ST420-15 Statistical Learning and Big Data

24/25

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

Description

Introductory description

This module runs in Term 2 and is available for students on a course where it is a listed option (subject to restrictions*) and as an Unusual Option to students who have completed the prerequisite modules.

Pre-requisites

- ST218 Mathematical Statistics A; or,
- ST228 Mathematical Methods for Statistics and Probability and ST229 Probability for Mathematical Statistic.
- ST231 Linear Statistical Modelling with R.

MSc students:

- ST903 Statistical Methods and ST952 Introduction to Statistical Practice; or,
- MA907 Simulation and Machine Learning.

Non-statistics students.

- ST121 Statistical Laboratory and ST232/ST233 Introduction to Mathematical Statistics.
- ST231 Linear Statistical Modelling with R.

Module web page

Module aims

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.

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.

Statistical Learning – an introduction to statistical learning theory, using simple ML methods to illustrate the various ideas:
From over-fitting to apparently complex methods which can work well, such as VC dimension and shattering sets.
PAC bounds. Loss functions. Risk (in the learning theoretic sense) and posterior expected risk. Generalisation error.
Supervised, unsupervised and semi-supervised learning.
The use of distinct training, test and validation sets, particularly in the context of prediction problems.
The Bootstrap revisited. Bags of Little Bootstraps. Bootstrap aggregation. Boosting.
Big Data and Big Model – issues and (partial) solutions:
The “curse of dimensionality”. Multiple testing; voodoo correlations, false-discovery rate and family-wise error rate. Corrections: Bonferroni, Benjamini-Hochberg.
Sparsity and Regularisation. Variable selection; regression. Spike and slab priors. Ridge Regression. The Lasso. The Dantzig Selector.
Concentration of measure and related inferential issues.
MCMC in high dimensions – preconditioned Crank Nicholson; MALA, HMC. Preconditioning. Rates of convergence.

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

View reading list on Talis Aspire

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	Optional
Lectures	30 sessions of 1 hour (20%)	2 sessions of 1 hour
Private study	90 hours (60%)	
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.

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 D4

	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 number of words noted 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. 500 words is equivalent to one page of text, diagrams, formula or equations; your ST420 Assignment 1 should not exceed 15 pages in length.

Assignment 2	10%	15 hours	No

The deadline for the assignment can be found in the Statistics Assessment Handbook (http://warwick.ac.uk/STassessmenthandbook).
The assignment will contain a number of questions for which solutions and / or written responses will be required.
The number of words noted 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. 500 words is equivalent to one page of text, diagrams, formula or equations; your ST420 Assignment 2 should not exceed 15 pages in length.

In-person Examination	80%		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

Assessment group R4

	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

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 TMAA-G1PE Master of Advanced Study in Mathematical Sciences
- Year 1 of TMAA-G1PD Postgraduate Taught Interdisciplinary Mathematics (Diploma plus MSc)
- Year 1 of TMAA-G1PF Postgraduate Taught Mathematics of Systems
- TESA-H1B1 Postgraduate Taught Predictive Modelling and Scientific Computing
  - Year 1 of H1B1 Predictive Modelling and Scientific Computing
  - Year 2 of H1B1 Predictive Modelling and Scientific Computing
- Year 1 of TSTA-G4P1 Postgraduate Taught Statistics
- Year 4 of USTA-G300 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics

This module is Option list A for:

- Year 4 of USTA-G300 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics
- Year 5 of USTA-G301 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics (with Intercalated
- Year 4 of USTA-G1G3 Undergraduate Mathematics and Statistics (BSc MMathStat)
- USTA-G1G4 Undergraduate Mathematics and Statistics (BSc MMathStat) (with Intercalated Year)
  - Year 4 of G1G4 Mathematics and Statistics (BSc MMathStat) (with Intercalated Year)
  - Year 5 of G1G4 Mathematics and Statistics (BSc MMathStat) (with Intercalated Year)

This module is Option list B for:

- Year 4 of USTA-G304 Undergraduate Data Science (MSci)
- Year 4 of UCSA-G4G3 Undergraduate Discrete Mathematics
- Year 5 of UCSA-G4G4 Undergraduate Discrete Mathematics (with Intercalated Year)
- Year 3 of USTA-G1G3 Undergraduate Mathematics and Statistics (BSc MMathStat)

This module is Option list D for:

- Year 4 of USTA-G300 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics

This module is Option list E for:

- Year 4 of USTA-G300 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics
- Year 5 of USTA-G301 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics (with Intercalated