Introduction to Computational Statistics
STAT 598G, Fall 2010

Instructor: S.V. N. Vishwanathan (email: vishy)
TA: Hyokun Yun (email: yun3)

https://learning.stat.purdue.edu/wiki/courses/598g/start

August 23, 2010
Class Details

- Classes: Tue/Thurs 10:30 am - 11:45 am MECH 117
- Recitation: Mon 4:30 pm - 5:00 pm BRNG B282
This is an introductory course in statistical computing.
You will learn how to code a number of statistics algorithms.
You will also learn a number of computer science concepts:
  - Data structures
  - Algorithms
Emphasis throughout the course will be on using C.
We will also review and use R for some tasks.
Ideal Audience

- Well versed with fundamental statistical concepts such as
  - Probability
  - Random Variables
  - Mean and Variance
  - etc.

- Comfortable with statistical algorithms such as
  - Linear and Logistic regression
  - k-means clustering
  - etc.

- Good familiarity with $R$ (or Matlab)

- Some familiarity with $C$ or any other such high level programming language (SAS, SPSS etc do not count!)

- Interested in learning how to *efficiently* code algorithms for *large scale* data analysis
Prerequisites

Required
- STAT 516: Basic Probability and Applications
- STAT 517: Statistical Inference
- MA 511: Linear Algebra
- Or equivalent . . .

Desirable
- STAT 598Z: Concepts in Computing with Data
- CS 190C: Introduction to Computational Thinking
- Or equivalent . . .
8 Assignments: 5 points each
Course project: 30 points
Midterm Quiz: 10 points
Finals: 20 points

Other policies on the course home page. Please review carefully.
Recitation and Office Hours

- Office Hours: 1:00 - 2:00pm Tue or by appointment at HAAS 232
- TA Office Hours: 11:00 am - 12:00 pm Fri or by appointment at HAAS 175
- Recitation: 4:30 - 5:30pm Mon at BRNG B282
Software Requirements

UNIX environment!

C programming
- GCC suite with gdb for debugging
- Any editor (preferably emacs)

R Programming
- R software package
- Visit [http://www.r-project.org](http://www.r-project.org)
- ESS for emacs

\LaTeX (optional)
- TexLive and related packages
- AucTeX mode for emacs
Frequently Asked Questions I

**Q:** Will I need to do lots of programming?

**Ans:** Yes. The homework problems will increasingly become tougher and involve more and more programming. Besides, you are expected to do a non-trivial project.

**Q:** Will I become a guru C or R programmer?

**Ans:** You may very well become one, but that is not the goal of this course. For most part we will stick to basic language constructs and simple syntax.

**Q:** Will you teach us how to use standard libraries e.g. for matrix manipulation or sorting?

**Ans:** No. We will learn to code everything we need from scratch. The aim is to learn the underlying data structures and algorithms. Besides, there are no really “standard” libraries for matrix manipulation.
Q: Will I need lots of maths to understand your lectures?
Ans: I expect familiarity with
- Linear Algebra
- Multivariate Calculus
- Probability Theory

as pre-requisites. There will be emphasis on rigor even when learning about algorithms and data structures. After all, Computer Science is all about discrete maths!

Q: Can I meet you anytime I want?
Ans: I will definitely be around during office hours. You are welcome to walk in any other time I am in my office, but do remember that I generally have busy days. To avoid disappointment it is best to book a slot via email.
Q: Do you reply to emails?
Ans: I try to reply to emails as promptly as possible. If you do not hear back from me within 3 - 4 days then please ping me during the class. Your email may have ended up in my junk mail folder!

Q: Can I solve the HW problems collaboratively?
Ans: The course policy clearly says:
*Group discussions are encouraged to further understand difficult topics. You may consult with other students about homework problems, provided that you 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.* Any violation will result in zero credit for the assignment.
Q: How do I submit my HW?

Ans: For problems which do not involve coding, neatly type or write the solution and submit in class. I strongly encourage the use of LaTeX and discourage the use of MS Word. For solutions which involve coding, submit a print out in class and send your code via email to the TA before the class.
Q: Is there a textbook for this course?

Ans: No. I draw upon a number of resources and books to teach this course. Some of them can be found below. However, I will regularly post handouts with material covered in the class.

Q: Will you post notes for all topics?

Ans: Yes for almost all topics except standard ones for which I will refer you to chapters in a text book or to other standard resources.

Q: Will you use slides (e.g. powerpoint) for your lectures?

Ans: No. I prefer to lecture on the blackboard. Class notes will be available for download from the course home page shortly after the class.
Topics (Tentative) I

- Review
  - Programming in R
  - Introduction to Computational Complexity (Oh notation)
  - Introduction to UNIX

- Programming in C
  - Compiled vs interpreted languages
  - Basic syntax
  - Functions
  - Loops and control flow
  - Arrays
  - Strings
  - Pointers and memory management
  - Input/Output
  - Interfacing C with R
Topics (Tentative) II

- Data structures and algorithms
  - arrays and linked lists
  - sorting, recursion, divide-and-conquer, and dynamic programming

- Statistical Applications
  - Basic linear algebra in C (dense and sparse vectors and matrices)
  - k-nearest neighbors
  - k-means clustering
  - Fuzzy k-means (Gaussian Mixture Models)
  - Expectation Maximization (EM)
  - Kernel density estimation
  - Linear regression
  - Optimization techniques: gradient descent and Newton’s method
  - Logistic regression
  - Sampling and MCMC methods
  - Hidden Markov Models (if time permits)
Please answer as truthfully as possible
Can help me tailor the lectures
Talk to me if you have any concerns or comments
Questions?