A FIRST COURSE IN DATA MINING: SUSTAINING STATISTICAL EDUCATION IN THE MODERN BUSINESS CURRICULUM

Deborah J. Gougeon
University of Scranton
gougeond1@scranton.edu

When National Security Agency contractor Edward Snowdon allegedly leaked classified documents that detail how the U.S. government uses the technique to track terrorists, this was the first time that most people heard of data mining (Pappalardo, 2013). Every business, agency, company, etc. benefits from collecting and analyzing data. With the proliferation of computers, data is easier and less expensive to collect and store. Although statistical education in the undergraduate curricula of most business schools is commonly limited to one or two semesters of Business Statistics, there is an ever-growing need to sustain and further develop students' skills in this important area. Whether trying to determine why a company's customer base is rising or falling, or analyzing buying habits of customers at a supermarket where items are scanned, or using a credit card to process a transaction on-line, or predicting what size clothes should go to what stores, more sophisticated statistical skills are required to analyze massive sets of data. 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), or Business Intelligence. Subjects include procedures that are used to summarize and interpret data, identify patterns and trends, and assist students in making the best possible decision from a business perspective.

A variety of definitions exist for Data Mining. One 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" (Hand et al, 2001). Another definition is "an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases" (Evangelos Simoudis in Cabena et al, 1998). Data Mining is basically a component of the Knowledge Discovery Process. As one expert puts it, "Merely finding patterns is not enough. You must respond to the patterns and act on them, ultimately turning data into information, information into action, and action into value. This is the virtuous cycle of data mining in a nutshell" (Berry & Linoff, 2004).

A statistical perspective on Data Mining is also provided in this course. Two courses in Business Statistics are a prerequisite. Topics covered include 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. 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.

REFERENCES

Berry, Michael and Gordon Linoff, (2004) *Data Mining Techniques*, 2nd Edition, Wiley Publishing,Inc.

Berry, Michael and Gordon Linoff, (2011) *Data Mining Techniques*, 3rd Edition, Wiley Publishing,Inc.

Cabena, Peter, Pablo Hadjinian, Rolf Stadler, Joap Verhees, and Alessandro Zanasi, (1998) *Discovering Data Mining: From Concept to Implementation*, Prentice Hall.

Hand, David, Heikki Mannila, and Padhraic Smyth, (2001) Principles of Data Mining, MIT Press.

Pappalardo, Joseph, (2013) *NSA Data Mining: How It Works*, Popular Mechanics.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org