Course curriculum for STAN45 Data Mining and Visualization

1. General information

1. Name: Data Mining and Visualization
2. Level: Advanced (A1N)
3. Credit points: 7.5; ECTS-credits: 7.5
4. Approved by the Board of Directors at the Department of Statistics, School of Economics and Management, Lund University on November 21, 2012.

2. Course placement within the educational system

1. Subject: Statistics
2. This is an advanced level course.
3. The course is offered in English.

3. Learning outcomes

On a general level the students should be able to understand and identify the challenges in analyzing massive data. They will learn to select a proper data mining algorithm for extracting meaningful information from such data. The student will also become familiar with a wide range of visualizations techniques from which he/she can select to properly present both to specialists and to general public results of performed data mining.

4. Course content

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 and financial world, very large databases on commercial and financial transactions have been generated by retailers, traders and banks. In science, 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. Another example, are climate and environmental data collected by satellites. The World Wide Web provides another example with billions pages of textual and multimedia information that are used every day 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 is a collection of more universal methods that address this problem by providing techniques and software to automate the analysis and exploration of large complex data sets. Research on data mining has been pursued by researchers in a wide variety of fields, including statistics, machine learning, database management and data visualization. This is an emerging and rapidly developing field that requires understanding both established method and newly adopted techniques.

This course on data mining and visualization covers methodology, major programming tools and applications in this field. By introducing principal ideas in statistical learning, the course helps students to understand methods in data mining and 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. The course also explores the question of what visualization is, and why one should use visualizations for quantitative data.

Students are required to work on projects to practice applying existing software and to a certain extent, developing their own algorithms. Classes are 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 growing range of new methodologies.

In particular, basics for classification and clustering, e.g., linear classification methods, prototype methods, decision trees, and hidden Markov models, are introduced. Roughly five course lab sessions are included with emphasis on understanding and using existing learning algorithms. Students will be encouraged to bring to discussion their own research problems with potential applications of data mining methods. Possible topics include image segmentation and image retrieval; text search, link analysis, and summarization; microarray data analysis; and recommender systems for books and movies. A variety of common and different digital visualization software tools are also presented. Lab sessions focus on providing practice using real-world data.

5. Teaching and assessment

The course is designed as a series of lectures, student presentations, and lab sessions with reports. Grading is based on individual performance, via written assignments, oral presentation as well as group activities.

Note

The university views plagiarism very seriously, and will take disciplinary actions against students for any kind of attempted malpractice in examinations and assessments. Plagiarism is considered to be a very serious academic offence. The penalty that may be imposed for this, and other unfair practice in examinations or assessments, includes suspension from the University for a specified period.

6. Grading scale

At the School of Economics and Management grades are awarded in accordance with a criterion-based grading scale A-F:

A: Excellent
B: Very good
C: Good
D: Satisfactory
E: Sufficient
F: Fail

Students have to receive a grade of E or higher in order to pass a course.

GRADE	CHARACTERISTIC	POINTS	CRITERIA
A	Excellent	100-85	A distinguished result that is excellent with regard to the following aspects – theoretical depth, practical relevance, analytical ability and independent thought.
B	Very good	84-75	A very good result with regard to the above mentioned aspects.
C	Good	74-65	The result is of a good standard with regard to the above mentioned aspects and lives up to expectations.
D	Satisfactory	64-55	The result is of a satisfactory standard with regard to the above mentioned aspects and lives up to expectations.
E	Sufficient	54-50	The result satisfies the minimum requirements with regard to the above mentioned aspects, but not more.
U (F)	Fail	49-0	The result does not meet the minimum requirements with regard to the above mentioned aspects.

7. Prerequisities

General prerequisities for the masters programme in Statistics.

8. Literature

Requested: *The Elements of Statistical Learning*, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, Springer, the 2nd edition.

Recommended: *Data Mining with R. Learning with Case Studies*, by Luis Torgo
Principles of Data Mining by H. Mannila, P. Smyth and D. J.Hand
Data Mining: Concepts and Techniques by J. Han and M. Kamber