ST420-15 Statistical Learning and Big Data

26/27

Department
Statistics
Level
Undergraduate Level 4
Module leader
Elke Thonnes
Credit value
15
Module duration
10 weeks
Assessment
Multiple
Study location
University of Warwick main campus, Coventry

Description

Introductory description

This module will introduce students to modern applications of Statistics in challenging modern data analysis contexts and provide them with the theoretical underpinnings to apply these methods.

This module is available for students on a course where it is a listed option and as an Unusual Option to students who have the background knowledge as indicated by the pre-requisite modules.

Pre-requisites:

UG students:

- Statistics students.
    - ST228 Mathematical Methods for Statistics and Probability and ST229 Probability for Mathematical Statistic.
    - ST231 Linear Statistical Modelling with R.
- Non-statistics students.
    - ST121 Statistical Laboratory and ST232/ST233 Introduction to Mathematical Statistics

or ST352 Introduction to Mathematical Statistics (for Finalists).
    - ST240 Linear Statistical Modelling or ST351 Linear Statistical Modelling (For Finalists).

We recommend taking ST340 Fundamentals of Machine Learning (formerly Programming for Data Science) before taking ST420 as it provides a good background for the module.

MSc students:

- ST963 Theory of Data Science, or
- MA907 Simulation and Machine Learning.

Module web page

Module aims

This module aims to provide an in-depth theoretical foundation for modern statistical machine learning concepts and procedures, while also introducing key methods and topics in advanced statistical machine learning relevant to the big data era.

Outline syllabus

This is an indicative module outline only to give an indication of the sort of topics that may be covered. Actual sessions held may differ.

1. Recap of fundamental concepts and core methods in statistical learning
2. Computational, algorithmic, statistical, and ethical considerations in big data
3. Multiple testing in big data
4. Statistical learning with sparsity in big data: ridge regression and the Lasso
5. Support vector machines, kernels and reproducing kernel Hilbert spaces
6. Bagging and boosting
7. Statistical learning theory including PAC bounds and VC theory
8. Selected topics in big data and large-scale models (varying from year to year): double descent, implicit bias of gradient descent and benign overfitting, privacy and fairness, MCMC in high dimensions

Learning outcomes

By the end of the module, students should be able to:

- Explain, critically discuss and apply fundamental concepts and analytic tools in Statistical Learning;
- Analyse and discuss issues and fundamental tools in the analysis of Big Data and Big Models;

- Implement and assess methods for prediction based on partitioning data;
- Apply fundamental tools based on sparsity, regularisation and the control of error rates to analyse large data sets.

Indicative reading list

Specific reading list for the module

Subject specific skills

- Evaluate, select and apply appropriate mathematical and/or probabilist techniques.
- Demonstrate knowledge of and facility with formal probability concepts, both explicitly and by applying them to the solution of problems.
- Create structured and coherent arguments communicating them in written form.
- Construct logical mathematical arguments with clear identification of assumptions and conclusions.
- Reason critically, carefully, and logically and derive (prove) mathematical results.

Transferable skills

- Problem solving: Use rational and logical reasoning to deduce appropriate and well-reasoned conclusions. Retain an open mind, optimistic of finding solutions, thinking laterally and creatively to look beyond the obvious. Know how to learn from failure.
- Self awareness: Reflect on learning, seeking feedback on and evaluating personal practices, strengths and opportunities for personal growth.
- Communication: Present arguments, knowledge and ideas, in a range of formats.
- Professionalism: Prepared to operate autonomously. Aware of how to be efficient and resilient. Manage priorities and time. Self-motivated, setting and achieving goals, prioritising tasks.

Study

Study time

Type	Required
Lectures	30 sessions of 1 hour (20%)
Private study	90 hours (60%)
Total	150 hours

Type	Required
Assessment	30 hours (20%)
Total	150 hours

Private study description

Weekly revision of lecture notes and materials, wider reading, practice exercises and preparing for examination.

Other activity description

Revision support.

Costs

No further costs have been identified for this module.

Assessment

You do not need to pass all assessment components to pass the module.

Students can register for this module without taking any assessment.

Assessment group D6

	Weighting	Study time	Eligible for self-certification
Assignment 1	10%	15 hours	No
The assignment will contain a number of questions for which solutions and / or written responses will be required. The study time refers to the amount of time in hours that a well-prepared student who has attended lectures and carried out an appropriate amount of independent study on the material could expect to spend on this assignment. Your ST420 Assignment 1 should not exceed 15 pages in length.			
Assignment 2	10%	15 hours	No
The assignment will contain a number of questions for which solutions and / or written responses will be required. The study time refers to the amount of time in hours that a well-prepared student who has attended lectures and carried out an appropriate amount of independent study on the material could expect to spend on this assignment. Your ST420 Assignment 2 should not exceed 15 pages in length.			
Centrally-timetabled examination (On-campus)	80%		No

	Weighting	Study time	Eligible for self-certification

The examination paper will contain four questions, of which the best marks of THREE questions will be used to calculate your grade.

- Students may use a calculator
- Answerbook Pink (12 page)

Assessment group R6

	Weighting	Study time	Eligible for self-certification
In-person Examination - Resit	100%		No

The examination paper will contain four questions, of which the best marks of THREE questions will be used to calculate your grade.

- Answerbook Pink (12 page)
- Students may use a calculator

Feedback on assessment

Assignments are marked and given feedback online within 20 working days of the submission deadline. Where appropriate, model solutions will be provided.

Solutions and cohort level feedback will be provided for the examination. Individual scripts are retained for external examiners and will not be returned.

Past exam papers for ST420

Availability

Courses

This module is Optional for:

- Year 1 of TIBS-N3G1 Postgraduate Taught Financial Mathematics
- TSTA-G4P1 Postgraduate Taught Statistics
  - Year 1 of G4P1 Statistics (Taught)
  - Year 1 of G40B Statistics with Data Science (Taught)
  - Year 1 of G40C Statistics with Finance (Taught)
  - Year 1 of G40A Statistics with Probability (Taught)

- Year 4 of USTA-G304 Undergraduate Data Science (MSci)
- Year 5 of USTA-G305 Undergraduate Data Science (MSci) (with Intercalated Year)
- USTA-G300 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics
    - Year 3 of G30A Master of Maths, Op.Res, Stats & Economics (Actuarial and Financial Mathematics Stream)
    - Year 3 of G30J Master of Maths, Op.Res, Stats & Economics (Data Analysis Stream)
    - Year 3 of G30B Master of Maths, Op.Res, Stats & Economics (Econometrics and Mathematical Economics Stream)
    - Year 3 of G30C Master of Maths, Op.Res, Stats & Economics (Operational Research and Statistics Stream)
    - Year 3 of G30C Master of Maths, Op.Res, Stats & Economics (Operational Research and Statistics Stream)
    - Year 3 of G30D Master of Maths, Op.Res, Stats & Economics (Statistics with Mathematics Stream)
    - Year 3 of G300 Mathematics, Operational Research, Statistics and Economics
    - Year 3 of G300 Mathematics, Operational Research, Statistics and Economics
    - Year 3 of G300 Mathematics, Operational Research, Statistics and Economics
    - Year 4 of G30A Master of Maths, Op.Res, Stats & Economics (Actuarial and Financial Mathematics Stream)
    - Year 4 of G30J Master of Maths, Op.Res, Stats & Economics (Data Analysis Stream)
    - Year 4 of G30B Master of Maths, Op.Res, Stats & Economics (Econometrics and Mathematical Economics Stream)
    - Year 4 of G30C Master of Maths, Op.Res, Stats & Economics (Operational Research and Statistics Stream)
    - Year 4 of G30C Master of Maths, Op.Res, Stats & Economics (Operational Research and Statistics Stream)
    - Year 4 of G30D Master of Maths, Op.Res, Stats & Economics (Statistics with Mathematics Stream)
    - Year 4 of G300 Mathematics, Operational Research, Statistics and Economics
    - Year 4 of G300 Mathematics, Operational Research, Statistics and Economics
    - Year 4 of G300 Mathematics, Operational Research, Statistics and Economics
- USTA-G301 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics (with Intercalated
    - Year 4 of G301 BSc Master of Mathematics, Operational Research, Statistcs and Economics (with Intercalated Year)
    - Year 4 of G30E Master of Maths, Op.Res, Stats & Economics (Actuarial and Financial Mathematics Stream) Int
    - Year 4 of G30K Master of Maths, Op.Res, Stats & Economics (Data Analysis Stream) Int
    - Year 4 of G30F Master of Maths, Op.Res, Stats & Economics (Econometrics and Mathematical Economics Stream) Int
    - Year 4 of G30G Master of Maths, Op.Res, Stats & Economics (Operational Research and Statistics Stream) Int
    - Year 4 of G30H Master of Maths, Op.Res, Stats & Economics (Statistics with Mathematics Stream)

    - Year 5 of G301 BSc Master of Mathematics, Operational Research, Statistcs and Economics (with Intercalated Year)
    - Year 5 of G30E Master of Maths, Op.Res, Stats & Economics (Actuarial and Financial Mathematics Stream) Int
    - Year 5 of G30K Master of Maths, Op.Res, Stats & Economics (Data Analysis Stream) Int
    - Year 5 of G30F Master of Maths, Op.Res, Stats & Economics (Econometrics and Mathematical Economics Stream) Int
    - Year 5 of G30G Master of Maths, Op.Res, Stats & Economics (Operational Research and Statistics Stream) Int
    - Year 5 of G30H Master of Maths, Op.Res, Stats & Economics (Statistics with Mathematics Stream)
- USTA-G1G3 Undergraduate Mathematics and Statistics (BSc MMathStat)
    - Year 3 of G1G3 Mathematics and Statistics (BSc MMathStat)
    - Year 4 of G1G3 Mathematics and Statistics (BSc MMathStat)