# Machine Learning and Computational Statistics (DS-GA 1003)

David Rosenberg

New York University

January 24, 2017

# Logistics

- Class webpage: `https://davidrosenberg.github.io/ml2017`
    - Syllabus on the website
- Piazza: `https://piazza.com/nyu/spring2017/dsga1003`
    - **All class announcements via Piazza**
    - Ask all questions on Piazza
- Class Times
    - Tuesdays "Lecture": 5:20 - 7pm (GSACL C95)
    - Wednesdays "Lab": 8:35 - 9:25pm (GSACL C95)
    - **(Both are required.)**

## Course Staff

- TAs:
  - Brett Bernstein (3rd year PhD Student in Math at Courant)
  - Vlad Kobzar (grad student in Math and ML at Courant)
- Graders:
  - Ben Jakubowski (Head Grader)
  - Hao Liu
  - Yuhao Zhao
  - Xinyi Gong
  - Lanyu Shang
  - Prithvi Krishna "PK" Gattamaneni
- Project Advisers:
  - Kurt Miller, Brian d'Alessandro, Bonnie Ray, Daniel Chen, and more TBD.

## Evaluation

- About 8 homeworks (35%)
- Two tests (45%)
  - One-Hour Test (15%) in Week 6
  - Two-Hour Test (30%) in Week 12
- Project (20%)
  - Poster session during lecture time in week after classes (Week 15)
- Extra Credit Opportunities
  - Optional homework problems
  - Significant contributions to Piazza and in-class discussions
  - Primarily used to boost a borderline grade
  - **At most, increases final grade by half a letter (e.g. B+ to A-)**

# Lab Sessions

- Some led by TA Brett Bernstein, some by me
- Most will be lecture format
- We'll have one test during lab session (and one during lecture)
- Meetings with project advisors

# Homework (45%)

- First assignment out Wednesday – Due week from Thursday 10pm
- Submit with Gradescope (details on website)
- Homeworks should be **submitted as a PDF document**.
- Late homework: Accepted up to 48 hours late with 20% penalty
- Collaboration is fine, but
  - Write up solutions and code on your own
  - List names of who you talked to about each problem

# Projects (20%)

- Find some new data or new approach to old data
- See notes on website.
- Logistics:
    - 3 students per group
    - First meeting with advisors (March 8)
    - Project proposal due after Spring Break (March 23)

# Prerequisites

- DS-GA 1001: Introduction to Data Science
- DS-GA 1002: Statistical and Mathematical Methods
- Math
  - Multivariate Calculus
  - Linear Algebra
  - Probability Theory
  - Statistics
- Python programming (numpy)

# High Level Goals of the Class

- Learn fundamental building blocks of machine learning

- Goal is to start seeing
  - **fancy new method A "is just" familiar thing B + familiar thing C + tweak D**

  - SVM "**is just**" ERM with hinge loss with $\ell_2$ regularization

  - Pegasos "**is just**" SVM with SGD with a particular step size rule

  - Random forest "**is just**" bagging with trees, with an interesting tweak on choosing splitting variables

# General Philosophy

- Mastery vs Performance
  - understanding vs just getting it done
- Don't confuse "kind of understanding" with "actual understanding"
  - See `https://github.com/davidrosenberg/mlcourse/blob/gh-pages/course-faq.md` about partial credit.

# Questions?

- What are you looking to get out of the course?
- Questions for me?