MSC1090 “Introduction to Computational BioStatistics with R”

SciNet HPC Consortium

September, 2018

Course Details

Objective

The goal of this class is to prepare graduate students to perform scientific data analysis. Successful students will learn how to use statistical inference tools to gain insight into their data, as well as be exposed to cutting-edge techniques and best practises to store, manage and analyze data.

Instructors:

Dr. Marcelo Ponce and Dr. Erik Spence (SciNet, Advanced Research Computing at the University of Toronto).

Proposed Dates:

Fall term, 2018.

Tuesdays and Thursdays: 1pm-2pm

Location:

MSB4279

Structure & Enrollment:

* Twelve weeks, with two 1-hour lectures per week.

* Final grades will be based on 10-12 approximately-weekly assignments and *a mid-term*. There will be no final exams.

* Passing mark: 70% of the final grade.

Grading Scheme

Most weeks, students will be given a programming assignment, with a due date one week after. These assignments are designed to help absorb the course material.

There will be between 10 and 12 assignments. The average of the assignments and the midterm will make up the grade. To ensure a timely reporting of student grades, we will adhere to the following policy:

Homework may be submitted up to one week after the due date, at a penalty of 0.5 point per day, out of the 10 points for each homework. Deviations of this rule will only be considered, on a case-by-case basis, in exceptional circumstances.

All sets of homework need to be handed in and the mid-term has to be taken for a passing grade, which is based on the average of the total sets. If, due to exceptional circumstances, an assignment was missed, a make-up assignment can be given at the end of the course. Rather than focusing on the topic of a specific week, the make-up assignment may involve any of the material of the course.

Syllabus: "Introduction to Computational BioStatistics with R"

Week 1: Introduction to bash programming. File manipulation, regular expressions, bash scripting, automation of data-analysis pipelines.

Week 2: Introduction to programming with R. IDEs and R standard console, basic programming concepts: conditionals, loops, variable types.

Week 3: Introduction to programming best practices in R. Functions and scripts, interactive versus batch processing, variable scope, modular programming, defensive programming, comments and documentation.

Week 4a: Introduction to version control. Motivation, implementation and use of version control, logs, rollback, branches.

Week 4b: Binary file input/output. Accessing, reading and writing binary data, file input and output strategies, and best practises.

Week 5 – 6: Basics statistics using R. Review of the basic concepts of probability and statistics, probability distributions, descriptive and inference statistics. Statistical modeling implementations using R: linear models, quadratic models, generalized linear model. Testing models: correlation and covariance, Pearson coefficient. Hypothesis testing: examples of null hypothesis tests implementations, T-Student test, two sample T-test, matched pair experiments, independence tests, ANOVA-like based tests. Model Diagnostics: graphical tools, leverage, influential points, Cook's distance, residuals, validation of assumptions. Clustering algorithms and decision trees examples.

Week 7: Introduction to machine learning. Regression, overfitting, bias-variance tradeoff, cross-validation, bootstrapping, LOESS, LOWESS.

Week 8: Advanced machine learning. Variable selection, dimensionality reduction, principal component analysis.

Week 9: Classification algorithms. Decision trees, confusion matrices, clustering, logistic regression, Naive Bayes.

Week 10: Visualization of data. Publication-quality figures, basic plotting, 1D (curves), 2D (contour maps) and 3D plots, interactive visualization, animations.

Week 11: Advance statistical topics: Generalized linear models, power analysis, survival analysis,

structural modeling equation, etc.

Week 12: High-performance R. Memory management, in-core processing, byte-compiling, C++ interfaces, parallel techniques.

Examples and assignments, presented and discussed within the course, will cover study cases based on clinical trials, drugs tests, medical cases and hospital treatments, differential genes expression, bioninformatics and *omics techniques, etc.