Where the Gold is: Data Mining and Improving the Curriculum in the 21st Century

Deborah J. Gougeon, BS, MS, PhD
Associate Professor of Statistics
Faculty of The Kania School of Management, University of Scranton, Scranton, Pennsylvania
gougeond1@scranton.edu

Abstract

With the proliferation of computers and computer technology in the last decade, decision makers in business and government have been inundated with massive amounts of data. Whether you shop at a supermarket where your items are scanned into a database, or mouse click on a web site that you are browsing, or use your credit card to purchase an item, or make a phone call to a friend overseas, more data are continually being collected. At a time when companies are attempting to determine why their customer base is rising or falling, or what new and promising drug will be effective in treating an illness, and when government intelligence agencies are trying to determine what combination of phone calls, e-mails, and foreign travel might signal terrorist activity, it seems imperative that we educate our college students regarding the statistical tools that are used in making these decisions. This paper focuses on the development of an undergraduate course in Data Mining, also known as Knowledge Discovery in Data (KDD), Exploratory Data Analysis (EDA), and Business Intelligence. Several technical procedures will be discussed that are used to summarize and interpret data, identify patterns and trends, and assist one in making the best decision.

Introduction

With the proliferation of computer technology in the last decade, decision makers in business and government have been inundated with massive amounts of data. When you shop at a supermarket where your items are scanned into a database, or you click on a web site and browse, or use your credit card to purchase an item, or make a phone call to a friend overseas, you contribute to an ever-growing mass of raw data. At a time when companies are attempting to determine why their customer base is rising or falling or what new and promising drug will be effective in treating an illness, and when government intelligence agencies are trying to determine what combination of phone calls, e-mails, and foreign travel might signal terrorist activity, it is imperative that we educate our college students regarding the statistical tools that are used in making these decisions. This paper provides a model for the development of an undergraduate elective course in Data Mining, (also known as Knowledge Discovery in Data [KDD], Exploratory Data Analysis [EDA], Data Pattern Processing, Knowledge Extraction, Data Archaeology, Information Discovery, or Business Intelligence). Data Mining is basically a component of the Knowledge Discovery Process which includes the following steps: 1) identifying the business problem, 2) performing the actual Data Mining, 3) evaluation and measurement, and 4) eventual deployment and integration of results that meet the needs of business and government. Most of the applications of Data Mining considered in this course deal with marketing and sales. Several of the technical procedures discussed are used to summarize and interpret data, identify patterns and trends, and assist in decision making. This course includes discussion of descriptive models dealing with summarization and various features of the data and predictive models covering the behavior of one variable by knowledge of another.

This course begins with a presentation of the history of Data Mining and how it evolved as a confluence of multiple disciplines such as database technology, information science, statistics, machine learning, and visualization. Included are a Data Mining timeline, which begins with

Bayes' Theorem in the late 1700s, regression analysis in the early 1900s, neural networks in the early 1940s, nearest neighbor in the early 1950s, decision trees in the mid 1960s, k-means clustering in the late 1970s, and OLAP (On-Line Analytical Processing) in the 1990s. Also included in this course are applications for many information repositories such as The World Wide Web, flat files, relational databases (a collection of tables with a unique name), data warehouses (organized around major subjects and collected from multiple sources), transactional databases (also known as "market basket analysis" consisting of files where each record represents a transaction), and advanced database systems handling spatial data and multimedia data (also referred to as database management systems [DBMS]).

One of the most common applications of Data Mining is in market basket analysis. This procedure uses association techniques to determine which products the consumer will purchase at a given time. The analysis is based on, "If a customer bought x, he would likely buy y, etc." One might think that someone who purchases beer at a supermarket will also purchase potato chips, but Data Mining techniques have shown that on Thursdays customers often purchase beer and diapers together. These are, apparently, young families who might want to just "stock up" for the weekend. Such information can be very helpful in determining store layout, offering coupons for matching products and, ultimately, stimulating sales. In addition, multi-dimensional data models are introduced along with visuals such as *data cubes*, *star schema*, *snowflake schema*, and *galaxy schema* (also referred to as a *fact constellation)*. The *data cube* allows the data to be viewed in multiple dimensions. For business, the data cube displays the multi-dimensionality of their records according to factors such as sales of a computer, computer printer, blackberry, or i-pod, their costs, and a certain season or quarter of the year. The *star schema* is a common visual resembling a starburst that consists of a fact table that includes most of the data and then smaller tables for each of the dimensions. The *snowflake schema* is essentially a star schema with additional splitting of the dimension tables through normalization. Redundancies are more likely to be reduced using the schema. More sophisticated applications use *fact constellations* or *galaxy schemas* where dimensional tables are shared. It should be noted that what works well in one-dimension does not necessarily work as well in multi-dimensions. Ironically, there is an exponential increase in the data needed for multi-dimensional analysis to maintain a specific level of accuracy. This is often referred to as the "curse of dimensionality." In addition, the concepts of *roll-up* (as an example, going from a city to a county when considering location), *drill-down* (as an example, going from a county to a city to a street), *slice and dice* (viewing a certain slice of a data cube), and *pivot* (rotating data on an axis to observe a different presentation) are also discussed in detail.

A statistical perspective on Data Mining is also being provided in this course. Two courses in Business Statistics are a prerequisite. Such topics as measures of central tendency including the mean, median, and mode; measures of dispersion including the range, average absolute deviation, standard deviation, variance, and interquartile range; and graphic representations such as histograms and scatter plots are covered. In addition, z-scores, skewness, and kurtosis are addressed. The concepts of confidence intervals and hypothesis testing, regression and correlation analysis, and chi-square analysis are thoroughly discussed. The concept of visualization will also be considered as an important tool in viewing the behavior of the data in order to identify key relationships and trends. Some of the visualization techniques included are graphical (frequency polygons and histograms), geometric (box plots and scatter diagrams), icon based (figures, colors, icons), pixel-based (each value seen as a color pixel), and hierarchical (dividing a display screen into regions). Several lectures are devoted to data preprocessing since the data may be inconsistent, that

is, containing discrepancies in the department that categorize them, or incomplete, lacking attributes of interest. Data selection and data cleaning are discussed as well. Data cleaning covers processes for removing noisy data that include outliers that deviate from what is expected. Some solutions such as regression (smoothing the data to a function), binning (consulting neighbors using means and medians), and clustering (similar values organized into cluster) are covered. The cleaning process may be somewhat confused or have a tendency to slow down because of large sets of redundant data. Data reduction is also covered which looks at smaller but representative groups of the larger data set. Some of the topics covered in reduction include data discretization (automatic generation of concept hierarchies), data aggregation (data cube building), generalization, (low-level concepts are replaced with higher level concept), and dimensionality reduction.

Two Data Mining approaches are addressed in this course. (1) *Supervised Methods*, where the response variable in the study guides the group. Classification and regression examples are used to illustrate this approach. (2) *Unsupervised Methods*, where no direction is given to guiding the groups in the study. Clustering techniques are used as an example for this approach. Clustering is a Data Mining technique where the data are grouped into clusters or related observations. This method has the advantage of being very flexible and can organize the data hierarchically or non-hierarchically. A disadvantage and possible limitation used in clustering is the time consumption usually involved with large sets of data. This procedure may also require repeated examination and interpretation of results.

Additional Data Mining techniques that are addressed in this course include *k-means clustering*, *association*, *decision trees*, *prediction*, and *neural networks*. *K-means clustering* groups data in a top-down approach where the observations are assigned to a single group. *Association* rules are often used in Data Mining retail transactions (market basket analysis) since they help to determine the links or associations among the transactions. These rules are unsupervised and easy to interpret but time-consuming. *Decision trees*, on the other hand, represent a supervised method often done by hand. They are easy to understand but often expensive with large sets of data. These *decision trees* consist of nodes (a point in the tree), a parent node (larger set divided into two or more smaller groups), the child node (stemming from a parent node), and, lastly, a leaf node (no further divisions). *Prediction* is also discussed through coverage of regression analysis and the development of mathematical models. *Neural networks* are used to represent a series of independent nodes or processors. Input, output, and hidden layers are covered as well. Several software packages for Data Mining, such as Clementine, CART, Intelligent Miner, Enterprise Miner, MineSet, and DB Miner, are discussed and evaluated.

In summary, it is important to remember that Data Mining is a relatively young, but extremely important and evolving discipline. It is somewhat daunting to think of the amount of data being collected each and every day on virtually every individual in our society. Whether it is what we purchase at the supermarket, what we pay on our mortgage, what banks and financial institutions we deal with, what credit cards we use on what items and where, how many phone calls we make and to whom and where, or what WEB surfing we do, all this information, and more, is being collected. As this discipline evolves, more and more individuals, companies, institutions, and government agencies will use Data Mining for their own purposes. It has often been said that we are "data rich, but information poor." It is, therefore, imperative in the 21st century that we expose our college and university students to this very important topic in order that we, as a productive society, may become "information rich" and make the best possible decisions collectively as well as individually.