BIO STATISTICS COLLABORATION OF AUSTRALIA

Study Guide

Machine Learning for Biostatistics (MLB)

Semestre 2, 2020

Prepared by:

Armando Teixeira-Pinto
School of Public Health, The University of Sydney

Copyright © School of Public Health, The University of Sydney

Machine Learning for Biostatistics (MLB)

Study Guide

Contents

Unit Overview

Recent years have brought a rapid growth in the amount and complexity of data in biostatistical applications. Among others, data collected in imaging, genomic, 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. 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 R package will be used throughout the unit.

Unit Objectives

At the completion of this unit you 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

Assumed Knowledge

Students should already be familiar with principles of statistical inference, linear and logistic regression.

Unit Structure

This unit is offered throughout Australia through the Biostatistics Collaboration of Australia. It is available in distance learning mode only, to students enrolled in postgraduate degrees in biostatistics coordinated by the BCA.

The unit consists of 8 modules comprising several topics. Each module is designed to take 1 or 2 weeks to complete (see timetable below). Each module comprises an introductory video, slides, study guide and the indication of readings from the textbook.

The last module is an elective topic. The student will be encouraged to research one additional topic to his/her choice.

There will be several exercises at the end of each module that students should complete.

Week begins	Aug 3	Aug 10	Aug 17	Aug 24	Aug 31	Sep 7	Sep 14	Sep 21	Sep 28	Oct 5 BREAK	Oct 12	Oct 19	Oct 26	Nov 2	Nov 9	Nov 16
Module																
Wk	1	2	3	4	5	6	7	8	9		10	11	12	13	14	15
1. Introduction to Machine Learning										B						
2. Regression and Classification										R						
3. Resampling methods										E						
4. Regularisation and model selection										A						
5. Beyond linearity										K						
6. Beyond additivity																
7. Unsupervised learning																
8. Elective topic																
Assignment released																
Assignment due																

Numbers in the shaded cells correspond to topic numbers, as shown in the table below.

Module descriptions

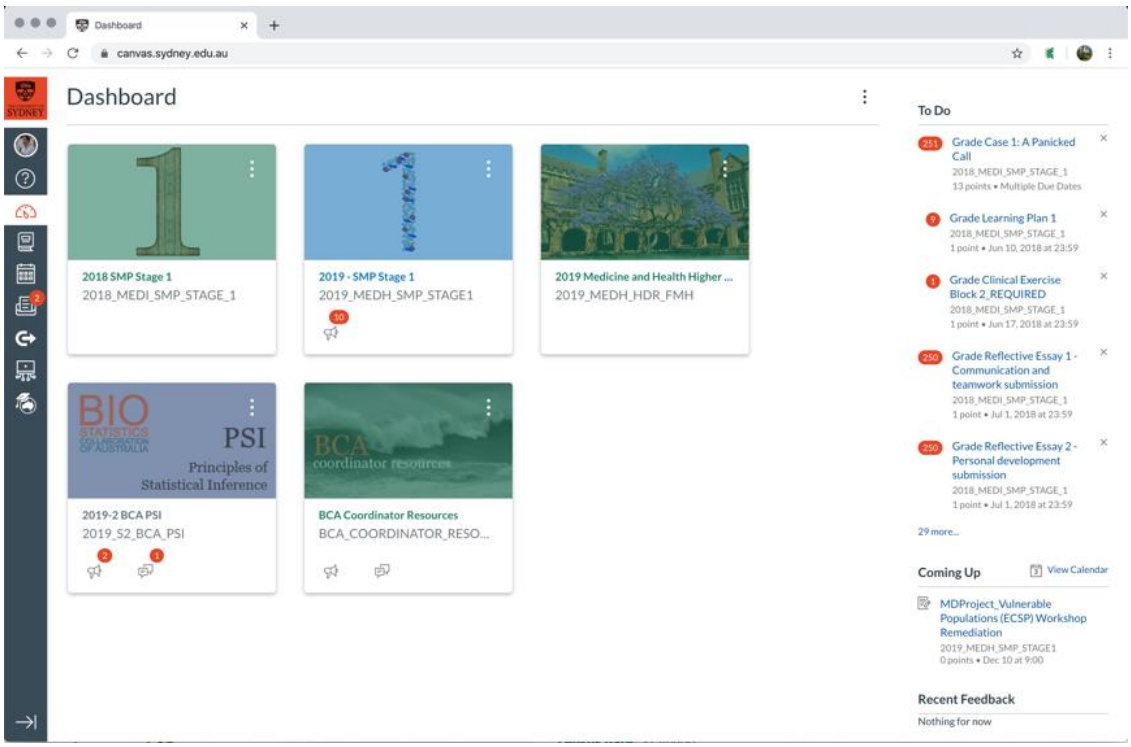
Module	Week		Book chapter
1	1	**Introduction to Machine Learning**	
		Basics of R language	Notes
		What is Machine Learning and Artificial Intelligence?	ISL 2.1-2.2
2	2-3	**Regression and Classification**	
		Linear regression and KNN regression	ISL 3.1, 3.2, 3.3, 3.5
		Logistic regression and KNN	ISL 4.1-4.3
		Discriminant analysis	ISL 4.4-4.5
3	4	**Resampling methods**	
		Bootstrap and Cross-validation	ISL 5.1-5.2
4	5-6	**Regularisation and model selection**	
		Subset selection	ISL 6.1
		Ridge regression and LASSO	ISL 6.2
5	7-8	**Beyond linearity**	
		Polynomial regression, step functions, basis	ISL 7.1-7.3
		Simple Semiparametric Models, Additive	ISL 7.4-7.7
6	9-10	**Beyond additivity**	
		Classification and regression trees	ISL 8.1
		Bagging, Random Forests, Boosting	ISL 8.2
7	11	**Unsupervised learning**	
		PCA review, K-means and hierarchical clustering	ISL 10.1-10.3
8	12-13	**Elective topic**	
		In this module, the student can choose one topic not covered in the previous modules and research it. Examples of topics: Neural networks, ensemble learning, Adaboost, support vector machines.	Research

Methods of Communication

Online eLearning

We will use the BCA eLearning site as the main means of communication:

http://elearning.sydney.edu.au



Make sure that you have the **correct email registered** in the eLearning platforms or you may miss important announcements.

An eLearning Guide, which gives basic information on how to use online eLearning is available from the Student Resources page http://www.bca.edu.au/currentstudents.html on the BCA website.

Solutions to the weekly exercises will be posted on the eLearning site. Assignments will be posted there too.

We will use the Discussions/Forum facility on 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 seem to need resolution.
- Any general Discussion items or questions, in particular on the study guide and notes, can be posted to the other Discussion areas.

About online discussions

Discussions form an important part of your learning and your assessment. Discussions are really quite similar to face-to-face tutorials, except that your discussion is in written rather than spoken form, and you can't see those you are talking to (in fact, you may never see them). Some things to think about:

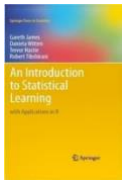
- Students will be invited by the coordinator, in a random fashion, to lead the discussion on specific exercises and share their solutions in the discussion board
- **Everybody's ideas and contributions are valuable**. We can all learn from each other's experience and insights. Don't be shy about contributing your ideas. The more ideas you contribute, the richer the discussion will be. And don't be afraid to be the first to contribute!
- **Your relationship with others in your group**: Make sure you contribute to, and read the postings in, the **Introductions** blog. Maintain good relations with the others by observing netiquette - avoiding overt criticism, flaming etc and being very careful with humour. When you can't see each other, it's easy to misunderstand something that's perhaps awkwardly worded. Learning is easier if everyone gets on well.
- **Don't be afraid to ask questions**. There may be someone else in the group wondering about just the same thing that puzzled you!
- **Interact** with the others in your group, just as you would face-to-face. Agree, disagree (politely of course, and giving reasons for your opinion). Ask for clarification, add ideas - all of this makes the discussion more interesting and worthwhile.
- **Check in often**: Get into the habit of accessing eLearning regularly and checking the 'Discussions' icon, to see if there have been any new postings. It's much easier to keep up if you check-in regularly.
- **Readings**: Be sure to do the required reading before you start the discussion, so that you can make a meaningful contribution.
- **Length of contributions**: We're not looking for assignment-length postings to discussions! We might indicate how much is needed, but if not, generally just one or two well-written paragraphs will be enough, or even one or two sentences in some cases. This is much kinder to the others in your group, who need to read what you've written or to have a chance to make their own contribution.
- **Getting it right**: You might like to create your posting in a word-processing program and check spelling and grammar before you post. Or type your contribution directly into eLearning and preview the message to have a look at it before you post it.
- **Not happy with your posting?** If you've posted something that you're not happy with, **you won't be able to remove it** – you'll need to ask your tutor to arrange this for you. It's better to make sure your posting is OK before you post it.
- **Adding an attachment**: You can add an attachment created in any program to your posting, but be aware that, if the people who are supposed to read the attachment don't have the same program on their computers, they won't be able to read it.

Email and Phone

If you have any questions during the semester, please email the unit coordinator, A/Prof Armando Teixeira-Pinto (armando.teixeira-pinto@sydney.edu.au) or call 02 9351 4369.

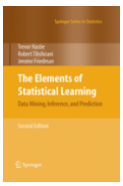
Textbook and course notes

Course notes/slides, homework assignments, and data sets will be posted on the course website. We will follow closely the textbook. The textbook **required** for the unit is:



*Gareth James, Daniella Witten, Trevor Hastie, Robert Tibshirani. "An Introduction to Statistical Learning", Springer Texts in Statistics. Electronic copy **available for free** at:*

http://www-bcf.usc.edu/~gareth/ISL/



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:*

http://statweb.stanford.edu/~tibs/ElemStatLearn/

Software

We will be using the statistical package R throughout the unit

Assessment

Assessment will be by 2 main assignments, each one counting 40% to the final mark, and 2 sets of practical exercises worth 10% each.

The main assignments are to be completed within approximately two weeks after their release. Assignments will be posted online and an email will be sent to you when the assignment is posted.

Please consult the document 'BCA Assessment Guide'[1] for details about submitting your assignments, and guidelines for written work. Also, read the "BCA Guide for reporting statistical results"[2]

All material submitted for assessment must be entirely your own work. Please see the note on 'Academic Dishonesty and Plagiarism' on pages 3-4 of the BCA Assessment Guide.

When you submit your assignment online, you will be required to complete a declaration, in the form of one either/or test question, certifying that you have read and understood the Academic Dishonesty and Plagiarism policy at the university in which you are enrolled. The assignment should not appear on the page until you have done this. This procedure is a compulsory requirement of all universities. See page 3 of the BCA Assessment Guide for more details.

1 http://www.bca.edu.au/linked%20docs/Student%20resources/BCA_assessment_guide_student.pdf

2 http://www.bca.edu.au/linked%20docs/Student%20resources/BCA%20Presenting%20Statistical%20Information%20Guide.pdf

If you don't submit your assignment via eLearning, you will need to complete the BCA Assignment Cover Sheet[3]

Assignments should preferably be submitted online. If this proves difficult then send by email as an attachment to armando.teixeira-pinto@sydney.edu.au

I strongly suggest that you keep a copy of your assignments.

It is planned that the assignments will be released and due as follows (see also Timetable) at the **end of each day**:

Assignment 1 will be released Monday 7th September, due Monday 21st September

Assignment 2 will be released Monday 2nd November, due 16th November

Extensions or late submissions policy

For various reasons, you may sometimes experience difficulties in getting your assignment submitted on the due date. Requests for an extension of the due date for an assignment **must** be made **in advance of the due date for that assignment**. The normal grounds for an extension being granted are bereavement, personal illness or illness in a family member requiring you to exercise a significant carer role. This request must be made directly to the unit coordinator by email. The unit coordinator will reply by email with the decision as to whether an extension has been granted and the new due date. Extensions will normally be no longer than three days.

Where a student is so incapacitated by a medical or other condition that he or she is unable to request an extension in advance, medical or other certification should explicitly note the severity of the disabling condition that precluded the advance request being made.

Late penalty

If no extension has been given, 5% of the earned mark for an assignment will be deducted for each day that an assignment is late, up to a maximum of 50%.

NOTE: It is not the intention of this late penalty policy to cause a student to fail the unit when otherwise they would have passed. If deductions for late assignments result in the final unit mark for a student being less than 50, when otherwise it would have been 50 or greater, the student's final mark will be exactly 50.

Contact details

For **enquiries about this unit**, contact the unit coordinator:

Dr Armando Teixeira-Pinto, A27, University of Sydney, NSW 2006
phone 02 9351 5424 email: armando.teixeira-pinto@sydney.edu.au
fax 02 9351 5049

In case of illness or extended absence of the unit coordinator, the deputy coordinator is:

3 http://www.bca.edu.au/Linked%20docs/Student%20resources/BCA_assignment_exam_cover_form.pdf

Patrick Kelly, School of Public Health, A27, University of Sydney, NSW 2006
phone 02 9351 4648 email: p.kelly@sydney.edu.au
fax 02 9351 5049

For **enquiries about receipt of assignments**, contact:

Biostatistics Administrative Officer, School of Public Health, University of Sydney
phone 02 9351 5994 email: sph.bsta@sydney.edu.au
fax 02 9351 5049

For **enquiries about the BCA** and about the various degrees towards which this unit contributes, contact the BCA Executive Officer:

Karolina Kulczynska-Le Breton, University of Sydney, NSW 2006
phone 02 9562 5076 email: karolina.kulczynska-lebreton@sydney.edu.au
mob 0430469669

For **enquiries about your degree program**, contact the university through which you are enrolled.