ST420-15 Statistical Learning and Big Data

20/21

Academic year
20/21
Department
Statistics
Level
Undergraduate Level 4
Module leader
Richard Everitt
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:
Statistics Undergraduate students: ST218 Mathematical Statistics A, ST219 Mathematical Statistics B and ST221 Linear Statistical Modelling.*
MSc in Statistics students: ST903 Statistical Methods and ST952 Introduction to Statistical Practice.*
Master's in Financial Mathematics students: MA907 Simulation and Machine Learning.
External Undergraduate students: ST220 Introduction to Mathematical Statistics and ST221 Linear Statistical Modelling.*

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

TBC

Transferable skills

TBC

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.

Assessment group D

	Weighting	Study time
Assignment 1	10%	15 hours

Due Term 2 Week 6.
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

	Weighting	Study time

Due Term 2 Week 9.
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.

2 hour examination (April) 80%

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

~Platforms - Moodle

Assessment group R

	Weighting	Study time
2 hour examination (September)	100%	

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

~Platforms - Moodle

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-G1P9 Postgraduate Taught Interdisciplinary Mathematics
- Year 1 of TMAA-G1P0 Postgraduate Taught Mathematics
- Year 1 of TMAA-G1PC Postgraduate Taught Mathematics (Diploma plus MSc)
- Year 1 of TSTA-G4P1 Postgraduate Taught Statistics
- USTA-G300 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics

    - Year 3 of G300 Mathematics, Operational Research, Statistics and Economics
    - Year 4 of G300 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
- 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)
- Year 4 of USTA-G1G4 Undergraduate Mathematics and Statistics (BSc MMathStat) (with Intercalated Year)

This module is Option list B for:

- Year 4 of USTA-G304 Undergraduate Data Science (MSci)

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