Statistics and Machine Learning 1
Module Handbook

M. Muldoon, D. Perez Ruiz and K. Chen

November 1, 2023

1 About this Handbook

This is a ‘module handbook’ that supplements and complements the overall programme handbook for the Data Science MSc. The information here was correct on the date given above, but is subject to revision as the term unfolds.

Course materials and updates will be distributed electronically via BlackBoard, so you should check the site regularly. We will also monitor the module’s Discussion Board on Blackboard.

2 Class Meetings

The module will have both online and face-to-face components. There will be:

- pre-recorded lectures and guided reading that students will need to study in their own time;
- a 90 minute lecture/review: 9:00–10:30 on Thursdays in the main Lecture Theatre of the Samuel Alexander Building;
- lab sessions that will happen at least a week after the associated lecture materials have become available. These 90 minute sessions will happen twice per week
  - 12:00–13:30 on Thursdays in Room 1.12, Crawford House
  - 13:30–15:00 on Thursdays in Room 1.12, Crawford House

  The lecture/review session will take place in the Samuel Alexander Building (number 67 on the campus map available `here`) while the lab sessions will be in Crawford House (number 31 on the campus map). Students should plan to attend *only one* of the lab sessions per week.

Details about what we’ll do in these sessions and when the materials will become available appear in Section 5 below.

3 Module Staff

The people listed below should be your first point of contact for any questions about the module.

Mark Muldoon
Department of Mathematics
mark.muldoon@manchester.ac.uk
Lecturer and Module Lead

Diego Perez Ruiz
Department of Social Statistics
diego.perezruiz@manchester.ac.uk
Lecturer

Ke Chen
Department of Computer Science
ke.chen@manchester.ac.uk
Lecturer

4 Assessment

There are two components to the formal assessment.

4.1 Coursework (20%)

A coursework assignment (1000 words) will be published on BlackBoard early enough that you will have at least two weeks to prepare your submission, which you should upload to BlackBoard as a PDF.

Coursework Assignment

Published	*Due*	*Length*	*Credit*
31 October	20 November	1000 words	20%

4.2 Written Exam (80%)

There will be a two-hour exam held during the University's Winter exam period, 15–26 January 2024. We will provide you with guidance about what to prepare later in the term.

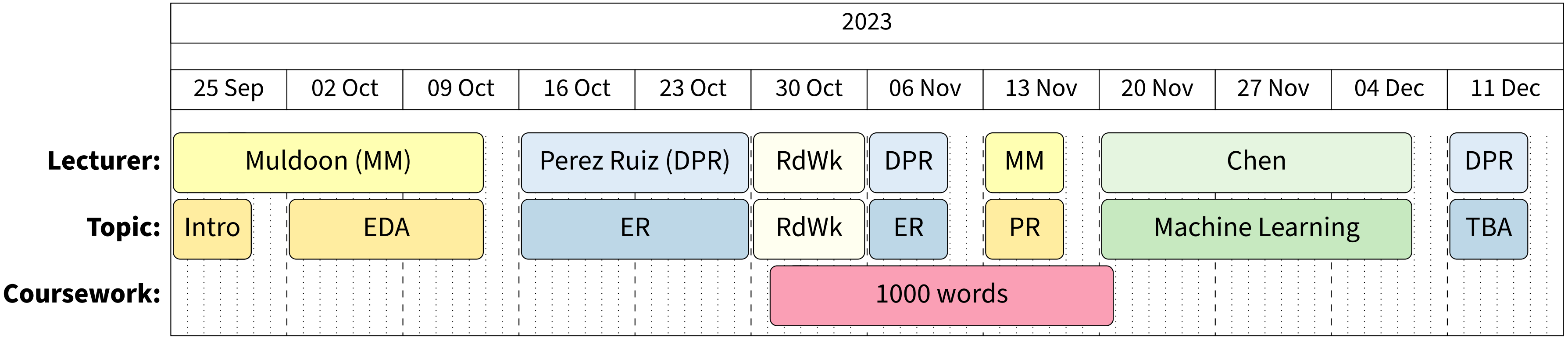

Figure 1: Timelines for the module: weeks are listed by the Monday with which they begin. **RdWk** stands for Reading Week—a one week period from 30 October–5 November during which there will not be any teaching—while, in the list of lecturers, **DPR** stands for Diego Perez Ruiz and **MM** stands for Mark Muldoon. In the list of topics, **EDA** stands for Exploratory Data Analysis, **ER** for Estimation & Regression, **PR** for Probabilistic Reasoning and **TBA** for To Be Arranged. Associated lecture materials will become available at least a week *before* a topic is listed so, for example, you should study the materials for the labs in the first week, which begins on Monday 25 September, during the week 18–24 September.

5 Lecture and Lab Schedule

Lecture and lab materials available by Monday 18 Sep.

Lecture 1 [Muldoon]: Randomness and Associated Algorithms.
9:00–10:30 Thurs. 28 Sep. Samuel Alexander Bldg, Lecture Theatre
Randomness as a way of representing unpredictable events. Definition of a random variable through its distribution function. Continuous and discrete random variables. Calculation of expectations. Commonly used parametric distributions. Computational generation of pseudo-random numbers. Data as a relatively small sample from a relatively large population. Data as a noisy observation of a process over time. A simple random model of data. First exposure to Monte Carlo methods for simulating data and bootstrapping.

Lab 1 [Muldoon]:
12:00–13:30 ***-or-*** *13:30–15:00*, Thurs. 28 Sep. Crawford House, Room 1.12

(1) No-credit, indicative online quiz covering basic material from probability and linear algebra.

(2) Generation of pseudo-random numbers.

(3) Estimation of π using Monte Carlo methods.

(4) Distribution of ages in census data—plotting distribution function and histogram, including bootstrap interval.

Lecture and lab materials available by Monday 25 Sep.

Lecture 2 [Muldoon]: Univariate Exploratory Data Analysis.
9:00–10:30 Thurs. 6 Oct. Samuel Alexander Bldg, Lecture Theatre
Introduction to the idea of Exploratory Data Analysis (EDA). Definition of lossy and lossless visualisations. Challenges of univariate data. Direct visualisation techniques including histograms, rug plots and ECDF. Sample expectations including mean, variance, skew and kurtosis. Estimation of modes using kernel density plotting. Order statistics. Implications of all techniques considered for what further analysis should be undertaken.

Lab 2 [Muldoon]:
12:00–13:30 ***-or-*** *13:30–15:00*, Thurs. 6 Oct. Crawford House, Room 1.12
Computational implementation of all techniques considered in Lecture 2 on real data.

Lecture and lab materials available by Monday 2 Oct.

Lecture 3 [Muldoon]: Multivariate EDA.
9:00–10:30 Thurs. 12 Oct. Samuel Alexander Bldg, Lecture Theatre
Challenges of multivariate data. Direct visualisation through scatter plots. Two-dimensional kernel density. Plot matrices. Calculation of correlation coefficients. Transforms including centering, standardisation, logarithms and Mahalanobis. Spaghetti plots for x-y data.

Lab 3 [Muldoon]:
12:00–13:30 ***-or-*** *13:30–15:00*, Thurs. 12 Oct. Crawford House, Room 1.12
Computational implementation of all techniques considered in Lecture 3 on real data.

Lecture and lab materials available by Monday 9 Oct.

Lecture 4 [Perez Ruiz]: Estimation.
9:00–10:30 Thurs. 19 Oct. Samuel Alexander Bldg, Lecture Theatre
Estimators, bias and consistency. Biased and unbiased estimation of population mean and variance from sample statistics. The maximum likelihood principle and its asymptotic unbiasedness and consistency. Confidence intervals for univariate maximum likelihood estimation based on a normal population distribution. Justification of this assumption using the central limit theorem.

Lab 4 [Perez Ruiz]:
12:00–13:30 ***-or-*** *13:30–15:00*, Thurs. 19 Oct. Crawford House, Room 1.12

(1) Simulate samples of univariate quantities of different sizes from different population distributions.

(2) Confirm the results about consistency, bias and confidence for these simulations.

(3) Apply methods to real univariate datasets where a normal distribution works (e.g. adult height) and where it does not (e.g. wealth).

Lecture and lab materials available by Monday 16 Oct.

Lecture 5 [Perez Ruiz]: Regression I.
9:00–10:30 Thurs. 26 Oct. Samuel Alexander Bldg, Lecture Theatre
Introduction to the concept of regression. Definition of linear regression. Derivation of least-squares fitting from a probability model and maximum likelihood. Generalised linear models including: logarithmic for strictly positive data; Poisson for count data; logistic for yes-no data. Standard measures of goodness-of-fit (e.g. R^2 for linear regression) and hypothesis testing using p values, together with critical discussion of these approaches.

Lab 5 [Perez Ruiz]:
12:00–13:30 ***-or-*** *13:30–15:00*, Thurs. 26 Oct. Crawford House, Room 1.12
Fitting and assessing fit using real data for linear, Poisson and logistic regression models.

Reading Week: *30 Oct.–5 Nov.*

Lecture and lab materials available by Monday 23 Oct.

Lecture 6 [Perez Ruiz]: Regression II.
9:00–10:30 Thurs. 9 Nov. Samuel Alexander Bldg, Lecture Theatre
Discussion of the ways in which reality fails to conform to the regression ideals. Standardisation of continuous covariates. Dummy variables for categorical covariates, as well as methods for ordinal integer variables. Polynomial, binned and interaction terms in regressions. Fixed and random effects, and an introduction to the concept of multilevel models.

Lab 6 [Perez Ruiz]:
12:00–13:30 ***-or-*** *13:30–15:00*, Thurs. 9 Nov. Crawford House, Room 1.12
Fitting regressions for more complex real data with integer covariates, strong non-linearity, interactions between covariates and random effects.

Lecture and lab materials available by Monday 6 Nov.

Lecture 7 [Muldoon]: Probabilistic Reasoning.
9:00–10:30 Thurs. 16 Nov. Samuel Alexander Bldg, Lecture Theatre
Conditional probabilities as the result of statistical prediction. Application to correlated multivariate data (e.g. if you know someone is tall, that gives information about their likely weight). Bayes' theorem. Conjugacy using the parametric distributions introduced in Lecture 1. Non-technical introduction to Gaussian process regression.

Lab 7 [Muldoon]:
12:00–13:30 ***-or-*** *13:30–15:00*, Thurs. 16 Nov. Crawford House, Room 1.12

- Probabilistic reasoning for the multivariate normal distribution applied to biometric data.
- Conjugate Bayesian inference for A/B testing.
- Gaussian process regression on a biological time series.

Lecture and lab materials available by Monday 13 Nov.

Lecture 8 [Chen]: Model Assessment and Selection I (concepts, motivation and empirical methods).
9:00–10:30 Thurs. 23 Nov. Samuel Alexander Bldg, Lecture Theatre
Key concepts — inductive bias/generalisation; overfitting vs. underfitting. Motivation — bias-variance trade-off; model assessment and selection. Held-out validation. Cross validation — k-fold and leave-one-out (LOO); practical issues regarding cross-validation; examples of cross-validation used in model assessment and selection.

Lab 8 [Chen]:
12:00–13:30 ***-or-*** *13:30–15:00*, Thurs. 23 Nov. Crawford House, Room 1.12

(1) Polynomial fits of increasing degree on x-y data for regression.

(2) Utilisation of polynomial models of different degrees trained on various samples of simulated data to understand key concepts such as overfitting, model complexity (flexibility) and the bias-variance trade-off.

(3) Trade-off between computational effort and accuracy for different cross-validation methods on a real dataset.

Lecture and lab materials available by Monday 20 Nov.

Lecture 9 [Chen]: Model Assessment and Selection II (analytic methods).
9:00–10:30 Thurs. 30 Nov. Samuel Alexander Bldg, Lecture Theatre
Background and general ideas behind the analytic methods. Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC) — general definition; procedure of applying AIC and BIC to model assessment and selection; Bayesian perspective on model selection. Specific AIC and BIC used in regression. Case study — feature (variable) selection in linear regression.

Lab 9 [Chen]:
12:00–13:30 ***-or-*** *13:30–15:00*, Thurs. 30 Nov. Crawford House, Room 1.12

(1) Comparison of AIC and BIC on simulated data with polynomial regression models as the model complexity and sample size vary.

(2) Apply AIC and BIC to feature (variable) selection in linear regression on a real dataset.

Lecture and lab materials available by Monday 27 Nov.

Lecture 10 [Chen]: Regularised Linear Models.
9:00–10:30 Thurs. 7 Dec. Samuel Alexander Bldg, Lecture Theatre
Background and Motivation behind regularisation. Different regularised linear models — ridge regression and the LASSO (loss functions, parameter estimation and hyperparameter tuning). Geometric interpretation of ridge regression and the LASSO. Variants of the LASSO — elastic net, bridge regression and group LASSO.

Lab 10 [Chen]:
12:00–13:30 ***-or-*** *13:30–15:00*, Thurs. 7 Dec. Crawford House, Room 1.12

(1) Investigation of linear regression, ridge regression and the LASSO on different simulated data regarding two situations — (a) the number of training examples (observations) is larger than that of features (variables) and (b) the number of training examples (observations) is comparable to (even smaller than) the number of features (variable).

(2) Apply the LASSO to feature (variable) selection in linear regression on a real dataset.

Lecture and lab materials available by Monday 6 Dec.

Lecture 11 [Perez Ruiz]: Special Topic.
9:00–10:30 Thurs. 14 Dec. Samuel Alexander Bldg, Lecture Theatre

Nonparametric methods:

- Introduction to the Bootstrap
- Linear Regression and Bootstrap
- Other types of Bootstraps
- Permutation Tests
- Smoothers

Lab 11 [Perez Ruiz]:
12:00–13:30 ***-or-*** *13:30–15:00*, Thurs. 14 Dec. Crawford House, Room 1.12
Practical application of ideas and tools from the lecture.