BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Machine Learning for Biostatistics (MLB)

Semester 2, 2025

Prepared by:

Andrew Grant
School of Public Health, The University of Sydney

THE UNIVERSITY OF SYDNEY

Contents

Machine Learning for Biostatistics (MLB)

Semester 2, 2025

Contact details

Andrew Grant

School of Public Health
University of Sydney

(02) 9351 9032

andrew.grant1@sydney.edu.au

If you have any **general BCA queries**, please contact the BCA Coordinating Office on 02 9562 5076/54 or email bca@sydney.edu.au

Background

Recent years have brought a rapid growth in the amount and complexity of data in biostatistical applications. Among others, data collected in imaging, genomics, and health registries, call for new statistical techniques in both predictive and descriptive learning. Statistical machine learning is a collection of algorithms and techniques for classification and prediction that complement classical statistical tools in the analysis of these data.

Context within the program

This unit covers other modern approaches to statistical modelling focusing on prediction. Students should already be familiar with principles of statistical inference, linear and logistic regression.

Prerequisites
Regression Modelling for Biostatistics 1 (RM1) or
Biostatistics: Statistical Modelling (PUBH5217)

Unit summary

This unit will introduce students to modern machine learning methods, particularly useful for large and complex data. Topics include classification trees, random forests, model selection, lasso, bootstrapping, cross-validation, generalised additive modelling, and regression splines. Some mathematical details will be covered but the primary

emphasis of the course will be on the intuition, implementation, and application of these methods. The statistical software package R will be used throughout the unit.

Workload requirements

The expected workload for this unit is 10-12 hours per week on average, consisting of guided readings, discussion posts, independent study, and completion of assessment tasks.

Learning Outcomes

At the completion of this unit students should be able to:

1. Recognise situations where machine learning methods can offer advantages over traditional statistical modelling approaches to data analyses in health applications

2. Recognise and explain the differences between the goals of description and prediction

3. Determine and implement appropriate machine learning approaches for description and prediction in real-world health applications

4. Measure and explain the uncertainty of the results of analyses using machine learning approaches

5. Interpret the results of analyses using machine learning in light of the assumptions required, the quality of input data, and the sensitivity to the specific technique implemented

6. Critically appraise published papers concerning machine learning applications for classification or prediction in health

7. Effectively communicate results of analyses in language suitable for a clinical or epidemiological journal

Unit content

The unit consists of 8 modules, summarised in more detail below. Each module is designed to take between 1 and 3 weeks to complete (see timetable below) and includes the following:

- An introductory video
- Slides used in the video
- Selected readings from the textbook
- Exercises - there will be several exercises at the end of each module that students should try to complete
- Optional tutorial – for every module the instructor will run an online session to discuss the solutions of the proposed exercises and other questions posed by the students. The schedule for these sessions will be agreed with the students.

The last module is an elective topic. The student will be required to research one additional topic of their choice.

All materials are accessed from the eLearning unit site. Assignments and supplementary material such as datasets will be available within each Assignment item. Please note that we are not able to post copies of copyright material (journal articles and book extracts)—for these you will have to rely on your home university's library.

Recommended approaches to study

Students should work through each module systematically, following the module notes and any readings referred to, and working through the accompanying exercises. *You will learn a lot more efficiently if you tackle the exercises systematically as you work through the notes.* You should also work through all the computational examples in the notes for yourself on your own computer.

Outline solutions to the exercises in each module (except those to be submitted for assessment, as described below) will be posted online at the midway point of the allocated time for the module. This is intended to encourage you to attack the exercises independently (or via the eLearning site), and yet not make you wait too long to see the sketch solutions.

Make the most of this unit by engaging with the instructor and fellow students on the Discussion Board and in online sessions. These are safe spaces to discuss the course material and related ideas and students are encouraged to make the most of them by engaging in respectful discussion.

Method of communication with coordinator(s)

We will use the BCA eLearning site as the main means of communication. We are using the Learning Management system, Canvas, hosted by the University of Sydney. You may be familiar with the system from previous BCA units and will receive any specific instructions on using the eLearning site this semester from the BCA Coordinating Office. Make sure that you check regularly for announcements.

For every module the instructor will run optional online sessions to discuss the solutions of the proposed exercises and other questions posed by the students. Solutions to the module exercises will be posted on Canvas. Assignments will be posted there too.

We will use the Discussion Board facility in Canvas. If you have a question or comment about the course material, post it to the relevant Discussion topic, where we can all access it and make a response. The instructor will generally let discussions flow between the students in each group, except where key points need resolution.

Questions about administrative aspects or course content can be emailed to the coordinator. Please use "MLB" in the Subject line of your email to assist in keeping track of our email messages. Please note that instructors are not necessarily available every day of the week and you should expect that it may take a day or so to respond to questions (possibly longer over weekends and during breaks). We strongly recommend that you post content-related questions to the Discussion Board in the

unit site. Questions about Assignments should be directed to the coordinator in the first instance

Module descriptions

Below is an outline of the study modules, followed by a timetable and assessment description table

Each module is scheduled to begin on a Monday and conclude on the Sunday of the relevant week.

Module 1: Introduction to Machine Learning	**Book Chapter**
• Basics of the R language	
• What is Machine Learning and Artificial Intelligence?	ISL 2.1 - 2.2
Module 2: Regression and Classification	
• Linear regression and KNN regression	ISL 3.1, 3.2, 3.3, 3.5
• Logistic regression and KNN	ISL 4.1 - 4.3
• Discriminant analysis	ISL 4.4 - 4.5
Module 3: Resampling Methods	
• Bootstrap and cross-validation	ISL 5.1 - 5.2
Module 4: Regularisation and model selection	
• Subset selection	ISL 6.1
• Ridge regression and LASSO	ISL 6.2
Module 5: Beyond linearity	
• Polynomial regression, step functions, basis functions	ISL 7.1 - 7.3
• Simple semiparametric models, additive models	ISL 7.4 - 7.7
Module 6: Beyond additivity	
• Classification and regression trees	ISL 8.1
• Bagging, random forests, boosting	ISL 8.2
Module 7: Unsupervised learning	
• PCA review, K-means and hierarchical clustering	ISL 12.1, 12.2, 12.4
Module 8: Elective topic	
In this module, the student will choose one topic not covered in the previous modules and research it. Examples of topics: support vector machines, neural networks, survival analysis with regularisation, survival trees.	Research

Unit schedule

Semester 2, 2025 starts on Monday 28th July

Week	Week commencing	Module	Topic	Assessment
1	28th July	Module 1	1. Introduction to Machine Learning	
2	4th August	Module 2	2. Regression and Classification	
3	11th August			
4	18th August	Module 3	3. Resampling methods	Practical Exercise 1 released 18th August
5	25th August	Module 4	4. Regularisation and model selection	
6	1st September			**Practical Exercise 1 due 1st September** Major Assignment 1 released 1st September
7	8th September	Module 5	5. Beyond linearity	
8	15th September			**Major Assignment 1 due 15th September**
9	22nd September	Module 6	6. Beyond additivity	
Mid-semester break 29th – 3rd October				
10	6th October	Module 6	6. Beyond additivity	
11	13th October			Practical Exercise 2 released 13th October
12	20th October	Module 7	7. Unsupervised learning	
13	27th October	Module 8	8. Elective topic	**Practical Exercise 2 due 27th October** Major Assignment 2 released 27th October
	3rd November			
	10th November			**Major Assignment 2 due 10th November**

Assessment

Assessment includes 2 major assignments worth 40% each and 2 practical exercises worth 10% each.

The assignments will be released approximately two weeks before their due date. Assignments will be posted online and an announcement made in Canvas.

Assignments are due by 11:59pm (AET) on the stated day.

Assessment name	Assessment type	Coverage	Learning objectives	Weight
Practical Exercise 1	Practical Exercises	Modules 1-3	1, 2, 3, 4, 5, 6, 7	10%
Major Assignment 1	Assignment	Modules 1-4	1, 2, 3, 4, 5, 6, 7	40%
Practical Exercise 2	Practical Exercises	Modules 1-6	1, 2, 3, 4, 5, 6, 7	10%
Major Assignment 2	Assignment	Modules 1-8	1, 2, 3, 4, 5, 6, 7	40%

The assessment tasks will require submission of a mix of written responses, coding scripts, and pre-recorded presentations.

In general, you are required to submit written work typed in Word or similar. We strongly recommend you become familiar with equation typesetting software such as Microsoft's Equation Editor for algebraic work. You may submit neatly handwritten work, however please note that marks will potentially be lost if the solution cannot be understood by the markers due to unclear or illegible writing. Handwritten work should be scanned and collated into a single pdf file and submitted via the eLearning site. See the BCA Assessment Guide for guidelines on acceptable standards for assessable work.

All material submitted for assessment must be entirely your own work. Please see the note on 'Academic Dishonesty and Plagiarism' in the Assessment Guide.

It is strongly suggested that you keep a copy of your assignments.

Students are encouraged to discuss relevant topics in the Discussion Board. However, please avoid posting questions relating directly to assessable material. These should be emailed to the Unit Coordinator in the first instance.

Explicit solutions to assessable exercises should not be posted for others to use. Each student's submitted work must be clearly their own, with anything derived from other students' discussion contributions clearly attributed to the source.

Submission and academic honesty policy

All assessment material should be submitted via the relevant Assessment module in Canvas unless otherwise advised. Turnitin plagiarism detection is applied to all submissions. For detailed information, please see the BCA Assessment Guide, which includes links to the Academic Honesty policies at member universities. Please familiarise yourself with the procedures and policies at your home university. You will

need to indicate your compliance with the plagiarism guidelines and policy at your home university.

A special note regarding "contract cheating" sites: Unfortunately, there have been instances in the past of students using such websites to post assignment questions and receive solutions (usually for a fee). We have arrangements with these sites to identify the student posting questions or accessing the solutions, and such students will be referred to and face disciplinary processes at their home university.

Use of ChatGPT and other generative AI tools in assessment tasks

The assessment tasks in this Unit have been designed to be challenging, authentic and complex. Although individual assessment components may provide specific guidance regarding the use of generative AI tools (e.g., ChatGPT), successful completion of these components will require students to critically engage in specific contexts and tasks for which artificial intelligence will provide only limited support and guidance. In all cases, a failure to reference the use of generative AI may constitute student misconduct under the Student Code of Conduct of your University of enrolment. To successfully complete assessment tasks, students will be required to demonstrate detailed comprehension of their written submission independent of AI tools.

Late submission and extension procedure

The standard BCA policy for late penalties for submitted work is a 5% deduction from the earned mark for each day the assessment is late, up to a maximum of 10 days (including weekends and public holidays). Extensions are possible, but these need to be applied for (by email) as early as possible. The Unit Coordinator can approve extensions up to three days; for extensions beyond three days, you must apply to your home university via their standard procedures.

Learning resources

Course notes/slides, tutorial exercises, and data sets will be posted on the course website. The textbook **required** for the unit is:

> *Gareth James, Daniella Witten, Trevor Hastie, Robert Tibshirani. "An Introduction to Statistical Learning with Applications in R", 2nd Edition, Springer Texts in Statistics. Electronic copy **available for free** at:*
> *https://www.statlearning.com*

For supplementary reading, a more in-depth treatment of similar material is provided in:

> *Trevor Hastie, Robert Tibshirani, Jerome Friedman. "Elements of Statistical Learning", Springer Texts in Statistics. Electronic copy **available for free** at: https://web.stanford.edu/~hastie/ElemStatLearn/download.html*

Software requirements and assumed knowledge

We will be using the statistical package R throughout the unit. For help with R, please see Learning R in the Student Resources site.

Required mathematical background

The unit is mainly computational and basic mathematical knowledge is assumed.

Feedback

Our feedback to you:

The types of feedback you can expect to receive in this unit are:

- Formal individual feedback on submitted exercises assignments
- Discussion during the online sessions
- Responses to questions posted on the Discussion Board

Your feedback to us:

One of the formal ways students provide feedback on teaching and their learning experience is through the BCA student evaluation survey at the end of each semester. The feedback is anonymous and provides the BCA with evidence of aspects that students are satisfied with and areas for improvement.

Unit changes, including response to recent student evaluation

MLB was last delivered in Semester 2, 2024. The feedback of the students was very positive, and no major changes were implemented for the current semester.

Acknowledgments

The course materials have been developed by Prof Armando Teixeira-Pinto with contributions from Prof Jarek Harezlak (Indiana University), Dr. Shuvo Bakar, and Associate Prof Tim Schlub.