

Data Mining and Statistical Learning

Single subject and programme course

15 credits

Data Mining and Statistical Learning

732A33

Valid from:

Determined by

The Quality Board at the Faculty of Arts
and Sciences

Date determined

2008-09-10

Revision date

2013-03-18

Main field of study

Statistics

Course level

Second cycle

Advancement level

A1X

Entry requirements

For acceptance to the course, the student must have a bachelor's degree with a total of at least 90 ECTS credits (1.5 years of full-time studies) in mathematics, applied mathematics, statistics, and computer science. The undergraduate courses in mathematics should include both calculus and linear algebra. Basic undergraduate courses in statistics and computer science are also required. Documented knowledge of English equivalent to Engelska B/Engelska 6. internationally recognized test, e.g. TOEFL (minimum scores: Paper based 575 + TWE-score 4.5, and internet based 90), IELTS, academic (minimum score Overall band 6.5 and no band under 5.5), or equivalent.

Intended learning outcomes

The course lays the foundation for professional work and research in which large amounts of data are explored, modified, modelled and assessed to uncover previously unknown patterns and trends.

Having completed the course, the student should be able to:

- account for the principles of statistical modeling, in particular for the analysis of large data sets
- utilize tools in SAS environment to explore large and complex data sets, derive data-based models, assess their outcomes and use such models for forecasting
- compare the performance of statistical and data mining models in order to select the most appropriate model in a given context

Course content

- basic concepts in statistical learning, in particular supervised learning,
- model selection strategies involving the use of training sets, validation sets, and test sets and model selection by cross-validation.
- linear regression technique and regression shrinkage methods
- spline smoothers and kernel smoothers,
- decision trees and classification methods, such as discriminant analysis and logistic regression,
- neural networks, support vector machines, and generalized additive models
- ensemble methods, including bagging and boosting.
- Bayesian approach in data mining

Teaching and working methods

The teaching comprises lectures, seminars, and computer exercises. Lectures are devoted to presentations of theories, concepts and methods. Computer exercises provide practical experience of data analysis in SAS environment (as a rule) or in other environments (in exceptional cases). The seminars comprise student presentations and discussions of computer assignments.
Language of instruction: English.

Examination

Written reports on the computer assignments. Obligatory attendance of the seminars. One final written examination.

Grades

ECTS, EC

Other information

Planning and implementation of a course must take its starting point in the wording of the syllabus. The course evaluation included in each course must therefore take up the question how well the course agrees with the syllabus. The course is carried out in such a way that both men's and women's experience and knowledge is made visible and developed.

Department

Institutionen för datavetenskap