School of Economics and Management

STAN45, Statistics: Data Mining and Visualization, 7.5 credits

Statistics: Data Mining and Visualization, 7,5 högskolepoäng

Second Cycle / Avancerad nivå

Details of approval

The syllabus was approved by The Board of the Department of Statistics on 2015-06-08 and was last revised on 2015-06-08. The revised syllabus applies from 2015-06-08, autumn semester 2015.

General Information

Language of instruction: English

Main field of studies	*Depth of study relative to the degree requirements*
Statistics	A1N, Second cycle, has only first-cycle course/s as entry requirements

Learning outcomes

Knowledge and understanding

For a passing grade the student must

- demonstrate knowledge of analytical methods for massive data that are difficult to process without modern computational tools (the Big Data problem).

Competence and skills

For a passing grade the student must

- demonstrate the ability to apply a data mining solution to a practical problem involving complex, big data sets focusing on economics and business oriented applications,
- demonstrate the ability to choose an algorithm that can efficiently handle very large data sets but also judge the scalability of the algorithm.

Judgement and approach

For a passing grade the student must

- demonstrate familiarity with statistical techniques useful for drawing patterns from multidimensional data, which help improving decision making.

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 analyse 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 analysed. 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 microarray data analysis. Lab sessions focus on providing practice using real-world data.

Course design

The course is designed as a series of lectures, tutorials, and lab sessions with reports. Teaching is offered in three forms: lectures, project discussions and an in-depth study.

Assessment

Grading is based on individual performance, via written assignments, oral presentation as well as group activities.

The examination consists of written assignments and a computer based exam.

Subcourses that are part of this course can be found in an appendix at the end of this document.

Grades

Marking scale: Fail, E, D, C, B, A.

A (Excellent) 85-100 points/percent. A distinguished result that is excellent with regard to theoretical depth, practical relevance, analytical ability and independent thought.

B (Very good) 75-84 points/percent. A very good result with regard to theoretical depth, practical relevance, analytical ability and independent thought.

C (Good) 65-74 points/percent. The result is of a good standard with regard to theoretical depth, practical relevance, analytical ability and independent thought and lives up to expectations.

D (Satisfactory) 55-64 points/percent. The result is of a satisfactory standard with regard to theoretical depth, practical relevance, analytical ability and independent thought.

E (Sufficient) 50-54 points/percent. The result satisfies the minimum requirements with regard to theoretical depth, practical relevance, analytical ability and independent thought, but not more.

F (Fail) 0-49 points/percent. The result does not meet the minimum requirements with regard to theoretical depth, practical relevance, analytical ability and independent thought.

To pass the course, the students must have been awarded the grade of E or higher.

Entry requirements

General prerequisites for the Master's programme in Statistics.

Subcourses in STAN45, Statistics: Data Mining and Visualization

Applies from H15

1301 Data Mining and Visualization, exam, 7,5 hp
Grading scale: Fail, E, D, C, B, A
1303 Data Mining and Visualization, lab, 0,0 hp
Grading scale: Fail, Pass

Applies from H13

1301 Data Mining and Visualization, exam, 7,5 hp
Grading scale: Fail, E, D, C, B, A
1302 Data Mining and Visualization, lab, 0,0 hp
Grading scale: Fail, E, D, C, B, A