Biostatistics 615 – Statistical Computing – Fall 2015

Instructor: Jian Kang, PhD. Assistant Professor of Biostatistics
Email: jiankang@umich.edu
Office: 3651 SPH I
Office hour: Thursday 10:15AM – Noon, or By Appointment
Class Time: Tuesday & Thursday 8:30AM - 10:00AM
Class Room: 1230 Undergraduate Science Building (USB)

Graduate Student Instructor: Pin Li, B.S. and Master Student of Biostatistics
Email: pinli@umich.edu
Office hour: TBD
Room: TBD

Objective

This course aims at providing students with a practical understanding of computational issues related to the implementation of statistical methods, from basics of programming language to inner workings of sophisticated statistical methods. C++ and R languages will be used.

Prerequisites

Linear algebra (matrix theory), multivariate calculus, basic statistics including probability distribution, linear model and hypothesis testing. Biostatistics 601 or equivalent is required prior to or in parallel to taking Biostatistics 615. Previous experience in programming is not required, but those who do not have previous programming experience should expect to spend additional time studying and learning to be familiar with C++ and R programming languages during the coursework.

Reference Books (Optional)

- Press, Teukolsky, Vetterling, and Flannery, *Numerical Recipes, 3rd Edition*, Cambridge University Press, 2007
- Randall and Kupresanin, *Statistical Computing in C++ and R.* CRC Press, 2011
- Hedley Wickham, *Advanced R*, CRC, 2014
- Lange, Kenneth, *Numerical analysis for statisticians*. Springer Science & Business Media, 2010

Grading

- Homework: 50%
  - Homework assignments will be given out at every two weeks
  - Encourage students to discuss homework problems with others
  - Must implement and write up the assignment individually
  - Plagiarism will NOT be tolerated and late homework will NOT be graded

- Quiz: 10%
  - Small quizzes (multiple choice questions) will be given during classes

- Course project: 40%
  - Two or three students form a team
  - Develop a C++ command-line tool or an R package for cutting-edge statistical methods with a detailed help document for users
  - Give an oral presentation to demonstrate the developed software package
  - Write a scientific report for introducing the developed software package
  - Excellent software and reports are encouraged to be submitted to journals for publication (E.g., *Journal of Statistical Software*, *Bioinformatics Application Notes, The R Journal*)

- Letter grades: A+ = [95,100], A = [85, 94], A- = [75, 84], B+ = [60, 74], B = [50, 59], B- = [40, 49]

Important Dates:

- September 15th: Homework 1 distributed
- September 28th: Drop / Add deadline for full term classes
- September 29th: Homework 1 due & Homework 2 distributed
- October 13th: Homework 2 due & Homework 3 distributed
- October 20th: No class for fall break
- October 27th: Homework 3 due & Homework 4 distributed
- November 3rd: Team members and topics for course project due
- November 10th: Homework 4 due & Homework 5 distributed
- November 24th: Homework 5 due
- November 26th: No class for Thanksgiving
- December 1st – December 10th: Course project oral presentations
- December 18th: Course project due

Topics

- Part I: Basics of C++, R, Data Structure and Algorithms
  - Introduction to C++ and R
  - Key Data Structure
  - Computational time complexity
  - Basic algorithms (Sorting, Divide and Conquer, Dynamic Programming)
  - Interfacing C++ and R languages
  - Introduction to parallel computing
- Part II: Numerical and Statistical Methods
  - Numerical methods (Matrix computing, Optimization, Root finding, Numerical integration)
  - Monte Carlo methods (Random number generation, Importance sampling, Bootstrap)
  - Special topics:
    - Expectation-Maximization (EM) algorithm
    - Markov chain Monte Carlo (MCMC) methods
    - Variable selection algorithms

Core Competencies

After taking this class, students are expected to be able to

- Understand core numerical and statistical algorithms for data analysis
- Analyze the computational time complexity of statistical algorithms
- Efficiently implement sophisticated statistical methods using C++ and R
- Develop C++ command-line tools and R packages
- Perform simulation studies to compare different methods
- Write scientific reports to introduce developed software

Standards of Academic Act

The following is an extract from the School of Public Health's Student Code of Conduct.

Student academic misconduct includes behavior involving plagiarism, cheating, fabrication, falsification of records or official documents, intentional misuse of equipment or materials, and aiding and abetting the perpetration of such acts. The preparation of reports, papers, and examinations, assigned on an individual basis, must represent each student's own effort. Reference sources should be indicated clearly. The use of assistance from other students or aids of any kind during a written examination, except when the use of books or notes has been approved by an instructor, is a violation of the standard of academic conduct.

In the context of this course, any work the student hand-in should be his/her own and any material that is a transcript (or interpreted transcript) of work by others must be clearly labeled as such.