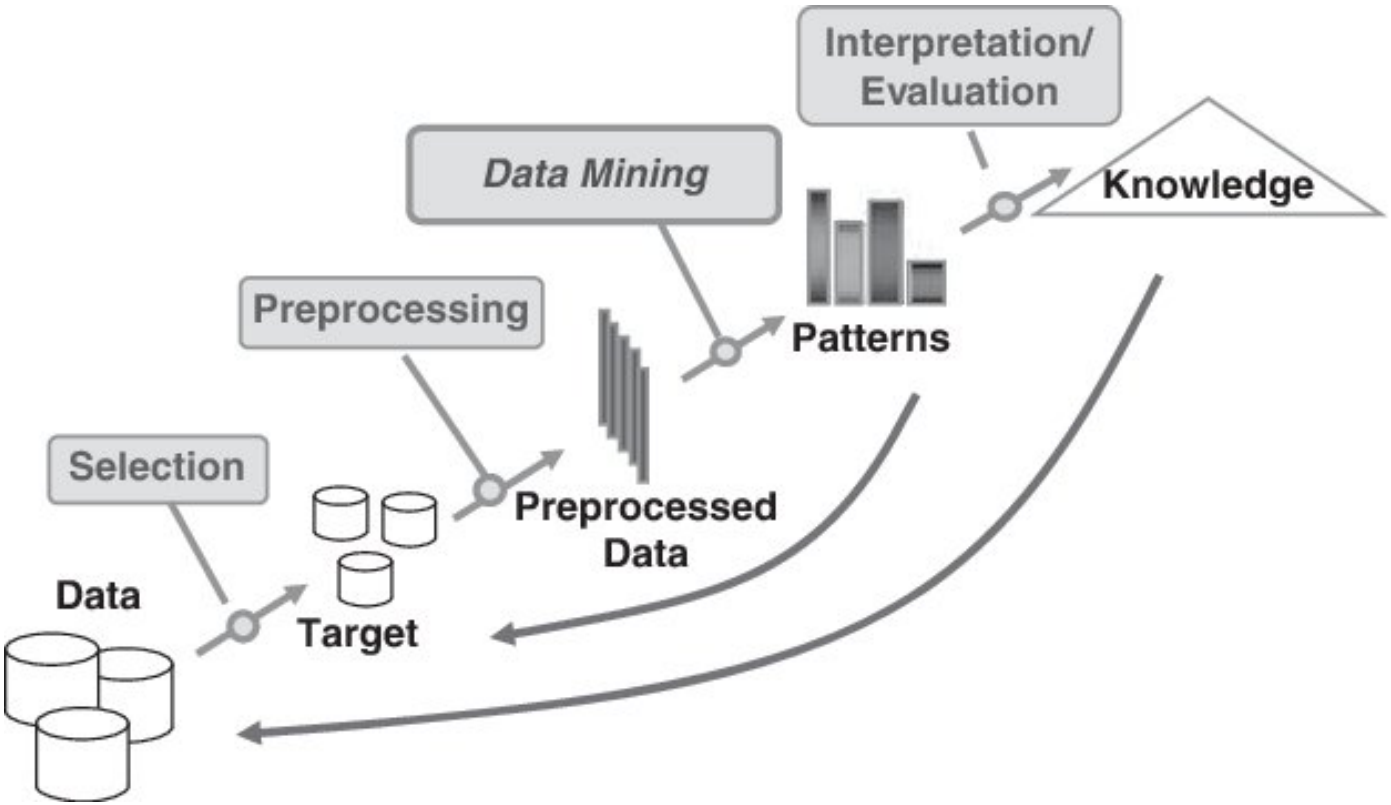

Figure 1: The process of knowledge discovery in databases (KDD)[4]

WHY DATA MINING?

As early as 1984, in his book Megatrends [5], John Naisbitt observed that "we are drowning in information but starved for knowledge." The problem today is not that there is not enough data and information streaming in. We are, in fact, inundated with data in most fields. Rather, the problem is that there are not enough trained human analysts available who are skilled at translating all of this data into knowledge, and thence up the taxonomy tree into wisdom. The ongoing remarkable growth in the field of data mining and knowledge discovery has been fueled by a fortunate confluence of a variety of factors:

1. The explosive growth in data collection
2. The storing of the data in data warehouses, so that the entire enterprise has access to a reliable current database
3. The availability of increased access to data from Web navigation and intranets
4. The competitive pressure to increase market share in a globalized economy
5. The development of off-the-shelf commercial data mining software suites
6. The tremendous growth in computing power and storage capacity

NEED FOR HUMAN DIRECTION OF DATA MINING

Automation is no substitute for human input. As we shall learn shortly, humans need to be actively involved at every phase of the data mining process.

Georges Grinstein of the University of Massachusetts at Lowell and AnVil, Inc., stated it like this [6]:

"Imagine a black box capable of answering any question it is asked. Any question. Will this eliminate our need for human participation as many suggest? Quite the opposite. The fundamental problem still comes down to a human interface issue. How do I phrase the question correctly? How do I set up the parameters to get a solution that is applicable in the particular case I am interested in? How do I get the results in reasonable time and in a form that I can understand? Note that all the questions connect the discovery process to me, for my human consumption."

CROSS-INDUSTRY STANDARD PROCESS

There is a temptation in some companies, due to departmental inertia and compartmentalization, to approach data mining haphazardly, to reinvent the wheel and duplicate effort. A cross-industry standard was clearly required that is industry-neutral, tool-neutral, and application-neutral. The Cross-Industry Standard Process for Data Mining (CRISP–DM) [7] was developed in 1996 by analysts representing DaimlerChrysler, SPSS, and NCR. CRISP provides a non-proprietary and freely available standard process for fitting data mining into the general problem-solving strategy of a business or research unit.

CROSS-INDUSTRY STANDARD PROCESS

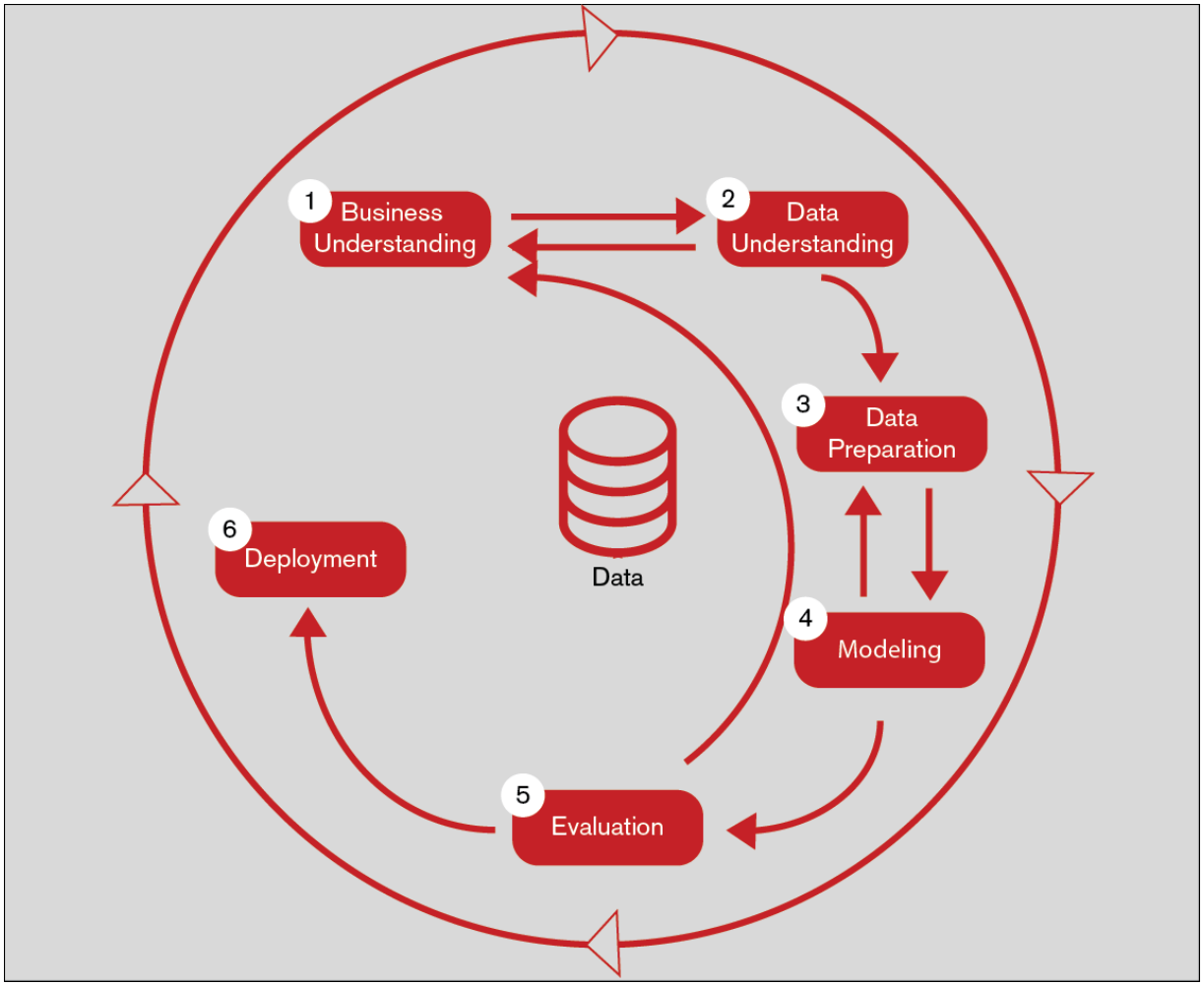

Figure 2: CRISP–DM is an iterative, adaptive process.

CROSS-INDUSTRY STANDARD PROCESS

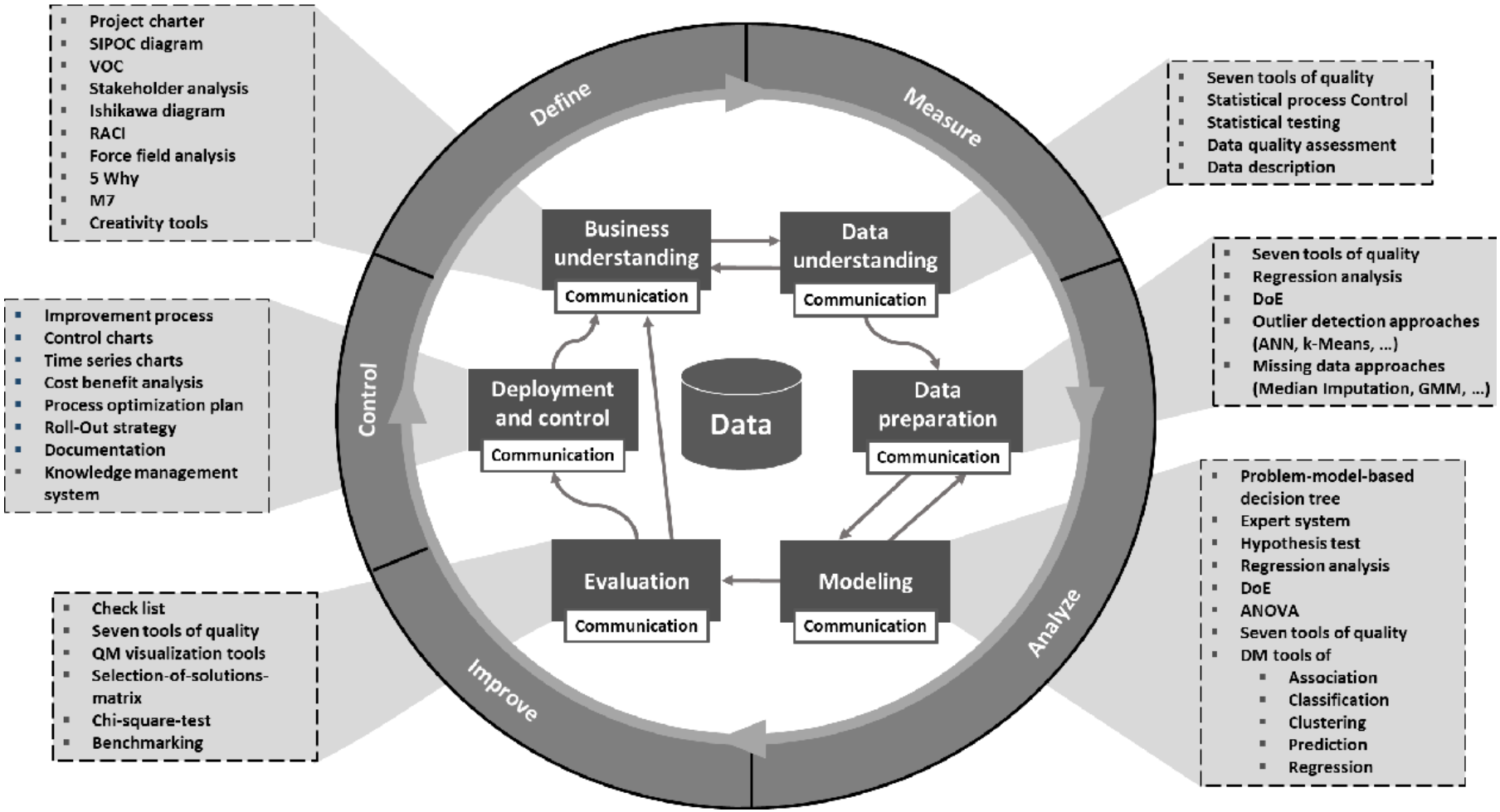

Figure 3: Breakdown of CRISP–DM

1 - Automobile Warranty Claims

ANALYZING AUTOMOBILE WARRANTY CLAIMS: EXAMPLE OF THE CRISP–DM INDUSTRY STANDARD PROCESS IN ACTION [8]

Quality assurance continues to be a priority for automobile manufacturers, including Daimler Chrysler. Jochen Hipp of the University of Tubingen, Germany, and Guido Lindner of DaimlerChrysler AG, Germany, investigated patterns in the warranty claims for DaimlerChrysler automobiles.

1. **Business Understanding Phase**
   DaimlerChrysler's objectives are to reduce costs associated with warranty claims and improve customer satisfaction. Through conversations with plant engineers, who are the technical experts in vehicle manufacturing, the researchers are able to formulate specific business problems, such as the following:
   - Are there interdependencies among warranty claims?
   - Are past warranty claims associated with similar claims in the future?
   - Is there an association between a certain type of claim and a particular garage?

   The plan is to apply appropriate data mining techniques to try to uncover these and other possible associations.

2. **Data Understanding Phase**

   The researchers make use of DaimlerChrysler's Quality Information System (QUIS), which contains information on over 7 million vehicles and is about 40 gigabytes in size. QUIS contains production details about how and where a particular vehicle was constructed, including an average of 30 or more sales codes for each vehicle. QUIS also includes warranty claim information, which the garage supplies, in the form of one of more than 5000 possible potential causes.

   The researchers stressed the fact that the database was entirely unintelligible to domain nonexperts: "So experts from different departments had to be located and consulted; in brief a task that turned out to be rather costly." They emphasize that analysts should not underestimate the importance, difficulty, and potential cost of this early phase of the data mining process, and that shortcuts here may lead to expensive reiterations of the process downstream.

3. **Data Preparation Phase**
   The researchers found that although relational, the QUIS database had limited SQL access. They needed to select the cases and variables of interest manually, and then manually derive new variables that could be used for the modeling phase. For example, the variable *number of days from selling date until first claim* had to be derived from the appropriate date attributes. They then turned to proprietary data mining software, which had been used at DaimlerChrysler on earlier projects. Here they ran into a common roadblock—that the data format requirements varied from algorithm to algorithm. The result was further exhaustive pre- processing of the data, to transform the attributes into a form usable for model algorithms. The researchers mention that the data preparation phase took much longer than they had planned.

4. **Modeling Phase**

The researchers chose to apply the following techniques: (1) Bayesian networks and (2) association rules. Bayesian networks model uncertainty by explicitly representing the conditional dependencies among various components, thus providing a graphical visualization of the dependency relationships among the components. As such, Bayesian networks represent a natural choice for modeling dependence among warranty claims.

One insight the researchers uncovered was that a particular combination of construction specifications doubles the probability of encountering an automobile electrical cable problem. DaimlerChrysler engineers have begun to investigate how this combination of factors can result in an increase in cable problems.

The researchers investigated whether certain garages had more warranty claims of a certain type than did other garages. Their association rule results showed that, indeed, the confidence levels for the rule "If garage X, then cable problem," varied considerably from garage to garage. They state that further investigation is warranted to reveal the reasons for the disparity.

5. **Evaluation Phase**
   The researchers were disappointed that the support for sequential-type association rules was relatively small, thus precluding generalization of the results, in their opinion. Overall, in fact, the researchers state: “In fact, we did not find any rule that our domain experts would judge as interesting, at least at first sight.” According to this criterion, then, the models were found to be lacking in effectiveness and to fall short of the objectives set for them in the business understanding phase. To account for this, the researchers point to the “legacy” structure of the database, for which automobile parts were categorized by garages and factories for historic or technical reasons and not designed for data mining. They suggest adapting and redesigning the database to make it more amenable to knowledge discovery.

6. **Deployment Phase**
   The researchers have identified the foregoing project as a pilot project, and as such, do not intend to deploy any large-scale models from this first iteration. After the pilot project, however, they have applied the lessons learned from this project, with the goal of integrating their methods with the existing information technology environment at DaimlerChrysler. To further support the original goal of lowering claims costs, they intend to develop an intranet offering mining capability of QUIS for all corporate employees.

2 - PREDICTING ABNORMAL STOCK MARKET RETURNS

PREDICTING ABNORMAL STOCK MARKET RETURNS USING NEURAL NETWORKS [9]

The goal of this case study is to analyze an existing dataset and eventually predicting abnormal returns using neural networks.

1. **Business/Research Understanding Phase**
   Alan M. Safer, of California State University–Long Beach, reports that stock market trades made by insiders usually have abnormal returns. Increased profits can be made by outsiders using legal insider trading information, especially by focusing on attributes such as company size and the time frame for prediction. Safer is interested in using data mining methodol- ogy to increase the ability to predict abnormal stock price returns arising from legal insider trading.

2. **Data Understanding Phase**

   Safer collected data from 343 companies, extending from January 1993 to June 1997 (the source of the data being the Securities and Exchange Commission). The stocks used in the study were all of the stocks that had insider records for the entire period and were in the S&P 600, S&P 400, or S&P 500 (small, medium, and large capitalization, respectively) as of June 1997. Of the 946 resulting stocks that met this description, Safer chose only those stocks that underwent at least two purchase orders per year, to assure a sufficient amount of transaction data for the data mining analyses. This resulted in 343 stocks being used for the study. The variables in the original data set include the company, name and rank of the insider, transaction date, stock price, number of shares traded, type of transaction (buy or sell), and number of shares held after the trade. To assess an insider's prior trading patterns, the study examined the previous 9 and 18 weeks of trading history. The prediction time frames for predicting abnormal returns were established as 3, 6, 9, and 12 months.

3. **Data Preparation Phase**
   Safer decided that the company rank of the insider would not be used as a study attribute, since other research had shown it to be of mixed predictive value for predicting abnormal stock price returns. Similarly, he omitted insiders who were uninvolved with company decisions. (Note that the present author does not necessarily agree with omitting variables prior to the modeling phase, because of earlier findings of mixed predictive value. If they are indeed of no predictive value, the models will so indicate, presumably. But if there is a chance of something interesting going on, the model should perhaps be given an opportunity to look at it. However, Safer is the domain expert in this area.)

4. **Modeling Phase**
   The data were split into a training set (80% of the data) and a validation set (20%). A neural network model was applied, which uncovered the following results:
   1. Certain industries had the most predictable abnormal stock returns, including:
      - *Industry group 36*: electronic equipment, excluding computer equipment
      - *Industry Group 28*: chemical products
      - *Industry Group 37*: transportation equipment
      - *Industry Group 73*: business services
   2. Predictions that looked further into the future (9 to 12 months) had increased ability to identify unusual insider trading variations than did predictions that had a shorter time frame (3 to 6 months).
   3. It was easier to predict abnormal stock returns from insider trading for small companies than for large companies.

5. **Evaluation Phase**

   Safer concurrently applied a multivariate adaptive regression spline (MARS, not covered here) model to the same data set. The MARS model uncovered many of the same findings as the neural network model, including results (a) and (b) from the modeling phase. Such a confluence of results is a powerful and elegant method for evaluating the quality and effectiveness of the model, analogous to getting two independent judges to concur on a decision. Data miners should strive to produce such a confluence of results whenever the opportunity arises. This is possible because often more than one data mining method may be applied appropriately to the problem at hand. If both models concur as to the results, this strengthens our confidence in the findings. If the models disagree, we should probably investigate further. Sometimes, one type of model is simply better suited to uncovering a certain type of result, but sometimes, disagreement indicates deeper problems, requiring cycling back to earlier phases.

6. **Deployment Phase**
   The publication of Safer's findings in Intelligent Data Analysis [9] constitutes one method of model deployment. Now, analysts from around the world can take advantage of his methods to track the abnormal stock price returns of insider trading and thereby help to protect the small investor.

FALLACIES OF DATA MINING

Speaking before the U.S. House of Representatives Subcommittee on Technology, Information Policy, Intergovernmental Relations, and Census, Jen Que Louie, presi- dent of Nautilus Systems, Inc., described four fallacies of data mining. Additional two are added to the existing list.

- ***Fallacy 1***. There are data mining tools that we can turn loose on our data repositories and use to find answers to our problems.
- ***Fallacy 2***. The data mining process is autonomous, requiring little or no human oversight.
- ***Fallacy 3***. Data mining pays for itself quite quickly
- ***Fallacy 4***. Data mining software packages are intuitive and easy to use.
- ***Fallacy 5***. Data mining will identify the causes of our business or research
- ***Fallacy 6***. Data mining will clean up a messy database automatically.

Tasks in Data Mining

Next, we investigate the main tasks that data mining is usually called upon to accomplish. The following list shows the most common data mining tasks.

- Description
- Estimation
- Prediction
- Classification
- Clustering
- Association

Description

Sometimes, researchers and analysts are simply trying to find ways to describe patterns and trends lying within data. For example, a pollster may uncover evidence that those who have been laid off are less likely to support the present incumbent in the presidential election. Descriptions of patterns and trends often suggest possible explanations for such patterns and trends. For example, those who are laid off are now less well off financially than before the incumbent was elected, and so would tend to prefer an alternative.

Estimation

Estimation is similar to classification except that the target variable is numerical rather than categorical. Models are built using "complete" records, which provide the value of the target variable as well as the predictors. Then, for new observations, estimates of the value of the target variable are made, based on the values of the predictors.
Examples of estimation tasks in business and research include:

- Estimating the amount of money a randomly chosen family of four will spend for back-to-school shopping this fall.
- Estimating the percentage decrease in rotary-movement sustained by a National Football League running back with a knee injury.
- Estimating the number of points per game that Patrick Ewing will score when double-teamed in the playoffs.
- Estimating the grade-point average (GPA) of a graduate student, based on that student's undergraduate GPA.

Prediction

Prediction is similar to classification and estimation, except that for prediction, the results lie in the future. Examples of prediction tasks in business and research include:

- Predicting the price of a stock three months into the future
- Predicting the percentage increase in traffic deaths next year if the speed limit is increased
- Predicting the winner of this fall's baseball World Series, based on a comparison of team statistics
- Predicting whether a particular molecule in drug discovery will lead to a profitable new drug for a pharmaceutical company

Classification

In classification, there is a target categorical variable, such as income bracket, which, for example, could be partitioned into three classes or categories: high income, middle income, and low income. The data mining model examines a large set of records, each record containing information on the target variable as well as a set of input or predictor variables.

Examples of classification tasks in business and research include:

- Determining whether a particular credit card transaction is fraudulent
- Placing a new student into a particular track with regard to special needs
- Assessing whether a mortgage application is a good or bad credit risk
- Diagnosing whether a particular disease is present
- Determining whether a will was written by the actual deceased, or fraudulently by someone else
- Identifying whether or not certain financial or personal behavior indicates a possible terrorist threat

Clustering

Clustering refers to the grouping of records, observations, or cases into classes of similar objects. A cluster is a collection of records that are similar to one another, and dissimilar to records in other clusters. Clustering differs from classification in that there is no target variable for clustering. The clustering task does not try to classify, estimate, or predict the value of a target variable. Instead, clustering algorithms seek to segment the entire data set into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized and the similarity to records outside the cluster is minimized.

Association

The association task for data mining is the job of finding which attributes "go together." Most prevalent in the business world, where it is known as affinity analysis or market basket analysis, the task of association seeks to uncover rules for quantifying the relationship between two or more attributes. Association rules are of the form "If antecedent, then consequent," together with a measure of the support and confidence associated with the rule.
Examples of association tasks in business and research include:

- Investigating the proportion of subscribers to a company's cell phone plan that respond positively to an offer of a service upgrade
- Examining the proportion of children whose parents read to them who are themselves good readers
- Predicting degradation in telecommunications networks
- Finding out which items in a supermarket are purchased together and which items are never purchased together
- Determining the proportion of cases in which a new drug will exhibit dangerous side effects

Further Reading

- Chapter 1 of Data Mining - Practical Machine Learning Tools and Techniques, Second Edition - Ian H. Witten, Eibe Frank
- Chapter 1 of DISCOVERING KNOWLEDGE IN DATA - An Introduction to Data Mining - DANIEL T. LAROSE
- Chapter 1 of Introduction to Data Mining (Second Edition) - Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar

References

[1] T. G. Group, "https://www.gartner.com/."

[2] M. Bramer, *Principles of Data Mining*, 2nd ed. Springer Publishing Company, Incorporated, 2013.

[3] K. K. Hirji, "Discovering data mining: From concept to implementation," *SIGKDD Explor. Newsl.*, vol. 1, no. 1, p. 44–45, Jun. 1999. [Online]. Available: https://doi.org/10.1145/846170.846181

[4] Y. Alsultanny, "Selecting a suitable method of data mining for successful forecasting," *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 19, 9 2011.

[5] J. Naisbitt, *Megatrends : ten new directions transforming our lives / by John Naisbitt*. Warner Books New York, 1982.

[6] M. Ankerst, "Report on the sigkdd-2002 panel the perfect data mining tool: Interactive or automated?" *SIGKDD Explor. Newsl.*, vol. 4, no. 2, p. 110–111, Dec. 2002. [Online]. Available: https://doi.org/10.1145/772862.772883

[7] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0: Step-by-step data mining guide," 2000.

[8] J. Hipp and G. Lindner, "Analysing warranty claims of automobiles," in *Internet Applications*, L. C. K. Hui and D.-L. Lee, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 31–40.

[9] A. M. Safer, "A comparison of two data mining techniques to predict abnormal stock market returns," *Intell. Data Anal.*, vol. 7, no. 1, p. 3–13, Jan. 2003.

Thank you.
Any Questions?