

STAT557/IST557: Data Mining

Fall 2009: August 24 - December 18

Instructor:

- [Jia Li](#)
417A Thomas Building, phone: 814-863-3074, email: jjiali@psu.edu
office hours: MW 1:30pm-3:00pm

Teaching Assistant:

- Eun Yeong Ahn
324 IST Building, phone: 814-571-6342, email: eu121@ist.psu.edu
office hours: TTh 1:30pm-3:30pm
- Hyang Min Lee
331A Thomas Building, phone: 814-863-1772, email: hul145@ist.psu.edu
office hours: T: 9:00-10:00am, Th: 9:00am-12:00pm

Lectures: MW 9:45am-11:00am 201A IST

Course homepage: <http://www.stat.psu.edu/~jjiali/course/stat557>

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 and learning techniques developed in fields other than statistics, e.g., machine learning and signal processing, will also be introduced.

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, project

discussion, and special topic survey. Project discussion will enable students to share and compare ideas with each other and to receive specific guidance from the instructors. Efforts will be made to help students formulate real-world problems into mathematical models so that suitable algorithms can be applied with consideration of computational constraints. By surveying special topics, students will be exposed to massive literature and become more aware of recent research.

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. Part II will extend Part I with more techniques in machine learning and large-scale data processing. The focus will be on the breadth of data mining and its applications in information technology. 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, 416, or similar courses that cover basics on probability, expectation, and conditional distribution.

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
- *Pattern Recognition and Machine Learning* by C. M. Bishop
- *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: 70%
- Survey of special topics: 20%
- Participation: 10%

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.

Lecture Notes & Other Course Materials:

Course notes, reading materials, data sets, and project description

Weekly Schedule of Topics

Lecture Dates, academic calendar

Special Fall 2009 Flu Protocols:

In compliance with Pennsylvania Department of Health and Centers for Disease Control recommendations, students should NOT attend class or any public gatherings while ill with influenza. Students with flu symptoms will be asked to leave campus if possible and to return home during recovery. The illness and self-isolation period will usually be about a week. It is very important that individuals avoid spreading the flu to others.

Most students should be able to complete a successful semester despite a flu-induced absence. Faculty will provide students who are absent because of illness with a reasonable opportunity to make up missed work. Ordinarily, it is inappropriate to substitute for the missed assignment the weighting of a semester's work that does not include the missed assignment or exam. Completion of all assignments and exams assures the greatest chance for students to develop heightened understanding and content mastery that is unavailable through the weighting process. The opportunity to complete all assignments and exams supports the university's desire to enable students to make responsible situational decisions, including the decision to avoid spreading a contagious virus to other students, staff, and faculty, without endangering their academic work.

Students with the flu do not need to provide a physician's certification of illness. However, ill students should inform their teachers (but not through personal contact in which there is a risk of exposing others to the virus) as soon as possible that they are absent because of the flu. Likewise students should contact their instructors as quickly as possible to arrange to make up missed assignments or exams.

----- Updated on 2009 -----