

STAT597E/CSE598E: Data Mining

Fall 2006: September 5 - December 15, Part I

Instructor:

- [Jia Li](#)
417A Thomas Building, phone: 814-863-3074, email: jiali@psu.edu
office hours: MF 2:30-3:30pm, or by appointment

Lectures: TTh 11:15am-12:30pm 106 Wartik

Course homepage: <http://www.stat.psu.edu/~jiali/course/stat597e>

Description of the course:

With rapid advances in information technology, we have witnessed an explosive growth in our capabilities to generate and collect data in the last decade. In the business world, very large databases on commercial transactions have been generated by retailers. Huge amount of scientific data have been generated in various fields as well. For instance, the human genome database project has collected gigabytes of data on the human genetic code. The World Wide Web provides another example with billions of web pages consisting of textual and multimedia information that are used by millions of people. How to analyze huge bodies of data so that they can be understood and used efficiently remains a challenging problem. Data mining addresses this problem by providing techniques and software to automate the analysis and exploration of large complex data sets. Research on data mining have been pursued by researchers in a wide variety of fields, including statistics, machine learning, database management and data visualization.

This course on data mining will cover methodology, major software tools and applications in this field. By introducing principal ideas in statistical learning, the course will help students to understand conceptual underpinnings of methods in data mining. Considerable amount of effort will also be put on computational aspects of algorithm implementation. To make an algorithm efficient for handling very large scale data sets, issues such as algorithm scalability need to be carefully analyzed. Data mining techniques developed in fields other than statistics, e.g., support vector machine in machine learning will be introduced as well.

Students will be required to work on projects to practice applying existing software and to a certain extent, developing their own algorithms. Classes will be provided in three forms: lecture, case study, and project discussion. In case study, students will be lead through practical problems addressed by data mining techniques. The aim is to provide a detailed view on how to convert real problems into models so that algorithms can be applied appropriately and how to solve possible computational issues. Project discussion will enable students to share and compare ideas with each other and to receive specific guidance from the instructors.

Provided with the rich content in data mining, we plan to cover this course as Part I and II in a consecutive fall semester and spring semester. In Part I, basics for classification and clustering, e.g., linear classification methods, prototype methods, decision trees, and hidden Markov models, will be introduced. Roughly five course projects will be included in this part with emphasis on understanding and using existing learning algorithms. Students are expected to use C, Matlab, or S-plus for moderate amount of programming. Relatively advanced techniques such as support vector machine, boosting, and dimension reduction will be covered in Part II of the course. Course projects in Part II will be designed at a larger scale than those in Part I; and will be related to current research in multimedia information systems and bioinformatics. Students will be encouraged to bring to discussion their own research problems with potential applications of data mining methods. Possible project topics include image segmentation and image retrieval; text search, link analysis, and summarization; microarray data analysis; and recommender systems for books and movies.

Prerequisites: Stat 414, 415, or equivalent.

Textbooks:

Required: [The Elements of Statistical Learning](#), by Trevor Hastie, Robert Tibshirani, and Jerome Friedman

Recommended:

- *Classification and Regression Trees* by L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone.
- *Pattern Recognition and Neural Networks* by B. Ripley
- *Principles of Data Mining* by H. Mannila, P. Smyth and D. J. Hand
- *Data Mining: Concepts and Techniques* by J. Han and M. Kamber

Grading:

- Projects: 80%
- Participation and discussion: 20%

Note: late projects are not allowed without written request submitted and approved one week ahead of the due date.

Academic Integrity:

All Penn State and Eberly College of Science policies regarding academic integrity apply to this course. See <http://www.science.psu.edu/academic/Integrity/index.html> for details.

Topic Schedule

Lecture Dates, [academic calendar](#)

 Lecture Notes & Other Course Materials

----- Updated on August 29, 2006 -----