ST340-15 Programming for Data Science

23/24

Department
    Statistics
Level
    Undergraduate Level 3
Module leader
    Jeremie Houssineau
Credit value
    15
Module duration
    10 weeks
Assessment
    Multiple
Study location
    University of Warwick main campus, Coventry

Description

Introductory description

This module runs is available for students on a course where it is a listed option and as an Unusual Option to students who have completed the prerequisite module ST221 Linear Statistical Modelling or ST231 Linear Statistical Modelling with R .

There is a cap on student numbers for this module and pre-registration is essential. Information about prioritisation and the pre-registration form can be found at http://go.warwick.ac.uk/ST340

Module web page

Module aims

To introduce students to algorithms suitable to the analysis of large datasets. In the modern world it is very easy to generate very large amounts of data. Capturing and exploiting the important information contained within such datasets poses a number of statistical challenges. It may not even be clear how much useful information the data contains. The module will cover a variety of algorithms developed to tackle some of these challenges.

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. Computational Complexity.
2. Principal components analysis and singular value decomposition.
3. Markov chains and PageRank.
4. Clustering, EM algorithm.
5. Bandit problems.
6. Supervised learning, k-Nearest neighbours.
7. Supervised and unsupervised learning. Penalised regression.
8. Support vector machines.
9. Artificial neural networks.
10. Gaussian processes.
11. Parallel and distributed algorithms.

Learning outcomes

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

- Understand how to use a variety of practical algorithms when dealing with data analysis problems.
- Use R to implement data analysis algorithms.
- Interpret the output of various algorithms when applied to data sets.

Indicative reading list

[View reading list on Talis Aspire](#)

Subject specific skills

- Demonstrate facility with rigorousdata science methods.

- Evaluate, select and apply appropriate techniques to a variety of situations.

- Demonstrate knowledge of and facility with data science 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..

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	20 sessions of 1 hour (13%)	2 sessions of 1 hour
Practical classes	10 sessions of 1 hour (7%)	
Private study	46 hours (31%)	
Assessment	74 hours (49%)	
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 C4

	Weighting	Study time
Assignment 3	17%	25 hours

| | **Weighting** | **Study time** |

You will use R to implement and run algorithms on large datasets in response to a set of questions. You will present, discuss and evaluate the results. The number of words noted below 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 ST340 Assignment 3 should not exceed 25 pages in length.

| Assignment 1 | 16% | 24 hours |

You will analyse algorithms. You will use R to implement algorithms in response to a set of questions. You will present, discuss and evaluate the results. The number of words noted below 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 ST340 Assignment 1 should not exceed 24 pages in length.

| Assignment 2 | 17% | 25 hours |

You will use R to implement and run algorithms in response to a set of questions. You will present, discuss and evaluate the results. The number of words noted below 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 ST340 Assignment 2 should not exceed 25 pages in length.

| In-person Examination | 50% | |

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)

Assessment group R3

| | **Weighting** | **Study time** |

| Assignment | 50% | |

You will be asked to complete this assignment if you failed the module and you failed the coursework component of the original assessment. The reassessment will be similar in nature to the original assignments. 500 words is equivalent to one page of text, diagrams, formula or equations; your Assignment should not exceed 25 pages in length.

| In-person Examination - Resit | 50% | |

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

	Weighting	Study time

- Answerbook Pink (12 page)

Feedback on assessment

Marked assignments will be available for viewing at the support office within 20 working days of the submission deadline. Cohort level feedback will be provided, and students will be given the opportunity to receive feedback via face-to-face meetings.

Cohort level feedback will be provided for the examination.

[Past exam papers for ST340](#)

Availability

Courses

This module is Optional for:

- 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:

- 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)
- 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)
- USTA-GG14 Undergraduate Mathematics and Statistics (BSc)
    - Year 3 of GG14 Mathematics and Statistics
    - Year 3 of GG14 Mathematics and Statistics
- Year 4 of USTA-GG17 Undergraduate Mathematics and Statistics (with Intercalated Year)
- USTA-Y602 Undergraduate Mathematics,Operational Research,Statistics and Economics
    - Year 3 of Y602 Mathematics,Operational Research,Stats,Economics
    - Year 3 of Y602 Mathematics,Operational Research,Stats,Economics
- Year 4 of USTA-Y603 Undergraduate Mathematics,Operational Research,Statistics,Economics (with Intercalated Year)

This module is Option list B for:

- USTA-G302 Undergraduate Data Science
  - Year 3 of G302 Data Science
  - Year 3 of G302 Data Science
- Year 3 of USTA-G304 Undergraduate Data Science (MSci)
- Year 4 of USTA-G303 Undergraduate Data Science (with Intercalated Year)

This module is Option list D for:

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

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

This module is Option list F for:

- Year 3 of USTA-G300 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics
- USTA-G301 Undergraduate Master of Mathematics,Operational Research,Statistics and Economics (with Intercalated
  - Year 3 of G30H Master of Maths, Op.Res, Stats & Economics (Statistics with Mathematics Stream)
  - Year 4 of G30H Master of Maths, Op.Res, Stats & Economics (Statistics with Mathematics Stream)