This article appeared on December 2, 2004, in "CRM Zine" (Vol. 48), published by CRM Today (#1 CRM Resource – www.crm2day.com).

A Statistical Perspective of Data Mining

Dr. Diego Kuonen

Statoo Consulting, PO Box 107, 1015 Lausanne 15, Switzerland

kuonen@statoo.com

The field of data mining, like statistics, concerns itself with "learning from data" or "turning data into information". In this article we will look at the connection between data mining and statistics, and ask ourselves whether data mining is "statistical déjà vu".

What is statistics and why is statistics needed?

Statistics is the science of learning from data. It includes everything from planning for the collection of data and subsequent data management to end-of-the-line activities such as drawing inferences from numerical facts called data and presentation of results. Statistics is concerned with one of the most basic of human needs: the need to find out more about the world and how it operates in face of variation and uncertainty. Because of the increasing use of statistics, it has become very important to understand and practise statistical thinking. Or, in the words of H. G. Wells: "*Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write*".

But, why is statistics needed? Knowledge is what we know. Information is the communication of knowledge. Data are known to be crude information and not knowledge by themselves. The sequence from data to knowledge is as follows: from data to information (data become information when they become relevant to the decision problem); from information to facts (information becomes facts when the data can support it); and finally, from facts to knowledge (facts become knowledge when they are used in the successful completion of the decision process). Figure 1 illustrates this statistical thinking process based on data in constructing statistical models for decision making under uncertainties. That is why we need statistics. Statistics arose from the need to place knowledge on a systematic evidence base. This required a study of the laws of probability, the development of measures of data properties and relationships, and so on.

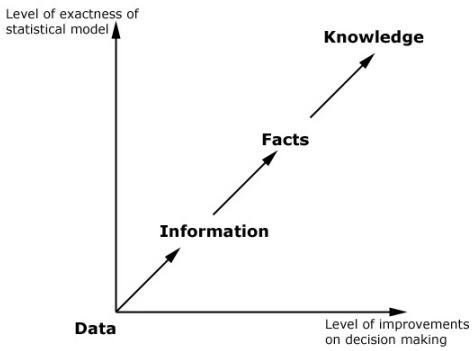


Figure 1. *The statistical thinking process based on data in constructing statistical models for decision making under uncertainties.*

What is data mining?

Data mining has been defined in almost as many ways as there are authors who have written about it. Because it sits at the interface between statistics, computer science, artificial intelligence, machine learning, database management and data visualization (to name some of the fields), the definition changes with the perspective of the user:

> "*Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.*" (M. J. A. Berry and G. S. Linoff)

> "*Data mining is finding interesting structure (patterns, statistical models, relationships) in databases.*" (U. Fayyad, S. Chaudhuri and P. Bradley)

> "*Data mining is the application of statistics in the form of exploratory data analysis and predictive models to reveal patterns and trends in very large data sets.*" ("Insightful Miner 3.0 User Guide")

We think of data mining as the process of identifying valid, novel, potentially useful, and ultimately comprehensible understandable patterns or models in data to make crucial business decisions. "Valid" means that the patterns hold in general, "novel" that we did not know the pattern beforehand, and "understandable" means that we can interpret and comprehend the patterns. Hence, like statistics, data mining is not only modelling and prediction, nor a product that can be bought, but a whole problem solving cycle/process that must be mastered through team effort.

Defining the right business problem is the trickiest part of successful data mining because it is exclusively a communication problem. The technical people analyzing data need to understand what the business really needs. Even the most advanced algorithms cannot figure out what is most important. Never forget that "garbage in" yields "garbage out". Data preprocessing or data cleaning or data preparation is also a key part of data mining. Quality decisions and quality mining results come from quality data. Data are always dirty and are not ready for data mining in the real world. For example,

- data need to be integrated from different sources;
- data contain missing values. *i.e.* incomplete data;
- data are noisy, *i.e.* contain outliers or errors, and inconsistent values (*i.e.* contain discrepancies in codes or names);
- data are not at the right level of aggregation.

The main part of data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. It is the computer which is responsible for finding the patterns by identifying the underlying rules and features in the data. The choice of a particular combination of techniques to apply in a particular situation depends on both the nature of the data mining task to be accomplished and the nature of the available data. The idea is that it is possible to strike gold in unexpected places as the data mining software extracts patterns not previously discernible or so obvious that no-one has noticed them before. The analysis process starts with a set of data, uses a methodology to develop an optimal representation of the structure of the data during which time knowledge is acquired. Once knowledge has been acquired this can be extended to larger sets of data working on the assumption that the larger data set has a structure similar to the sample data. This is analogous to a mining operation where large amounts of low grade materials are sifted through in order to find something of value.

This sounds familiar, doesn't it? First, recall that we defined statistics as the science of learning from data. Second, remember that the main sequence from data to knowledge is: from data to information, and from information to knowledge. Let us briefly illustrate this sequence. Data are what we can capture and store (*e.g.* customer data, store data, demographical data, geographical data), and become information when they become relevant to our decision problem. Information relates items of data (*e.g.* X lives in Z; S is Y years old; X and S moved; W has money in Z), and becomes knowledge when it is used in the successful completion of the decision process. Hence knowledge relates items of information (*e.g.* a quantity Q of product A is used in region Z; customers of class L use N% of C during period D). The latter is indeed a fragment of the so-called "business intelligence" chain: from

data to information, from information to knowledge, from knowledge to decision, and from decision to action (*e.g.* decisions: promote product A in region Z; mail ads to families of profile P; cross-sell service B to clients E). As we see, the main problem is to know how to get from data to knowledge, or, as J. Naisbitt said: "*We are drowning in information but starved for knowledge*". The remedy to this problem is data mining and/or statistics. With data mining, companies can analyze customers' past behaviours in order to make strategic decisions for the future. Keep in mind, however, that the data mining techniques and tools are equally applicable in fields ranging from law enforcement to radio astronomy, medicine and industrial process control (to name some of the fields).

Why data mining?

Data mining got its start in what is now known as "customer relationship management" (CRM). It is widely recognized that companies of all sizes need to learn to emulate what small, service-oriented businesses have always done well – creating one-to-one relationships with their customers. In every industry, forward-looking companies are trying to move towards the one-to-one ideal of understanding each customer individually and to use that understanding to make it easier for the customer to do business with them rather than with a competitor. These same companies are learning to look at the lifetime value of each customer so they know which ones are worth investing money and effort to hold on to and which ones to let drop.

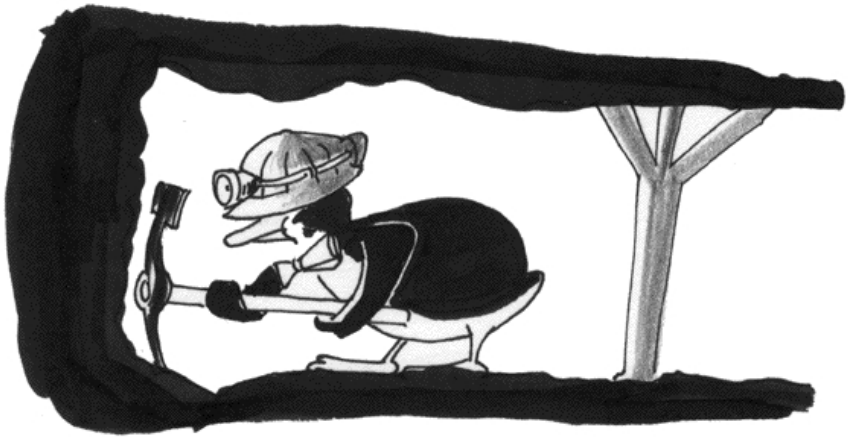
As noted, a small business builds one-to-one relationships with its customers by noticing their needs, remembering their preferences, and learning from past interactions how to serve them better in the future. How can a large enterprise accomplish something similar when most customers may never interact personally with company employees? What can replace the creative intuition of the sole proprietor who recognizes customers by name, face, and voice, and remembers their habits and preferences? In a word: nothing. But through the clever application of information technology, even the largest enterprise can come surprisingly close. In large commercial enterprises, the first step - noticing what the customer does - has already largely been automated. On-line transaction processing (OLTP) systems are everywhere, collecting data on seemingly everything. These days, we all go through life generating a constant stream of transaction records.

The customer-focused enterprise regards every record of an interaction with a client or prospect as a learning opportunity. But, learning requires more than simply gathering data. In fact, many companies gather hundreds of gigabytes or terabytes of data from and about their customers without learning anything. Data is gathered because it is needed for some operational purpose, *e.g.* inventory control or billing. And, once it has served that purpose, it languishes on tape or gets discarded. For learning to take place, data from many sources must first be gathered together and organized in a consistent and useful way. This is called data warehousing.

Hence data warehousing allows the enterprise to remember what it has noticed about its customers. *Data warehousing provides the enterprise with a memory*. But, memory is of little use without intelligence. That is where data mining comes in. Intelligence allows us to comb through our memories noticing patterns, devising rules, coming up with new ideas to try, and making predictions about the future. The data must be analyzed, understood, and turned into actionable information. Data mining provides tools and techniques that add intelligence to the data warehouse. *Data mining provides the enterprise with intelligence*. Using several data mining tools and techniques that add intelligence to the data warehouse, an enterprise will be able to exploit the vast mountains of data generated by interactions with its customers and prospects in order to get to know them better.

- What customers are most likely to respond to a mailing?
- Are there groups (or segments) of customers with similar characteristics or behaviour?
- Are there interesting relationships between customer characteristics?
- Who is likely to remain a loyal customer and who is likely to jump ship?
- What is the next product or service this customer will want?

Answers to such questions lie buried in the enterprise's corporate data, but it takes powerful data mining tools to get at them, *i.e.* to dig user info for gold.



The main data mining tasks

Let us define the main tasks well-suited for data mining, all of which involve extracting meaningful new information from the data. Knowledge discovery (learning from data) comes in two flavours: directed (supervised) and undirected (unsupervised) learning from data. The six main activities of data mining are:

- *classification* (examining the feature of a newly presented object and assigning it to one of a predefined set of classes);
- *estimation* (given some input data, coming up with a value for some unknown continuous variable such as income, height, or credit-card balance);
- *prediction* (the same as classification and estimation except that the records are classified according to some predicted future behaviour or estimated future value);
- *affinity grouping* or *association rules* (determine which things go together, also known as dependency modelling, *e.g.* in a shopping cart at the supermarket - market basket analysis);
- *clustering* (segmenting a population into a number of subgroups or clusters); and
- *description and visualization* (exploratory or visual data mining).

The first three tasks – classification, estimation and prediction – are all examples of directed knowledge discovery (supervised learning). In supervised learning the goal is to use the available data to build a model that describes one particular variable of interest, such as income or response, in terms of the rest of the available data ("class prediction"). The next three tasks – affinity grouping or association rules, clustering, and description and visualization – are examples of undirected knowledge discovery (unsupervised learning). In unsupervised learning no variable is singled out as the target; the goal is to establish some relationship among all the variables ("class discovery"). Unsupervised learning attempts to find patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes. This is similar to looking for needles in haystacks.

In fact, hardly any of the data mining algorithms were first invented with commercial applications in mind. Although most of the data mining techniques have existed, at least as academic algorithms, for years or decades, it is only in the last several years that commercial data mining has caught on in a big way. This is due to the convergence in the 1990s of a number of factors: the data are being produced; the data are being warehoused; the computing power is affordable; the competitive

pressure is strong; and commercial data mining software products have become available. The commercial data miner employs a grab bag of techniques borrowed from statistics, computer science and artificial intelligence research. Moreover, no single data mining tool or technique is equally applicable to all the tasks. The choice of a particular combination of data mining techniques to apply in a particular situation depends on both the nature of the data mining task to be accomplished and the nature of the available data. From a statistical perspective, many data mining tools could be described as flexible models and methods for exploratory data analysis. In other words many data mining tools are nothing else than multivariate (statistical) data analysis methods. Or, in the words of I. H. Witten and E. Franke: "*What's the difference between machine learning and statistics? Cynics, looking wryly at the explosion of commercial interest (and hype) in this area, equate data mining to statistics plus marketing*".

Data mining myths versus realities

Do not let contradictory claims about data mining keep you from improving your business. A great deal of what is said about data mining is incomplete, exaggerated, or wrong. Data mining has taken the business world by storm, but as with many new technologies, there seems to be a direct relationship between its potential benefits and the quantity of (often) contradictory claims, or myths, about its capabilities and weaknesses. When you undertake a data mining project, avoid a cycle of unrealistic expectations followed by disappointment. Understand the facts instead and your data mining efforts will be successful. Simply put, data mining is used to discover patterns and relationships in your data in order to help you make better business decisions. Data mining cannot be ignored – the data are there, the methods are numerous and the advantages that knowledge discovery brings to a business are tremendous. Companies whose data mining efforts are guided by "mythology" will find themselves at a serious competitive disadvantage to those organizations taking a measured, rational approach based on facts. Finally, let me cite A. Onassis, "*The secret of success is to know something that nobody else knows*", and J. Bigus, "*If you are not mining your data for all it is worth, you are guilty of underuse of one of your company's greatest assets*".

Conclusion: challenges for data miners, statisticians and clients

The field of data mining, like statistics, concerns itself with "learning from data" or "turning data into information". So we asked ourselves whether data mining is "statistical déjà vu". As seen, answering "yes" to the latter would be absurd. Rather, it is important to note that data mining can learn from statistics – that, to a large extent, statistics is fundamental to what data mining is really trying to achieve. There is the opportunity for an immensely rewarding synergy between data miners and statisticians. However, most data miners tend to be ignorant of statistics and client's domain; statisticians tend to be ignorant of data mining and client's domain; and clients tend to be ignorant of data mining and statistics. Unfortunately, they also tend to be inhibited by myopic points of view: computer scientists focus upon database manipulations and processing algorithms; statisticians focus upon identifying and handling uncertainties; and clients focus upon integrating knowledge into the knowledge domain. Moreover, most data miners and statisticians continue to sarcastically criticise each other. This is detrimental to both disciplines. Unfortunately, the anti-statistical attitude will keep data mining from reaching its actual potential – data mining can learn from statistics. Data mining and statistics will inevitably grow toward each other in the near future because data mining will not become knowledge discovery without statistical thinking, statistics will not be able to succeed on massive and complex datasets without data mining approaches. Remember that knowledge discovery rests on the three balanced legs of computer science, statistics and client knowledge: it will not stand either on one leg or on two legs, or even on three unbalanced legs. Hence, successful knowledge discovery needs a substantial to collaboration from all three. All parties should widen their focus until true collaboration and the mining for gold becomes reality. A maturity challenge is for data miners, statisticians and clients to recognize their dependence on each other and for all of them to widen their focus until true collaboration becomes reality. The critical challenge for us all is to view the challenges as opportunities for our joint success. All parties should widen their focus until true collaboration and the mining for gold becomes reality.

About the author

Diego Kuonen, PhD in Statistics, is founder and CEO of Statoo Consulting, Lausanne, Switzerland. Statoo Consulting is a vendor-independent Swiss consulting firm specialized in statistical consulting and training, data analysis, data mining, analytical CRM and bioinformatics services. Dr. Diego Kuonen has several years of experience in statistical consulting, in computing and in data mining, and also in teaching and training. Currently, he is also vice president of the “Swiss Statistical Society” and president of its Section “Statistics in Business and Industry”.

Have you already been Statooed? If not, please find further information on how to get Statooed at www.statoo.info/.