University of St.Gallen

Course and Examination Fact Sheet: Spring Semester 2022

10,365: Computational Statistics

ECTS credits: 4

Overview examination/s

(binding regulations see below)
Decentral - examination paper written at home (in groups - all given the same grades) (100%)
Examination time: term time

Attached courses

Timetable -- Language -- Lecturer
10,365,1.00 (GSERM) Computational Statistics -- Englisch -- Audrino Francesco

Course information

Course prerequisites

Advanced knowledge in statistics and econometrics.

Learning objectives

- Students will gain an advanced knowledge on the statistical aspects related to the use of machine learning techniques needed to analyze large or high-dimensional datasets.
- Students will learn how to apply machine learning tools in a responsible way and will properly apply the methods on a concrete dataset of their choice using the statistical R software and prepare a research paper summarizing their results.

Course content

Computational Statistics is the area of specialization within statistics that includes statistical visualization and other computationally-intensive methods of statistics for mining large, nonhomogeneous, multi-dimensional datasets so as to discover knowledge in the data. As in all areas of statistics, probability models are important, and results are qualified by statements of confidence or of probability. An important activity in computational statistics is model building and evaluation.

First, the basic multiple linear regression is reviewed. Then, some nonparametric procedures for regression and classification are introduced and explained. In particular, Kernel estimators, smoothing splines, classification and regression trees, additive models, projection pursuit and eventually neural nets will be considered, where some of them have a straightforward interpretation, other are useful for obtaining good predictions.

The main problems arising in computational statistics like the curse of dimensionality will be discussed. Moreover, the goodness of a given (complex) model for estimation and prediction is analyzed using resampling, bootstrap and cross-validation techniques.

Course structure and indications of the learning and teaching design

Outline:

1. Overview of supervised learning

Introductory examples, two simple approaches to prediction, statistical decision theory, local methods in high dimensions, structured regression models, bias-variance tradeoff, multiple testing and use of p-values.

2. Linear methods for regression

Multiple regression, analysis of residuals, subset selection and coefficient shrinkage.

3. Methods for classification

Bayes classifier, linear regression of an indicator matrix, discriminant analysis, logistic regression.

4. Nonparametric density estimation and regression

Histogram, kernel density estimation, kernel regression estimator, local polynomial nonparametric regression estimator, smoothing splines and penalized regression.

5. Model assessment and selection

Bias, variance and model complexity, bias-variance decomposition, optimism of the training error rate, AIC and BIC, cross-validation, boostrap methods.

6. Flexible regression and classification methods

Additive models; multivariate adaptive regression splines (MARS); neural networks; projection pursuit regression; classification and regression trees (CART).

7. Bagging and Boosting

The bagging algorithm, bagging for trees, subagging, the AdaBoost procedure, steepest descent and gradient boosting.

8. ***Introduction to the idea of a Superlearner***

Course literature

Main references:

- F. Audrino, Lecture Notes.
- Hastie T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning: data mining, inference and prediction*, Springer Series in Statistics, Springer, Canada.
- Bühlmann, P. and van de Geer, S. (2011).*Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer.
- van der Laan, M.J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer.

References to related published papers / chapters of books will be given during the course.

Additional course information

Due to the current pandemic and the limits on the lecture rooms' capacity to maintain the proper social distancing in place, if it is not possible for all students to follow the course in class in SpringS22, the course is conducted entirely online via the platform Zoom.

In case of a full lockdown there are no changes necessary to the examination information.

Only for PhD students of the University of St.Gallen

PEF & PEcon students may register via regular bidding for the courses offered together by PEcon and Global School in Empirical Research Methods (GSERM). Enrolment in a course is binding: students have to attend the course and take the exam. The credits will be shown on the scorecard.

All other PhD students should register for the courses offered by Global School in Empirical Research Methods (GSERM), both via bidding and via GSERM for:

-courses for the curriculum and

-optional courses with an examination. These will be listed on the scorecard under optional work (only possible if all required elective courses have already been completed).

University of St.Gallen

Please register only via GSERM for:

-optional courses without an examination and

-optional courses if not all required elective courses have been completed (not shown on the scorecard)

Examination information

Examination sub part/s

1. Examination sub part (1/1)

Examination time and form
Decentral - examination paper written at home (in groups - all given the same grades) (100%)
Examination time: term time

Remark
--

Examination-aid rule
Term papers

Written work must be written without outside help according to the known citation standards, and a declaration of authorship must be attached, which is available as a template on the StudentWeb.

Documentation (quotations, bibliography, etc.) must be carried out universally and consistently according to the requirements of the chosen/specified citation standard such as e.g. APA or MLA.

The legal standard is recommended for legal work (cf. by way of example: FORSTMOSER, P., OGOREK R., SCHINDLER B., Juristisches Arbeiten: Eine Anleitung für Studierende (the latest edition in each case), or according to the recommendations of the Law School).

The reference sources of information (paraphrases, quotations, etc.) that has been taken over literally or in the sense of the original text must be integrated into the text in accordance with the requirements of the citation standard used. Informative and bibliographical notes must be included as footnotes (recommendations and standards e.g. in METZGER, C., Lern- und Arbeitsstrategien (latest edition)).

For all written work at the University of St.Gallen, the indication of page numbers is mandatory, regardless of the standard chosen. Where page numbers are missing in sources, the precise designation must be made differently: chapter or section title, section number, article, etc.

Supplementary aids
--

Examination languages
Question language: English
Answer language: English

Examination content

Outline:

1. Overview of supervised learning

Introductory examples, two simple approaches to prediction, statistical decision theory, local methods in high dimensions, structured regression models, bias-variance tradeoff, multiple testing and use of p-values.

2. Linear methods for regression

Multiple regression, analysis of residuals, subset selection and coefficient shrinkage.

3. Methods for classification

Bayes classifier, linear regression of an indicator matrix, discriminant analysis, logistic regression.

4. Nonparametric density estimation and regression

Histogram, kernel density estimation, kernel regression estimator, local polynomial nonparametric regression estimator, smoothing splines and penalized regression.

5. Model assessment and selection

Bias, variance and model complexity, bias-variance decomposition, optimism of the training error rate, AIC and BIC, cross-validation, boostrap methods.

6. Flexible regression and classification methods

Additive models; multivariate adaptive regression splines (MARS); neural networks; projection pursuit regression; classification and regression trees (CART).

7. Bagging and Boosting

The bagging algorithm, bagging for trees, subagging, the AdaBoost procedure, steepest descent and gradient boosting.

8. Introduction to the idea of a superlearner

Examination relevant literature

F. Audrino, Lecture Notes, available on Canvas before the beginning of the course.

Please note

Please note that only this fact sheet and the examination schedule published at the time of bidding are is binding and takes precedence over other information, such as information on StudyNet (Canvas), on lecturers' websites and information in lectures etc.

Any references and links to third-party content within the fact sheet are only of a supplementary, informative nature and lie outside the area of responsibility of the University of St.Gallen.

Documents and materials are only relevant for central examinations if they are available by the end of the lecture period (CW21) at the latest. In the case of centrally organised mid-term examinations, the documents and materials up to CW 12 are relevant for testing.

Binding nature of the fact sheets:

- Course information as well as examination date (organised centrally/decentrally) and form of examination: from bidding start in CW 04 (Thursday, 27 January 2022);
- Examination information (regulations on aids, examination contents, examination literature) for decentralised examinations: in CW 12 (Monday, 21 March 2022);
- Examination information (regulations on aids, examination contents, examination literature) for centrally organised mid-term examinations: in CW 12 (Monday, 21 March 2022);
- Examination information (regulations on aids, examination contents, examination literature) for centrally organised examinations: two weeks before the end of the registration period in CW 15 (Monday, 11 April 2022).