1 Machine learning: the problem setting

In general, a learning problem considers a set of n samples of data and try to predict properties of unknown data. If each sample is more than a single number, and for instance a multi-dimensional entry (aka multivariate data), is it said to have several variables, also known as attributes or ***features***.
We can separate learning problems in a few large categories:

- **Supervised learning**, in which the data comes with additional attributes that we want to predict.
  This problem can be either:

  Classification: samples belong to two or more classes and we want to learn from already labeled data how to predict the class of unlabeled data.
  An example of classification problem would be the digit recognition example, in which the aim is to assign each input vector to one of a finite number of discrete categories.

  Regression: if the desired output consists of one or more continuous variables, then the task is called regression.
  An example of a regression problem would be the prediction of the weight of a pony as a function of its age and height.

- **Unsupervised learning**, in which the training data consists of a set of input vectors x without any corresponding target values.
  The goal in such problems may be

  - to discover groups of similar examples within the data, where it is called ***clustering***,
  - to determine the distribution of data within the input space, known as ***density estimation***,
  - to project the data from a high-dimensional space down to two or thee dimensions for the purpose of visualization

2 Data Science

Data science incorporates varying elements and builds on techniques and theories from many fields, including math, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products.

Data science is a novel term that is often used interchangeably with competitive intelligence or business analytics, although it is becoming more common. Data science seeks to use all available and relevant data to effectively tell a story that can be easily understood by non-practitioners. Some areas of research are:

- Cloud computing
- Databases and information integration
- Learning, natural language processing and information extraction
- Computer vision
- Information retrieval and web information access
- Knowledge discovery in social and information networks

Data scientists use an extensive understanding of business, combined with technical skills and statistical knowledge, to create methods for organizations to collect and interpret data. A data scientist helps an organization determine the questions that need answers, develops the methodology and technological tools needed to collect pertinent data, and builds the statistical models needed to derive answers from the data collected. In short, these professionals have the technical skills and business understanding needed to help an organization make decisions or develop useful products and services for customers, based on the analysis of data. The exact particulars of a data scientist's job vary based on the industry in which the professional works. One data scientist might focus on the programming needed to gather specific data, while another uses existing tools in unique ways to further enhance the accuracy or effectiveness of data. Still another data scientist might combine existing tools and specially made tools to collect and analyze data in a way that helps the organization offer a new service or product. For example, data scientists help develop many of the convenience applications used on social networking websites. Data is collected about each individuals employment and educational history, then this information is compared with current affiliations. Based on each individuals history and current connections, an application makes recommendations for additional connections, possible job leads, or products and services of interest to individual members. Results typically display on the users home page or main profile screen. Using technical skills and creativity, the data scientists who developed these applications helped each website create a more useful user experience. Similar applications allow a data scientist to collect, analyze, and report information about website visitors, in-store shoppers, and other customer information. Depending on the goals of the organization, such information may be used to create custom shopping experiences or test various marketing strategies. Many websites, for example, have applications that display tailored advertisements based on customer behavior. Before launching these applications, a professional data scientist had to program a means to collect customer information, analyze it, and produce an appropriate result to display. Different industries and, in fact, different companies within the same industry, have different needs when it comes to daily tasks completed by a data scientist. While the tasks may differ, the skills needed remain the same. Professionals in this line of work need programming and other technical skills in order to develop appropriate tools to collect and manipulate data. Additionally, such professionals need creativity and critical thinking skills, as well as the ability to understand business needs, in order to know what data to collect and the different ways to interpret information.

Data Sciences

- Data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about that information. Data analytics is used in many industries to allow companies and organization to make better business decisions and in the sciences to verify or disprove existing models or theories.
- Data analytics is distinguished from data mining by the scope, purpose and focus of the analysis. Data miners sort through huge data sets using sophisticated software to identify undiscovered patterns and establish hidden relationships. Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known by the researcher.
- The science is generally divided into exploratory data analysis (EDA), where new features in the data are discovered, and confirmatory data analysis (CDA), where existing hypotheses are proven true or false. Qualitative data analysis (QDA) is used in the social sciences to draw conclusions from non-numerical data like words, photographs or video.
- In information technology, the term has a special meaning in the context of IT audits, when the controls for an organization's information systems, operations and processes are examined. Data analysis is used to determine whether the systems in place effectively protect data, operate efficiently and succeed in accomplishing an organization's overall goals.

Analytics

The term "analytics" has been used by many business intelligence (BI) software vendors as a buzzword to describe quite different functions. Data analytics is used to describe everything from online analytical processing (OLAP) to CRM analytics in call centers. Banks and credit cards companies, for instance, analyze withdrawal and spending patterns to prevent fraud or identity theft. Ecommerce companies examine Web site traffic or navigation patterns to determine which customers are more or less likely to buy a product or service based upon prior purchases or viewing trends. Modern data analytics often use information dashboards supported by real-time data streams. So-called real-time analytics involves dynamic analysis and reporting, based on data entered into a system less than one minute before the actual time of use.

Data Modeling

Data modeling is the formalization and documentation of existing processes and events that occur during application software design and development. Data modeling techniques and tools capture and translate complex system designs into easily understood representations of the data flows and processes, creating a blueprint for construction and/or re-engineering.

A data model can be thought of as a diagram or flowchart that illustrates the relationships between data. Although capturing all the possible relationships in a data model can be very time-intensive, it's an important step and shouldn't be rushed. Well-documented models allow stake-holders to identify errors and make changes before any programming code has been written.

Data modelers often use multiple models to view the same data and ensure that all processes, entities, relationships and data flows have been identified. There are several different approaches to data modeling, including:

- **Conceptual Data Modeling** - identifies the highest-level relationships between different entities.
- **Enterprise Data Modeling** - similar to conceptual data modeling, but addresses the unique requirements of a specific business.
- **Logical Data Modeling** - illustrates the specific entities, attributes and relationships involved in a business function. Serves as the basis for the creation of the physical data model.
- **Physical Data Modeling** - represents an application and database-specific implementation of a logical data model.

Predictive Modeling

- Predictive modeling is a process used in predictive analytics to create a statistical model of future behavior. Predictive analytics is the area of data mining concerned with forecasting probabilities and trends.
- A predictive model is made up of a number of predictors, which are variable factors that are likely to influence future behavior or results. In marketing, for example, a customer's gender, age, and purchase history might predict the likelihood of a future sale.
- In predictive modeling, data is collected for the relevant predictors, a statistical model is formulated, predictions are made and the model is validated (or revised) as additional data becomes available. The model may employ a simple linear equation or a complex neural network, mapped out by sophisticated software.
- Predictive modeling is used widely in information technology (IT). In spam filtering systems, for example, predictive modeling is sometimes used to identify the probability that a given message is spam. Other applications of predictive modeling include customer relationship management (CRM), capacity planning, change management, disaster recovery, security management, engineering, meteorology and city planning.

3 Machine Learning

Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it. Many researchers also think it is the best way to make progress towards human-level AI.

4 Introduction to Data Mining

Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification withvalidation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions).

Stage 1: Exploration. This stage usually starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records and - in case of data sets with large numbers of variables ("fields") - performing some preliminary feature selection operations to bring the number of variables to a manageable range (depending on the statistical methods which are being considered). Then, depending on the nature of the analytic problem, this first stage of the process of data mining may involve anywhere between a simple choice of straightforward predictors for a regression model, to elaborate exploratory analyses using a wide variety of graphical and statistical methods (see Exploratory Data Analysis (EDA)) in order to identify the most relevant variables and determine the complexity and/or the general nature of models that can be taken into account in the next stage.

Stage 2: Model building and validation. This stage involves considering various models and choosing the best one based on their predictive performance (i.e., explaining the variability in question and producing stable results across samples). This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. There are a variety of techniques developed to achieve that goal - many of which are based on so-called "competitive evaluation of models," that is, applying different models to the same data set and then comparing their performance to choose the best. These techniques - which are often considered the core of predictive data mining - include: Bagging (Voting, Averaging), Boosting, Stacking (Stacked Generalizations), and Meta-Learning.

Stage 3: Deployment. That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

The concept of Data Mining is becoming increasingly popular as a business information management tool where it is expected to reveal knowledge structures that can guide decisions in conditions of limited certainty. Recently, there has been increased interest in developing new analytic techniques specifically designed to address the issues relevant to business Data Mining (e.g., Classification Trees), but Data Mining is still based on the conceptual principles of statistics including the traditional Exploratory Data Analysis (EDA) and modeling and it shares with them both some components of its general approaches and specific techniques.

However, an important general difference in the focus and purpose between Data Mining and the traditional Exploratory Data Analysis (EDA) is that Data Mining is more oriented towards applications than the basic nature of the underlying phenomena. In other words, Data Mining is relatively less concerned with identifying the specific relations between the involved variables. For example, uncovering the nature of the underlying functions or the specific types of interactive, multivariate dependencies between variables are not the main goal of Data Mining. Instead, the focus is on producing a solution that can generate useful predictions. Therefore, Data Mining accepts among others a "black box" approach to data exploration or knowledge

discovery and uses not only the traditional Exploratory Data Analysis (EDA) techniques, but also such techniques as Neural Networks which can generate valid predictions but are not capable of identifying the specific nature of the interrelations between the variables on which the predictions are based.

Data Mining is often considered to be *"a blend of statistics, AI (artificial intelligence), and data base research"* (Pregibon, 1997, p. 8), which until very recently was not commonly recognized as a field of interest for statisticians, and was even considered by some "a dirty word in Statistics" (Pregibon, 1997, p. 8). Due to its applied importance, however, the field emerges as a rapidly growing and major area (also in statistics) where important theoretical advances are being made (see, for example, the recent annual International Conferences on Knowledge Discovery and Data Mining, co-hosted by the American Statistical Association).

For information on Data Mining techniques, review the summary topics included below. There are numerous books that review the theory and practice of data mining; the following books offer a representative sample of recent general books on data mining, representing a variety of approaches and perspectives:

1. Berry, M., J., A., & Linoff, G., S., (2000). Mastering data mining. New York: Wiley.

2. Edelstein, H., A. (1999). Introduction to data mining and knowledge discovery (3rd ed). Potomac, MD: Two Crows Corp.

3. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery & data mining. Cambridge, MA: MIT Press.

4. Han, J., Kamber, M. (2000). Data mining: Concepts and Techniques. New York: Morgan-Kaufman.

5. Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). The elements of statistical learning : Data mining, inference, and prediction. New York: Springer.

6. Pregibon, D. (1997). Data Mining. Statistical Computing and Graphics, 7, 8.

7. Weiss, S. M., & Indurkhya, N. (1997). Predictive data mining: A practical guide. New York: Morgan-Kaufman.

8. Westphal, C., Blaxton, T. (1998). Data mining solutions. New York: Wiley.

9. Witten, I. H., & Frank, E. (2000). Data mining. New York: Morgan-Kaufmann.

5 Statistical Data Mining

Statistical data mining, also known as knowledge or data discovery, is a computerized method of collecting and analyzing information. The data-mining tool takes data and categorizes the information to discover patterns or correlations that can be used in important applications, such as medicine, computer programming, business promotion, and robotic design.

Statistical data mining techniques use complex mathematics and complicated statistical processes to create an analysis.

Data mining involves five major steps.

1. The first data mining application collects statistical data and places the information in a warehouse-type program.

2. Next, the data in the warehouse is organized and creates a management system.

3. The next step creates a way to access the managed data.

4. Then, the fourth step develops software to analyze the data, also known as data mining regression,.

5. The final step facilitates using or interpreting the statistical data in a practical way.

Generally, data mining techniques integrate analytical and transaction data systems. Analytical software sorts through both types of data systems using open-ended user questions. Open-ended questions allow countless answers so programmers are not influencing the results of the sorting. Programmers create lists of questions to assist in categorizing the information using an overall focus.

Sorting is then based on developing classes and clusters of data, associations found in the data, and attempts to define patterns and trends based on the associations. For example, Google collects information on users' purchasing habits to assist in placing online advertising. Open-ended questions used to sort this buyer data focus on buying preferences or viewing habits of Internet users.

Computer scientists and programmers focus on the analysis of the statistical data that is collected. Creation of decision trees, artificial neural networks, nearest neighbor method, rule induction, data visualization, and genetic algorithms all use the statistically-mined data.

These classification systems assist in interpreting the associations discovered by the analytical data programs. Statistical data mining involves small projects that can be done on a small scale on a home computer, but most data mining association sets are so large and the data mining regression so complicated that they require a supercomputer or a network of high-speed computers.

Statistical data mining collects three general types of data, including operational data, non-operational data, and meta data. In a clothing store, operational data is basic data used to run the business, such as accounting, sales, and inventory control.

Non-operational data, which is indirectly related to the business, includes estimates of future sales and general information about the national clothing market.

Meta data concerns the data itself. A program using meta data might sort store customers into classifications based on gender or geographic location of the clothing buyers or the customers favorite color, if that data was collected.

5.1 Applications of Data Mining

A data mining application can be extremely sophisticated and the statistical data mining tool may have widespread practical applications. The study of disease outbreaks is one example. A 2000 data mining project analyzed the disease outbreak of cryptosporidium in Ontario, Canada to determine the causes of the increase in disease cases. The results of the data mining assisted in linking the bacteria outbreak to local water conditions and the lack of proper municipal water treatment. A field called "biosurveillance" uses epidemiological data mining to identify outbreaks of a single disease.

Computer programmers and designers also employ the study of probability and statistical data analysis to develop machines and computer programs. The Google Internet search engine was designed using statistical data mining. Google continues to collect and use data mining to create program updates and applications.