STA141C: Big Data & High Performance Statistical Computing

Lecture 0: Course information

Cho-Jui Hsieh
UC Davis

April 4, 2017
Course Information

- Website: http://www.stat.ucdavis.edu/~chohsieh/teaching/STA141C_Spring2017/main.html
- My office: Mathematical Sciences Building (MSB) 4232
- Office hours: 4pm-5pm Wednesday
- My email: chohsieh@ucdavis.edu
- TA:
  - Huan Zhang (ecezhang@ucdavis.edu)
  - Clark Fitzgerald (clarkfitzg@gmail.com)
The goal of this:
- How to write a good program for data analytics
- Learn to implement statistical models for big data
- Learn use some open source tools
- How to parallelize your code

We’ll use **python** for this course

Homework will be solving real world data mining problems:
- Data from Kaggle or KDDCup.
Course Structure

- **Statistical Programming (in python)**
  - Basic python programming (including numpy, scipy, etc)
  - Analyze the time and memory usage of your program
  - Basic algorithms and data structure, and how to use them in python

- **Advanced statistical computing**
  - Linear algebra and applications (clustering, regression, dimensional reduction)
  - Optimization and applications (classification, regression)
  - Call some existing algorithms from scikit learn

- **Parallel computing**
  - Multicore programming
  - Distributed computing
Prerequisites

- Basic python programming skill
- Basic math and statistics
Grading Policy

- Homework (60%)
- Final project (30%)
- Class participation (10%)

Homeworks:
- We will have about 6-7 homeworks, each one has some programming problems.
- You’ll need to write a report for each homework.
- Use python to write the programming part.

Final project:
- Form a group of 2 people
- Work on a real data mining problem or a data mining contest.
- Project proposal due at the 5-th week (TBD)
- Final project report due at the end of this quarter (TBD)
Piazza and Smartsite

- We will use Piazza for online discussions and Q/A:
  piazza.com/uc_davis/spring2017/sta141c
  Access code: sq17sta141c
- Homework turn-in policy:
  - Homework will due Thursday in class.
  - Please turn in a hard copy (including code and report) in class, and also submit the code and report on smartsite.
No discussion sessions for the first week
Next week we’ll have discussion sessions on “introduction to python”
Later on TAs will be giving some tutorial or reviewing homework solutions in discussion sessions.
First homework will be related to the following Kaggle competition: Quora question pairs challenge (Kaggle):
https://www.kaggle.com/c/quora-question-pairs
We will implement a simple algorithm for this problem as homework 1 (will officially announce later).
Coming up

- Basic python programming

Questions?