STAT 598G: Introduction to Computational Statistics

Jian Zhang

WEBSITE:  http://www.stat.purdue.edu/~jianzhan/STAT598G/

Description:  Successful statistical data analysis relies increasingly on using computers. As datasets and dimensionality increase in size the computational element of data analysis takes on a critical role. In these cases, statistical inference requires carefully crafted solutions that are computationally efficient and numerically accurate. This course starts with an introduction to programming in R, a specialized statistical language for data analysis, and C. This introduction will progress at a fast pace and will assume that the students have had some previous exposure to programming. Following that, we will cover several fundamental data structures and algorithms that are directly related to statistical analysis. In the later part of the course we will study some numerical methods and their applications in statistics such as penalized Maximum Likelihood Estimation, Expectation Maximization, Generalized Linear Models, Monte Carlo and bootstrap. Students will develop their computing skills via programming exercises in R and C, and a medium-scale programming project.

Required Textbook: None. Lecture notes will be provided online.

Topics:

• Programming for statistics
  – Basics of the R and C languages
  – Interfacing with the operating system; calling C from within R programs
  – Data visualization using R

• Data structures and algorithms for statistics
  – Introduction to computation and time complexity
  – Elementary data structures for statistics: linked lists, trees, graphs
  – Elementary algorithms for accessing and manipulating statistical data: divide-and-conquer, searching, sorting, dynamic programming, graph algorithms
• Numerical Methods for Statistics
  – Optimization techniques: gradient descent, Newton’s method, conjugate gradient, quasi-Newton
  – Monte Carlo methods, Metropolis-Hasting and Gibbs sampling

• Statistical Applications
  – Linear and logistic regression
  – Nonparametric density estimation
  – Mixture model and the EM algorithm
  – Hidden Markov Model
  – Bootstrap methods
  – Bayesian analysis using sampling

Prerequisites:

• STAT 516, STAT 517, and some familiarity with computing. In particular, the students should have some programming experience using a language such as C, C++, Pascal, FORTRAN, Java (the students should be able to write, debug and compile a simple program in one of the above languages).

Policies:

• Grading: assignments: 40%, programming project 20%, final 40%.

• Homework: Assignments will be posted on the course web. You should start the homework early. Late homework will not be accepted unless it is pre-approved by the instructor or an emergent situation. Please talk to the instructor beforehand if you want to get an extension due to special reasons.

• Participation: Please attend the lectures, as exams might contain material covered in class but not in the textbook/reference. Class notes will be posted on the web regularly. You are strongly recommended to read the notes and suggested readings quickly before lecture. A second read after class is necessary to fully understand the material and prepare you for the homework.

• Collaboration: Group discussions are encouraged to further understand difficult topics. You may consult with other students about homework problems, provided that you must indicate such information (whom you consulted with, which problem, to which extent) on your solution sheet. However, you must refrain from getting direct answers from others.