GHENT UNIVERSITY

Course Specifications

Valid as from the academic year 2020-2021

Predictive Modelling (I002091)

Due to Covid 19, the education and evaluation methods may vary from the information displayed in the schedules and course details. Any changes will be communicated on Ufora.

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

Credits 6.0 Study time 150 h Contact hrs 60.0 h

Course offerings and teaching methods in academic year 2020-2021

A (semester 2)	English	Gent	guided self-study	10.0 h
			microteaching	7.5 h
			seminar: practical PC room classes	20.0 h
			lecture	22.5 h
B (semester 2)			seminar: practical PC room classes	20.0 h
			lecture	22.5 h
			microteaching	7.5 h

Lecturers in academic year 2020-2021

Waegeman, Willem	LA26	lecturer-in-charge

Offered in the following programmes in 2020-2021	crdts	offering
Master of Science in Bioinformatics (main subject Bioscience Engineering)	6	A
Master of Science in Bioinformatics (main subject Systems Biology)	6	A
Master of Science in Bioscience Engineering: Forest and Nature Management	5	B
Master of Science in Bioscience Engineering: Cell and Gene Biotechnology	5	B
Master of Science in Bioscience Engineering: Land and Water Management	5	B
Master of Science in Bioscience Engineering: Environmental Technology	5	B
Exchange Programme in Bioinformatics (master's level)	6	B
Exchange Programme in Bioscience Engineering: Agricultural Sciences (master's level)	5	B
Exchange Programme in Bioscience Engineering: Cell and Gene Biotechnology (master's level)	5	B
Exchange Programme in Bioscience Engineering: Chemistry and Bioprocess Technology (master's level)	5	B
Exchange Programme in Bioscience Engineering: Environmental Technology (master's level)	5	B
Exchange Programme in Bioscience Engineering: Food Science and Nutrition (master's level)	5	B
Exchange Programme in Bioscience Engineering: Land and Forest management (master's level)	5	B

Teaching languages

English

Keywords

Classification, regression, advanced prediction problems, applications in the life sciences

Position of the course

This course aims to give the students an introduction to the field of predictive modelling

(aka machine learning), aiming at the recognition and prediction of complex patterns in data. Predictive computational and statistical models are needed in many applications in bioinformatics and the life sciences in general. Think in this context at the prediction of diseases from genetic data, the forecast of natural events from climate data, the modelling of ecological migration from geografic data, etc.

Both classification methods (output = class label) and regression methods (output = real value) will be discussed thoroughly. Moreover, more involved prediction problems are touched upon as well, such as learning structured objects, multi-label classification, ordinal regression, etc. The students will apply the considered machine learning methods by means of existing software on several case studies that are situated in the life sciences. In the theory part as well, the course instructors are specifically focusing on predictive methods for the life sciences, by dealing with subjects such as predicting from high-dimensional and non-vectorial data.

Contents

The course focuses on introducing machine learning principles, techniques and applications. Every method will be practiced during computer exercise sessions on real biological problems.
1 Theoretical background on machine learning (classification vs. regression, training vs. test data, overfitting, cross validation, performance measures, modelling high-dimensional and non-vector data, complexity control and regularization, relation with bias-variance decomposition.
2. Basic regression methods (least squares linear regression, ridge regression and lasso, robust regression, non-linear regression through basic expansions, nearest neighbor methods)
3. Basic classification methods (least squares classification, logistic regression, linear discriminant analysis).
4. Specialized machine learning methods (neural networks, bagging, boosting, random forests, kernel methods).
5. Modelling of structured and non-vectorial data (text, DNA sequences, graphs, images, spectral data, etc.)
6. The role of the data scientist: pipelines, pitfalls, data preprocessing, visualization
7. Big data science (large-scale learning, optimization techniques)
8. Advanced prediction problems (multidimensional prediction, structured prediction, graph learning, semi-supervised learning, etc.)

Initial competences

It is important that the students have already hands-on experience with programming (Matlab, Python, R, etc.). We will use Python in the PC-classes.
Basic knowledge of mathematics, informatics, probability and statistics is recommended, in particular the following topics:
- vector and matrix algebra
- least squares problems
- singular value decomposition
- quadratic forms
- the gradient and partial derivatives
- extreme values of functions
- Lagrange multipliers
- Bayes' rule
- probability distributions (in particular the Gaussian distribution)
- linear regression

Final competences

The student must be able to:
1. select the most appropriate method for a given classification or regression problem;
2. apply these methods and interpret the results;
3. understand recent literature in machine learning, process and apply the methods presented in these articles.

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

Guided self-study, lecture, microteaching, seminar: practical PC room classes

Learning materials and price

1. Presentations are available on Minerva in PDF format.
2. Background study material, software code and data are made available on Minerva.
3. The course book can be freely downloaded by internet.

References

James et al. An introduction to statistical learning, Springer 2013.

Course content-related study coaching

1. The lecturer announces office hours for problems related to the theory.
2. The teaching assistant guides the practical exercises.

Evaluation methods

end-of-term evaluation and continuous assessment

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

Oral examination, participation, assignment, report

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

Oral examination, participation, assignment, report

Examination methods in case of permanent evaluation

Oral examination

Possibilities of retake in case of permanent evaluation

examination during the second examination period is possible in modified form

Extra information on the examination methods

Course offering A: The student will be evaluated based on 3 projects.
Course offering B: The student will be evaluated based on the first 2 projects.
Project 1 (Weight 2/6): Presentation of an advanced machine learning method, to be chosen from a list of topics.
Project 2 (Weight 3/6): Data mining competition, organized among the students, with submission of the results and a short report.
Project 3 (Weight 1/6): Application of the knowledge acquired on an elaborated case study in the field of bioinformatics, based on an exhaustive literature study. For the final exam, the student has to give an oral presentation (+/- 20 minutes + questions) and submit a profound report.

Course offering B: The student will be evaluated based on 3 projects, of which project 3 differs from course offering A:
Project 1 (Weight 2/5): Presentation of an advanced machine learning method, to be chosen from a list of topics.
Project 2 (Weight 3/5): Data mining competition, organized among the students, with submission of the results and a short report.

Calculation of the examination mark

Students who eschew period aligned and/or non-period aligned evaluations for this course unit may be failed by the examinator.