Statistical Espresso for Biologists
Can't stop the signal: machine learning, big data & summarizing variation
BIOL–6750–08 — Fall 2017

Instructor: William D. Pearse

1 Course description and learning objectives

There are two kinds of "problem" in statistics that everyone wishes they had: so many observations that you don't know what's going on (many rows), and so many variables you don't know what's important (many columns). In both cases it can be difficult to even see what's going on in your data, let alone statistically model it. By the end of this course, you will:

- Understand the basic principles of machine learning and management of 'big data'
- Be able to use principal components analysis, ordination, and cluster analysis to identify groups and associations among variables
- Be able to use regression trees and least angle regressions to summarize complex datasets
- Be able to manipulate, visualize, and perform regressions on large (>10Gb) datasets in `R` (*e.g.*, all known global observations of plants)

The class emphasizes both theoretical concepts and practical applications equally: students are encouraged to 'follow-along' with the examples given in the lecture in the `R` statistical environment. This will give you the practical skills you need to make use of the statistical approaches covered in the class in your own work.

As a graduate student, much of your grade will be determined by a written project, using the skills you develop in the class to investigate a dataset of your choosing. Through one-on-one meetings with the instructor during and after the class, you will learn to apply the concepts you practiced in the classroom in an entirely novel setting. This is an excellent opportunity to get help with your thesis...

2 Course materials

No text-books are required for this course, but a laptop computer (Windows, Mac, or Linux) with RStudio version $\geq$ 1.0.143 (using `R` version $\geq$ 3.4.0) and SQLite version $\geq$ 3.19.3. Advice and help getting these tools installed will be provided ahead of the start of class.

3 Requirements

Basic statistical classes (*i.e.*, knowledge of how to perform a t-test and simple linear regression) are required for this class. No prior experience with `R` whatsoever is required, as the class will focus entirely on how to interpret `R` output and students will have no need to perform any programming in the class. Students who are concerned about their level of `R` experience are encouraged to study the USU `baseR` website (`http://learnr.usu.edu/`), which covers more detail than is required for this course. The starting statistical requirements for this course are very low, but, nevertheless, students who are concerned are encouraged to contact Dr Pearse directly (`will.pearse@usu.edu`).

4 Attendance & participation

Attendance of all lectures is compulsory; we are covering a lot of ground in a compressed period of time and students who miss one session may not be able to catch back up. Students who miss more than one session may be dropped

from the course, unless they can give a good reason (with evidence if necessary) for missing that session.

5 Assessment

You will be assessed through a combination of coursework and a short exam delivered a few weeks after the end of the class. Coursework will be started in the classroom, and will consist of completing an analysis of one or more datasets on which the instructor will give guidance, and then analyzing and interpreting additional dataset(s) in a similar vein. The exam will be short, and test the retention of key statistical concepts—it will not require detailed calculations or rote memorization of arcane statistical facts. Example questions (and answers) will be given to the students ahead of the exam.

In addition to this, graduate students will be required to complete a short paper describing the use of the statistical techniques developed within this class in a novel dataset. Graduate students are *strongly* encouraged to use this as an opportunity to develop one of their thesis chapters. The marking breakdown for the undergraduate 4XXX series, and graduate 6XXX series, classes are given below:

Section	4XXX %	6XXX %
Coursework	55	27.5
Exam	35	17.5
Paper	0	45
Participation	10	10

The "Honor System" of Utah State University applies to the coursework and paper. I strongly advise you *not* to copy-paste answers to exercises from the Internet or from the work of other students. Not only are solutions from outside sources (and other students) often wrong, they are also quite simple for me to detect and those who cheat will be punished in accordance with Utah State University regulations.

6 Schedule

The course is split into three main sections. In the first (weeks 1–2), you will learn the classical concepts of data summary and basic commands in the `R` statistical language, and in the second (weeks 3–4) we will learn about the most common kinds of modern machine learning tools. In the third and final section (week 5), we will review modern data manipulation and storage techniques, and learn about best-practices with data. No coursework will be due until the end of the class, but students are *strongly* encouraged to complete exercises as the class proceeds.

Week	Title	1st session	2nd session
1	**Classical tools**	PCA	factor analysis
2	**Clustering**	hierarchical	k-means
3	**Machine Learning 1**	regression trees	least angle regression
4	**Machine Learning 2**	support vectors	artificial neural networks
5	**Data & synthesis**	`sql(ite)`	pipelines

Each week follows the same format: two sessions, each comprised of one 40-minute lecture and a 40-minute session to work on coursework and practical examples. Each lecture will give equal weight to core theoretical concepts and practical demonstrations of those within `R`: this is not a theoretical class, it is a practical class where you will pick up skills and experience. It is likely that students will want to have whatever they use for note-taking to hand and their laptops to try things with the instructor in real-time in `R`. The coursework practicals will be an opportunity for students to continue working through the examples given in the preceding lectures, and also to begin work on their coursework exercises and receive feedback and help with those exercises.

7 Miscellanea

- ADA compliance: Students with physical, sensory, emotional or medical impairments may be eligible for reasonable accommodations in accordance with the Americans with Disabilities Act and Section 504 of the Rehabilitation Act of 1973. All accommodations are coordinated through the Disability Resource Center in Room 101 of the University Inn, 797-2444 voice, 797-0740 TTY, or toll free at 1-800-259-2966. Please contact the DRC as early in the semester as possible. Alternate format materials (Braille, large print or digital) are available with advance notice.
- Sexual harassment is defined by the Affirmative Action/Equal Employment Opportunity Commission as any "unwelcome sexual advances, requests for sexual favors, and other verbal or physical conduct of a sexual nature." If you feel you are a victim of sexual harassment, you may talk to or file a complaint with the Affirmative Action/Equal Employment Opportunity Office located in Old Main, Room 161, or call the AA/EEO Office at 797-1266.
- Students whose religious activities conflict with the class schedule should contact me at the beginning of the semester to make alternative arrangements.
- Course scheduling and structure may change at short notice with no warning.