Machine Learning and Computational Statistics (DS-GA 1003)

David Rosenberg

New York University

January 27, 2016

Logistics

- Class webpage: `https://davidrosenberg.github.io/ml2016`
  - Syllabus on the website
- Piazza: `https://piazza.com/nyu/spring2016/dsga1003`
  - Ask questions here
- Class Times
  - Wednesdays "Lecture": 7:10pm - 9pm (WWH 109)
  - Thursdays "Lab": 7:10pm - 8pm (WWH 109)
  - **(Both are required.)**

Lab Sessions

- Some led by TA, some by me
- Most will be lecture format
- Sometimes we'll review homeworks or have test prep
- One-hour test during lab session
- Meetings with project advisors

Course Staff

- TA: Levent Sagun
- Graders:
  - Peter Li (Head Grader)
  - Lucy Wang
  - Jacqueline Gutman
  - Tian Wang
- Project Advisers:
  - Kurt Miller, Kush Varshney, Brian d'Alessandro, and more TBD.

Evaluation

- About 8 homeworks (40%)
- Two tests (40%)
  - One-Hour Test (15%) in Week 6
  - Two-Hour Test (25%) in Week 11
- Project (20%)
  - Poster session during exam week (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-)

Homework (40%)

- First assignment out Thursday – Due in one week.
- Submit with NYU Classes: `https://newclasses.nyu.edu`
- 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
- Project philosophy the same as in these courses:
  - `http://cs.nyu.edu/~dsontag/courses/ml14/assignments/projects.html`
  - `http://web.stanford.edu/class/cs221/project.html`
- Logistics:
  - 3 students per group
  - First meeting with advisors in March (date TBD)
  - Project proposal due around Spring Break (date TBD)

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)

General Philosophy

- Mastery vs Performance
  - (understanding vs "getting the grade")
- Don't confuse "kind of understanding" with "actual understanding"
- From Quora: "Why is L1 regularization supposed to lead to sparsity than L2?"

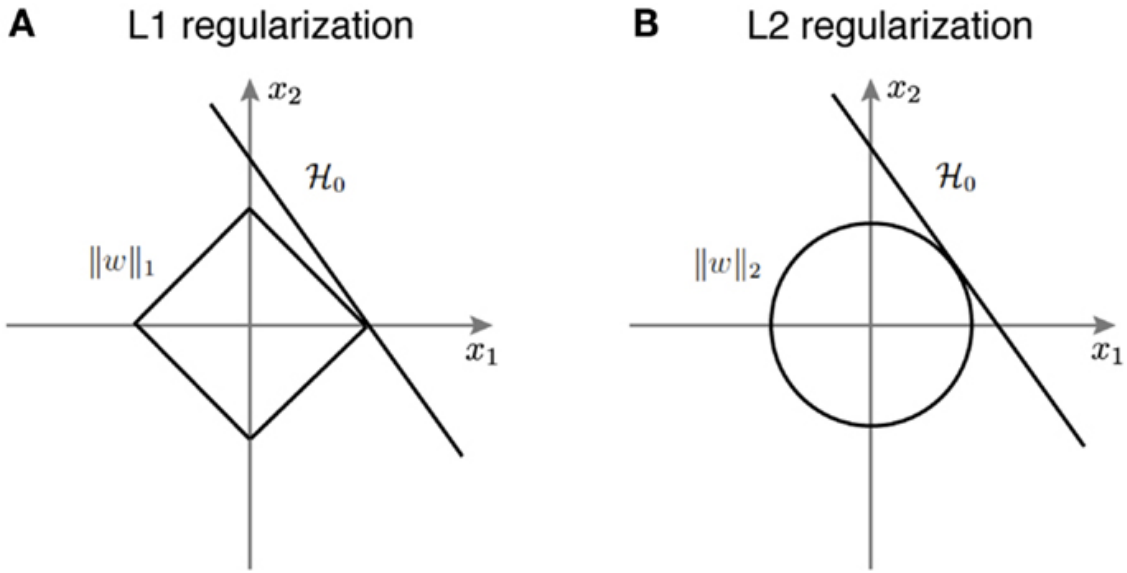

Figure from `https://www.quora.com/Why-is-L1-regularization-supposed-to-lead-to-sparsity-than-L2`.

Course Topics

- Frequentist Approaches
  - ERM, regularization, SVM, kernels, ensemble methods
- Probabilistic Models
  - GLM, Bayesian networks, Gaussian mixture models, EM algorithm
- Bayesian Approaches
  - priors/posteriors, hierarchical models, Bayesian regression
- Misc. and Advanced Topics
  - dimensionality reduction, structured prediction

Questions?

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