A Data Mining Applications Area in the Department of Mathematics

Patricia B. Cerrito
Department of Mathematics
Jewish Hospital Center for Advanced Medicine
pcerrito@louisville.edu

Data Mining

Data Mining represents a process developed to examine large amounts of data routinely collected. The term also refers to a collection of tools used to perform the process. Data mining is used in most areas where data are collected-marketing, health, communications, etc.

For example, retail stores routinely use data mining tools to learn about purchasing habits of its customers.

Amazon.com uses data mining to provide customers with purchase suggestions:

Customers who bought this book also bought:
Seven Methods for Transforming Corporate Data Into Business Intelligence by Vasant Dhar, Roger Stein
Building Data Mining Applications for CRM by Alex Berson, et al
Data Preparation for Data Mining by Dorian Pyle *Kellogg on Integrated Marketing* by Dawn Iacobucci (Editor), et al *Multivariate Data Analysis (5th Edition)* by Joseph F. Hair (Editor), et al
Explore similar items

The use of association rules has increased sales by 15%. SAS, Inc. developed the process for Amazon.com. Banks and the Federal Reserve use data mining to investigate the flow of money. Federal agencies use data mining to monitor cell phone communications via satellite. Compaq uses data mining to examine calls made to customer service to find patterns of complaints.

Data Mining Courses

- **Math 566 Nonparametric Statistical Methods.** Rank tests for comparing two or more treatments or attributes, the one-sample problem, tests of randomness and independence, nonparametric estimation, graphic methods, and computer programs.
- **Math 665 Advanced Linear Statistical Models.** Distribution of quadratic forms, estimation and hypothesis testing in the general linear model, special linear models, applications.
- **Math 667 Methods of Classification.** Classification methods used in the industry to handle large databases.Logistic regression, structural equation modeling, multivariate analysis, data mining.
- **CECS 535 Introduction to Databases.** Prerequisites: CECS 335. Introduction to database management systems, covering the field: database design, SQL, query processing and optimization, transactions. The emphasis will be placed on engineering design and implementation of relational systems.
- **CECS 632 Data Mining.** Prerequisites: IE 360, CECS 535. Data mining concepts, methodologies, and techniques, including statistical and fuzzy inference, cluster analysis, artificial neural networks and genetic algorithms, rule association and decision trees, N-dimensional visualization, Web and text mining, and advanced topics.
- **CECS 630 Advanced Databases and Data Warehousing.** Prerequisite: CECS 535. Object-relational databases, extended relational databases, and semi-structured data. Design, query languages, query processing, and optimization in data warehousing. Integration of heterogeneous data.

Other elective courses are also available. The data mining certificate program represents the first year of a 2-year master's program. However, any student who has the needed prerequisite material either from undergraduate courses or from work experience is eligible to register at the University of Louisville with non-degree status to complete the program.

Use of Statistical Software

Spreadsheets have some statistical tools but are extremely limited and should not be used as statistical packages. Small statistical packages can be purchased for use on 1 desktop at cost < $500. Their use is limited and can only perform relatively simple, routine statistical methods.

The two main statistical packages used in research are SPSS and SAS. SPSS was developed for use with the social sciences. It still sells primarily to an academic market. SAS now incorporates all the social science methodology of SPSS + database management. It was developed for use in the natural sciences. It is now the primary package used in the business market. **The ability to use SAS has now become a very marketable skill.** Businesses advertise for SAS experience. For example, FDA statistical guidelines were in part derived with SAS language. Pharmaceutical companies all use SAS as the statistical software package. Other statistical packages (such as S-plus) might be used as add-ons to SAS, but SAS remains the primary package.

One of the reasons that SAS remains so popular is that each year in early April, the SAS User's Group International holds an annual conference attracting approximately 4000 attendees. SAS developers interact with SAS users (mostly from the business environment) to improve the product. **Most of the innovations in using statistical methods are now coming from the business world rather than the academic world.**

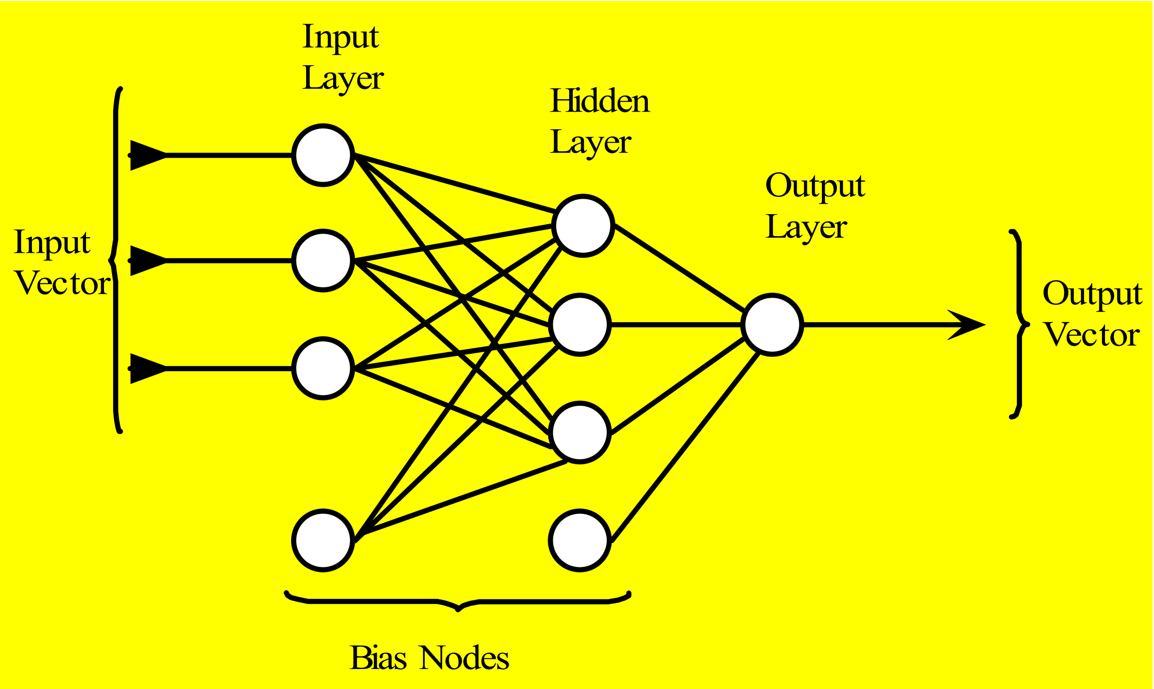
Data Mining Software

There are a number of data mining software packages, including Intelligent Miner by IBM. However, for good data mining software combined with good statistical software, there are two: Clementine by SPSS and Enterprise Miner by SAS. Enterprise Miner as the best integration of statistics with data mining. It contains methods that SPSS does not include: link analysis, text mining, memory-based reasoning. It is the data mining used in the business world.

Data Mining Techniques

- Artificial neural networks
- Rule induction
- Logistic regression
- Association rules
- Data visualization

Artificial Neural Networks are used to classify the observations into categories. The final result is a "black box" in that there is no statistical function that defines the classification rule.



Text Mining

Text mining can be used to investigate coded information Inventories generally are identified by codes. To investigate customer choices, the standard statistical tool is to identify each inventory code as a separate category. The result is too many categories. In healthcare, patient diagnoses are also categorized by codes that can be analyzed using text mining.

Text mining uses the process of "stemming". Words that have similar root stems are considered the same. Related inventory items generally have the same code stem. Therefore, similar items can be grouped using text mining to reduce the number of categories to a manageable amount.

Research Application: To investigate a database of patients undergoing open heart surgery.

Of the 122 patients in the pharmacy order database taking medications for diabetes, 8 (7%) are not coded in the clinical database as having diabetes. Conversely, there are 349 of 1459 (23%) patients listed in the clinical database as having diabetes without any order for diabetes medication. The 8 indicate under-reporting of patients with diabetes using manual extraction from patient records. The 349 either have their diabetes under control through diet, or they are not currently undergoing treatment for diabetes.

Suggested Master's Degree

Course	Description
Math 566 Nonparametric Statistics	Rank tests for comparing two or more treatments or attributes, the one-sample problem, tests of randomness and independence, nonparametric estimation, graphic methods, and computer programs.
Math 661 Probability Theory	A measure-theoretic approach to topics in probability theory; conditional probability, conditioned expectation, types of convergence, strong law of large numbers, characteristic functions, and the central limit theorem.
Math 662 Advanced Mathematical Statistics	Classical theory of statistical inference, asymptotic theory and robustness, Bayesian inference, and statistical decision theory.
Math 665 Advanced Linear Statistical Models	Distribution of quadratic forms, estimation and hypothesis testing in the general linear model, special linear models, applications.
Math 667 Methods of Classification	Classification methods used in the industry to handle large databases. Logistic regression, structural equation modeling, multivariate analysis, and data mining.
Math 635 Modeling Theory I	Introduction to mathematical modeling including optimization and linear dynamical systems, both discrete and continuous. Topics will include models from genetics, population studies and "battle" problems.
Math 636 Modeling Theory II	Continuation of MATH 635 with discrete and continuous nonlinear models. Topics will include Poincaré theory, chaotic models and elementary catastrophes.
CECS 535 Introduction to Databases.	Introduction to database management systems, covering the basic issues in the field: database design, SQL, query processing and optimization, transactions. The emphasis will be placed on engineering design and implementation of relational systems. A written project report is required.
CECS 632 Data Mining.	Data mining concepts, methodologies, and techniques, including statistical and fuzzy inference, cluster analysis, artificial neural networks and genetic algorithms, rule association and decision trees, N-dimensional visualization, Web and text mining, and advanced topics.
CECS 545 Artificial Intelligence.	Topics covered will include rationale and use of heuristic approach to engineering problem solving; information processing models as an explanation of human perceptual, cognitive, and affective behaviors. Applications involving the concepts and problems in artificial intelligence engineering.
Math 695	Graduate Thesis

For the PhD program, the following additional courses (after the Master's list) are recommended:

Course	Description
Math 567 Sampling Theory	Random, systematic, stratified, and cluster sampling techniques. Ratio and proportion estimates. Sample size and strata determination.
Math 601 Real Analysis I	Basic set theory and real topology, Lebesgue measure and integration on the real line, differentiation of integrals, L(p) spaces.
Math 602 Real Analysis II	Elementary Halberd space theory, abstract measure spaces and integration, product spaces. Applications to other areas.
Math 681 Combinatorics and Graph Theory I	Fundamental topics in Graph Theory and Combinatorics through Ramsey theory and Polya's theorem respectively. Motivation will be through appropriate applications.
Math 682 Combinatorics and Graph Theory II	Fundamental topics in Graph Theory and Combinatorics through Ramsey theory and Polya's theorem respectively. Motivation will be through appropriate applications.
CECS 619 Design and Analysis of Computer Algorithms.	The engineering design of efficient computer algorithms. A study of the interrelationships between algorithmic statements, data structures, and the resulting computational complexity of the algorithms. An engineering analysis of the effect of the computer implementation of the algorithmic statement on the computational complexity. Categorization of algorithms into complexity classes.
CECS 630 Advanced Databases and Data Warehousing.	Object-relational databases, extended relational databases, and semi-structured data. Design, query languages, query processing, and optimization in data warehousing. Integration of heterogeneous data.

Joint BS/MA with a concentration in data mining:

Course	Description	Number of Hours
Core Requirements	Math 205, 206, 301, 311, 325, 405, 501 or 521	24
BS Option Probability and Statistics Requirements	Math 560, 561, 562, 564	12
Computer Requirements	CECS 121, 230	6
Mathematics Electives	Math 502, 566	6
Computer Science Electives (Applications Area)	CECS 303, 335, 535, 545, 632	15

With the above courses, students can complete the proposed certificate in data mining. The following schedule is suggested for completing the requirements:

Course	Description	Number of Hours
Year 1	Math 205, 206 CECS 121, 230 16 hours of general education	14
Year 2	Math 301, 311, 325 CECS 303, 335 15 hours of general education, science requirements	15
Year 3	Math 560, 561, 562 CECS 545, 632 12 hours of science and general education	15
Year 4	Math 405, 501-502, 564, 665-667 Thesis	18
Computer Science Electives (Applications Area)	Math 566, 567, 660-662, 635-536	18

Summary

- Statistics is a marketable and profitable field with a large number of possible fields of specialization.
- Data mining is a process of data analysis that is used greatly in business but rarely in medicine.
- There are many opportunities available to analyze health data using data mining tools.
- The Data Mining Applications Area can be a part of the BS/MA, MA, and PhD curricula.