Course Information

- Website: http://www.stat.ucdavis.edu/~chohsieh/teaching/STA141C_Spring2018/main.html
- My office: Mathematical Sciences Building (MSB) 4232
- Office hours: 1pm–2pm Tuesday
- My email: chohsieh@ucdavis.edu
- TA:
  - Chun-Jui (Gary) Chen (cjuchen@ucdavis.edu)
  - Justin Wang (jstwang@ucdavis.edu)
Course Information

- The goal of this:
  - How to write a good program for data analytics
  - Learn to implement statistical models for big data
  - Learn to use some open source (python) tools
  - How to parallelize your code
- We’ll use **python** for this course
Course Structure

- Statistical Programming (in python)
  - We assume you already know how to use python
  - Basic algorithms and data structure, and how to use them in python

- Numerical Optimization for Statistical Problems
  - Briefly review basic optimization algorithms
  - How to use them to solve real world problems
    linear regression, classification, neural network

- Parallel Computing in Python
  - Multicore computing
  - Distributed computing

- Numerical Linear Algebra for Statistics
  - Matrix decomposition for huge matrices
    PageRank, Clustering, Word2vec

- Feature generation
  - Text, images
Prerequisites

- Basic python programming skill
  (STA 141B or ECS python course)
- Basic math and statistics
  (linear algebra, matrix multiplication, eigen-decomposition)
Grading Policy

Homework (65%)

Final project (35%)

Homeworks:

- Homework will be 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 \( \leq 4 \) 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)
Discussion sessions

- No discussion sessions for the first week
- Later on TAs will be giving some tutorial or reviewing homework solutions in discussion sessions.
- Statistical Programming (in python)
- Numerical Optimization for Statistical Problems
- Parallel Computing in Python
- Numerical Linear Algebra for Statistics
- Feature generation
Optimization:

\[
\min_w f(w)
\]

\(f\): objective function to be optimized
Optimization:

\[ \min_{\mathbf{w}} f(\mathbf{w}) \]

\( f \): objective function to be optimized

Example: linear regression

\[ \min_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i - y_i)^2 \]
Optimization:

\[
\min_w f(w)
\]

\(f\): objective function to be optimized

Example: linear regression

\[
\min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2
\]

Generalize to arbitrary loss:

\[
\min_w \sum_{i=1}^{n} \text{loss}(w^T x_i - y_i)^2
\]
Optimization:

\[
\min_w f(w)
\]

\(f\): objective function to be optimized

Example: linear regression

\[
\min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2
\]

Generalize to arbitrary loss:

\[
\min_w \sum_{i=1}^{n} \text{loss}(w^T x_i - y_i)^2
\]

Generalize to arbitrary function:

\[
\min_w \sum_{i=1}^{n} \text{loss}(g_w(x_i) - y_i)^2
\]
(Iterative) algorithms:
- Gradient descent
- Stochastic gradient descent
- Block-coordinate descent
- ...
(Iterative) algorithms:

- Gradient descent
- Stochastic gradient descent
- Block-coordinate descent
- ...

Implement to solve large problems
Preview: Numerical Optimization for Statistical Problems

- Statistical Programming (in python)
- Numerical Optimization for Statistical Problems
- Parallel Computing in Python
- Numerical Linear Algebra for Statistics
- Feature generation
Basic concept for multicore/distributed parallelism
How to implement (multi-core) parallelism in python
Preview: Numerical Optimization for Statistical Problems

- Statistical Programming (in python)
- Numerical Optimization for Statistical Problems
- Parallel Computing in Python
- Numerical Linear Algebra for Statistics
- Feature generation
Numerical Linear Algebra for Statistics

- SVD, eigen-decomposition
- How to decompose huge matrices
- Applications
  - PageRank, Clustering, Word2vec
Preview: Numerical Optimization for Statistical Problems

- Statistical Programming (in python)
- Numerical Optimization for Statistical Problems
- Parallel Computing in Python
- Numerical Linear Algebra for Statistics
- Feature generation
Feature generation

- Features for text data
- Features for image data
Coming up

- Numpy programming, basic algorithm/data structure

Questions?