STA 250 :: Syllabus

Advanced Statistical Computation (Baines)

UCD, Fall Quarter, 2013

(Syllabus last updated: 09/14/13) The course is organized around the following key topics:

1. "Complex" modeling: Bayesian inference, computational methods, applications
2. "Big" Data: understanding, approaches, tools, applications
3. "Fast" computation: methodology, technology, tools, applications

To cover these topics, the course will be broken into four *modules*: (i) Bayesian Inference and Computation, (ii) Statistics with "Big Data", (iii) Optimization and the EM Algorithm, and, (iv) Efficient Computing: Parallelization and GPUs.

The course is designed to equip students with the basic skills required to tackle challenging problems at the forefront of modern statistical applications. For statistics PhD students, there are many rich research topics in these areas. For masters students, and PhD students from other fields, the course is intended to cultivate practical skills that are required of the modern statistician/data scientst, and can be used in your own field of research or future career.

Before we get into the fun stuff, the first few classes will serve as a "boot camp" to make sure everyone has the mathematical and programming background to tackle the challenges later in the course. We will also use the first few weeks to become familiar with some of the key datasets that we will use throughout the course.

Logistics

- Instructor: Prof. Paul Baines (http://www.stat.ucdavis.edu/~pdbaines/)
- Lectures: Mon, Wed: 4:10pm-6:00pm (Bainer 1130)
- Course CRN: 53615
- Units: 4.0
- Office Hours:

  - Mon: 3:10-4:00pm (MSB 4105)
  - Fri: 12:00-1:00pm (MSB 4105)

- Pre-requisites (official): Course 131A; Course 232A is recommended but not required.
- Pre-requisites (unofficial): Comfort with the foundations of statistical inference. The ability to program in a statistical software environment such as R, Python, Matlab (or similar).
- Course Website: https://piazza.com/ucdavis/fall2013/sta250
- SmartSite URL: https://smartsite.ucdavis.edu/xsl-portal/site/pdbsta250
- Course GitHub: https://github.com/STA250
- Example GitHub repo: https://github.com/STA250/Stuff

Evaluation

Grading for the course will be broken down with the following weighting:

- Homeworks: 60% (4 x 15% each)
- Final project: 40%

There is no final exam for the course.

Course Topics

- Using R and Python for Statistics
- Introduction to Bayesian Inference
- Markov Chain Monte Carlo: Theory and Practice
- Batch Computing using Gauss: Simulation Studies, Model Validations
- Introduction to "Big" Data: Types, Philosophy, Computational Models
- Methodology for "Big" Data: Bag of Little Bootstraps & Others

- Databases for Statistics
- Working with "Big" data: Hadoop, MapReduce, Hive
- Cloud Computing: Amazon EC2, S3, EMR
- Optimization and the EM Algorithm
- Introduction to GPUs: Structure, CUDA, OpenCL
- High-Level Programming for GPUs: RCUDA, PyCUDA, Applications

Tentative Course Schedule

Lecture	Date	Topic	Notes
01	Mon 30th Sep:	Course Overview, Demos	
02	Wed 2nd Oct:	Boot Camp – Basics, R, Python	
03	Mon 7th Oct:	Boot Camp – Gauss, Linux, Stats	
04	Wed 9th Oct:	Bayes I – Introduction to Bayes	Homework 0 Due
05	Mon 14th Oct:	Bayes II – MCMC/Bayesian Computing	
06	Wed 16th Oct:	Bayes III – Inference/Model Checking	
07	Mon 21st Oct:	Bayes IV – Applications/Extras	
08	Wed 23rd Oct:	Big Data I – Types of "Big" Data	
09	Mon 28th Oct:	Big Data II – "Big" data strategies	Homework 1 Due
10	Wed 30th Oct:	Big Data III – "Big" data computation	
11	Mon 4th Nov:	Big Data IV – Applications/Extras	
12	Wed 6th Nov:	EM I – Introduction to EM	
–	Mon 11th Nov:	**NO CLASS – VETERANS DAY**	
13	Wed 13th Nov:	EM II – Variations on EM	Homework 2 Due
14	Mon 18th Nov:	EM III – Parametrization, Convergence	
15	Wed 20th Nov:	EM IV – Efficient algorithms	
16	Mon 25th Nov:	GPUs I – Overview of GPUs	Homework 3 Due
17	Wed 27th Nov:	GPUs II – Programming GPUs	
18	Mon 2nd Dec:	GPUs III – High-level GPU interfaces	
19	Wed 4th Dec:	GPUs IV – Applications/extras	
–	Fri 6th Dec:		Homework 4 Due
–	Mon 9th Dec:		Final Project Due

Assignments: The basic outline for homeworks is below. All due dates are subject to change.

- Homework 00: *Boot Camp Computing Basics* (Not for credit: due Mon Oct 14th)

- Homework 01: *Bayesian Inference*: Theory, MCMC, Validation using Gauss (due Mon Oct 28th)

- Homework 02: *Big Data*: Bag of Little Bootstraps, Databases, Hive, EMR (due Wed Nov 13th)

- Homework 03: *The EM Algorithm*: Theory, Applications (due Mon Nov 25th)

- Homework 04: *Programming with GPUs*: Applications using RCUDA/PyCUDA (due Fri Dec 6th)

and, last, but by no means least:

- Final Project: There will be a choice of four final projects, corresponding to the four modules of the course. Students will be allowed to select the project of their choice. Due during final exam period (Monday December 9th). Final projects will be posted early in the course to allow students to work on them throughout the quarter.

Other Course Duties:

- **Note-taking:** Each lecture will have two assigned note-takers. Each note-taker will be required to produce an electronic copy of the lecture material presented. Preferred formats are `LaTeX` and `Markdown`. Other formats that can be converted to `.pdf` such as `.docx`, `.txt.` etc., are fine as well. Scanned versions of handwritten notes are discouraged as these cannot be easily edited by other students. With the anticipated course enrollment it is expected that each student will only need to take notes once during the quarter. Lecture notes will be available to all students.

Projects: Below are very basic project descriptions to provide you a flavor of what to expect. Full final project descriptions and datasets will be provided later in the course. All final projects require a written report, and possibly an oral presentation.

1. **Bayesian Project** This project will allow you the opportunity to apply your newly acquired knowledge of Bayesian statistics and computational strategies to a complex model for real-world data (provided by the instructor). You will be required to demonstrate that you are able to effectively solve the problem using simulation results, and also to draw conclusions based on real data.

2. **Big Data Project** This project will extend some of the skills developed in the "Big" data module. It will involve a computationally challenging analysis of a "big" dataset: including model development, refinement, verification and application.

3. **EM Project** Throughout the course you will be introduced to the Expectation-Maximization (EM) algorithm, and many more sophisticated extensions of it. This final project will require you to derive several of these algorithms for a specific statistical model. Once derived, your job will be to implement the algorithms and run simulations to compare competing performance. The final report will detail your algorithms, explain any implementation decisions made, and summarize your findings.

4. **GPU Project** In the fourth and final module of the course you will be introduced to Graphics Processing Units (GPUs) and how they can be used for statistical computation. Using the tools you have learned in class, this final project will require you to implement a statistical analysis that makes use of the power of the GPU. You will be required to implement, debug, test, optimize and evaluate your code.

(: Happy Coding! :)