Modern BioStatistics and Data Mining

Instructor: Prof. Rafal Kustra

Lecture: 2h once per week

Lab: 1h roughly every other week

Objective: Introduce students to the statistical methods suitable for analysing large observational data, data constructed from multiple institutional databases, web-based data, and any data that may benefit from non-classical approaches. The theory will be presented as an extension of classical tools such as linear and logistic regression, parametric hypothesis testing, multivariate Gaussian theory, to make it more intuitive and accessible. Evaluation will comprise of theoretical exercises and practical data projects. We will be using R statistical language with appropriate packages. At the end of the course students should be aware of:

1. distinction between, and application of, supervised and unsupervised statistical learning problems,
2. classification problems, similarities between classifiers and regression models
3. non-classical regression and classification tools: loess and spline smoothing, tree-based methods, and kernel-based methods,
4. Importance and implementation of prediction error control in statistical modelling using v-fold cross-validation and leave-one-out bootstrap
5. Importance of, and tools for, complex data handling, testing, and manipulation

Textbook: Hastie T, Tibshirani R, and Friedman J. (2011). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction - Second Edition*. Springer New York.

(book available online at Prof Robert Tibshirani's web page at Stanford)

Format: Roughly 1.5h lecture combined with 1h discussion of assigned reading. Lab hour every other week will consist of simple hands-on exercises

Prerequisites: undergraduate-level courses: mathematical statistics, linear regression, linear algebra

Topics:

1. Data Mining, Big Data and BioStatistics: what's common and different?
2. Supervised vs Unsupervised problems in statistical learning (Chapter 1 and 2)
3. Review of linear regression as a setup for Supervised Learning. Basis expansion and regularization (Week 3: parts of chapter 3 and 5)
4. Classification problem and two classical methods (LDA and logistic regression) (Week 4 and 5, Chap 4)
5. Kernel-based methods and Support Vector Machine (week 6: parts of Chap 12)
6. Prediction error and its estimates (including quick Bootstrap introduction) (Week 7 and 8, Chap 7)

7. Decision Trees and Ensemble models (week 9 and 10)
8. Neural Networks and Deep Learning systems (week 11)
9. Data clustering tools and uses (Week 12 Section 3.6, Parts of Chap 14)
10. Data Checking and Preparation for Analysis: Iterative Quality Control (week 13, Intructor notes)

Evaluation:

	Covers	**Due/Date**	**Weight**
Assignment 1	Topics 1,2,3	Mar 3, 2016	25%
Midterm	Topics 1,2,3,4	Mar 17, 2016	25%
Assignment 3	Topics 4,5,6	Apr 7, 2016	25%
Paper presentation		During regular sessions. Presentation date assigned when topics assigned (by week 3)	15%
Participation	Being prepared with assigned reading, in-class activity, attendance		10%

Assignments: These will be roughly 5 questions each with a mixture of mathematical/methodological questions and applied/computational questions where discussed methods will be needed to be applied in R on a provided dataset and results briefly discussed.

Paper presentation: Groups of 2-3 students will be assigned a paper/handout to read covering some aspects/extensions of a topic from the list above and they will be required to present it at appropriate date.

Late Assignment Policy: Assignments are due by 4pm the due day. The late penalty is 5% per day. Assignments handed in later than 4 days after the due day will not be accepted. For the purposes of this policy day ends at 4pm and weekend and holiday days count. So Assignment due Mar 11, handed in after 4pm on March 12, will be assessed 10%. On weekends and days when University is closed, late assignments may only be submitted electronically but the responsibility for ensuring proper delivery and my ability to open it rests with the student.

Important note on academic integrity:

Students in graduate studies are expected to be familiar with the University's policies on academic integrity and commit to the highest standards of academic practice. This includes understanding the importance of protecting and acknowledging intellectual property. Students are expected to know how to cite references appropriately, thereby avoiding plagiarism. Please refer to following documents for guidance:
How Not to Plagiarize:
http://www.writing.utoronto.ca/advice/using-sources/how-not-to-plagiarize
The Code on Behavior and Academic Matters:
http://www.governingcouncil.utoronto.ca/Assets/Governing+Council+Digital+Assets/Policies/PDF/ppjun011995.pdf.

Plagiarism **(the presentation or paraphrasing of the work of another author as if it was one's own) is a form of academic fraud with potentially serious consequences. All university policies regarding plagiarism will be upheld in this course. The instructor reserves the right to submit student papers to Turnitin.com, a computer-based service which checks for originality in submitted papers. Students who do not wish their papers to be submitted to Turnitin.com should keep all draft versions of their work for review by the instructor and the TA, if needed.**

Student Presentation Evaluation

Date:	**Evaluator:**	
Presenter:		
Criteria	**Scores**	**Comments**
Content and Scientific Merits	/ 60	
Introduction: • Provides sufficient background information and rationale of research • States clearly the objectives and relevant questions Methodology: • Demonstrates knowledge in the subject areas, and includes sufficient technical details and pertinent examples • Responds adequately to methodological questions with explanations and elaborations Conclusion/Discussion: • Summarizes the main conclusions and ideas with evidence • Discusses potential limitations or weaknesses		
Delivery	*/20*	
• Speaks clearly at a steady pace • Follows a clear and logical sequence for the presentation, and uses language appropriate for the intended audience • Captures and maintains the audience’s attention • Handles questions professionally • Observes the time limits		
Audio/Visual	*/20*	
• Provides clear, legible and concise presentation slides • Utilizes appropriately figures and tables • Uses effectively multimedia materials		
Overall Score	**/100**	

General Comments: