GHENT UNIVERSITY

Course Specifications

Valid as from the academic year 2024-2025

Data Mining (C002703)

Course size *(nominal values; actual values may depend on programme)*

Credits 3.0 **Study time 80 h**

Course offerings and teaching methods in academic year 2024-2025

A (semester 1)	English	Gent	lecture
			seminar

Lecturers in academic year 2024-2025

Saeys, Yvan	WE02	lecturer-in-charge

Offered in the following programmes in 2024-2025	**crdts**	**offering**
Master of Science in Biochemistry and Biotechnology	3	A
Exchange programme in Biochemistry and Biotechnology (master's level)	3	A

Teaching languages

English

Keywords

Bioinformatics, Data mining, Data analysis, Pattern recognition, Classification, Regression, Feature selection, Sequence and microarray data

Position of the course

This course provides the student with the basic principles of data mining, and its applications in bioinformatics. The analysis of high-diemnsional and complex data sets is a problem that becomes more and more persistent in biotechnology and biology, requiring the use of advanced computer methods to analyze these data sets, and get insight into the processes that are being modelled.
The course discusses both classification and regression methods, acquinting the students with these methods in practical sessions where they can apply the methods that were seen in the theory, using existing data mining software, and applying the techniques to real world data sets.
This course contributes to the following program competencies: Ma.WE.BB.1.2, Ma.WE.BB. 1.3,Ma.WE.BB.2.5, Ma.WE.BB.2.6

Contents

- Overview of data mining techniques
  - Design cycle of data mining algorithms
  - Relations between bias, variance and model complexity
- Classification Methods:
  - Nearest Neighbors methods
  - Classification Trees
  - Bayesian classifiers
  - Linear Discriminant Analysis
  - Kernel methods (including SVM)
  - Applications in Bioinformatics
- Clustering Methods:
  - Hierarchical clustering
  - K-means
  - Self-organizing maps
  - Applications in Bioinformatics
- Regression methods:
  - Regression Trees
  - Principal Component Regression and Partial Least Squares

- Other regression methods (?)
- Model building, selection and inference:
  - Methods to estimate prediction error
  - Cross-validation
  - Bootstrap
  - Ensembles
  - Bagging - Boosting
- Methods for dealing with high dimensional data:
  - Principal Component Analysis
  - Independent Component Analysis
  - Feature selection

Initial competences

Basic knowledge in bioinformatics, computer science and statistics

Final competences

1 The student is able to propose the appropriate method for a giving data mining problem to realize a specific objective.
2 The student is able to understand, assimilate, and apply recent literature on data mining in bioinformatics.

Conditions for credit contract

Access to this course unit via a credit contract is determined after successful competences assessment

Conditions for exam contract

This course unit cannot be taken via an exam contract

Teaching methods

Group work, Seminar, Lecture

Extra information on the teaching methods

Theory: software demonstrations
Exercises: computer exercises, implementation as well as use of existing software packages, project work in small groups

Study material

Type: Handouts

Name: Powerpoint presentations of the theory classes, supporting material, program code and data'
Indicative price: € 10
Optional: no
Additional information: available on Ufora

References

T. Mitchell (1997). Machine Learning. McGraw-Hill
R.O. Duda, P.E. Hart, and D.G. Stork. (2001) Pattern Classification, Wiley, New York
P. Baldi and S. Brunak (1998) Bioinformatics, the machine learning approach, MIT Press

Course content-related study coaching

The lecturer announces office hours for problems related to the theory.
Supervised practical sessions.

Assessment moments

end-of-term and continuous assessment

Examination methods in case of periodic assessment during the first examination period

Oral assessment

Examination methods in case of periodic assessment during the second examination period

Oral assessment

Examination methods in case of permanent assessment

Assignment

Possibilities of retake in case of permanent assessment

examination during the second examination period is possible

Extra information on the examination methods

Periodical evaluation: oral presentation of the project work and oral examination

Calculation of the examination mark

Periodical evaluation (theory) (50%) + non-periodical evaluation (exercises) (50%) In case a student has not passed the non-periodical evaluation, a second chance is offered by means of a compensatory activity between the first and the second examination period.