August 31, 2020

AMSC 808N / CMSC 828V

Numerical Methods

for

Data Science

and

Machine Learning

Maria Cameron

Department of Mathematics

Department of Computer Science (affiliate)

What is Data Science?

- **Wiki:** is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.
- **Briefly:** ***Data Science is a science about learning from data.***
- **Birth:** J. Tukey's paper "The Future of Data Analysis" (1962)

Father-founder:
John Tukey (1915—2000)
(best known for the FFT algorithm and box-plot)

“50 Years of Data Science” (2015)

Activities of data Science

1. Data Exploration and Preparation;
2. Data Representation and Transformation;
3. Computing with Data;
4. Data Modeling;
5. Data Visualization and Presentation;
6. Science about Data Science.

David Donoho
Anne T. And Robert M. Bass Professor of Humanities and Sciences
Professor of Statistics

Successes of DS&ML that struck me personally

- **ML for solving PDEs in high dimensions arising in the study of rare events in stochastic systems:**

Weinan E and Bing Yu, ***Deep Ritz Method***, https://arxiv.org/pdf/1710.00211.pdf

Yuehaw Khoo, Jianfeng Lu, Lexing Ying, ***Solving for high dimensional committor functions using artificial neural networks***, https://arxiv.org/pdf/1802.10275.pdf

Qianxiao Li, Bo Lin, Weiqing Ren, ***Computing committor functions for the study of rare events using deep learning***, https://arxiv.org/pdf/1906.06285.pdf

- **ML for geophysics:**

Sergey Fomel, ***Automating seismic data analysis and interpretation***, https://seg.org/Education/Lectures/Distinguished-Lectures/2020-DL-Fomel

- **ML for molecular dynamics:**

Yihang Wang, Joao Ribeiro, Pratyush Tiwary, ***Machine learning approaches for analyzing and enhancing molecular dynamics simulations***, https://www.sciencedirect.com/science/article/pii/S0959440X19301551

Inspiration for this course

David Bindel

Associate Professor of Computer Science

Director, Center for Applied Math

CS, applied math, math, civil engineering, and CSE, and data science.
Confusing rabbits since 2003.

425 Gates Hall
Dept of Computer Science
Cornell University
Ithaca, NY 14853-5169

OH: By appointment
Scheduler link
bindel@cs.cornell.edu

Research highlights

Optimizing stellarators
Advancing magnetic confinement fusion through optimization and hidden symmetries.

Home page Announcement

Teaching highlights

Numerical Methods for Data Science
Revisiting numerical methods teaching for modern data science applications.

UChicago (June 2019) SJTU (May 2019) UMD (April 2019)
SJTU (June 2018) Cornell (Spring 2018)

Background

- Linear algebra: vectors, matrices, maps, norms, inner products
- Multivariable Calculus: $f(x+p) = f(x) + \nabla f(x)^T p + (1/2)p^T \nabla^2 f(x) p + \ldots$
- CS: notation $O(n^p)$, operation count, sparse vs dense
- Matlab: basic syntax
- Floating point arithmetic: machine epsilon, roundoff
- Conditioning: ill- vs well-conditioned, condition number

See Bindel's note:
http://www.cs.cornell.edu/~bindel/class/sjtu-summer18/lec/background.pdf

Logistics

- **Lectures:** Tuesday and Thursday 11:00 — 12:15 AM: in person at ITV1100 and via zoom
- **Homework:** latex it, upload your pdf file on ELMS (30% of your grade)
- ***Homework will be due in one week***
- **Homework 1 will be posted on 09/09 and due on 09/16**
- **Programming projects:** (35% of your grade)
  - Working groups (2-3 students)
  - Any suitable language, I recommend **Matlab** or **Python**
  - Each group member should develop his/her code, then verify results and write a common report with other group members; link all codes to it; upload a pdf file with the report on ELMS.
  - ***Projects will be assigned at the end of each chapter and will be due in two weeks***
- **Final exam:** asynchronous, due in one week after posting (35% of your grade)

Asking questions and getting answers

- Piazza for asking all course-related questions. If you haven't received an email from Piazza, let me know. I will enroll you.
- Office hours starting Sept. 7: via zoom
  - Tuesday 2:30—3:30 PM
  - Wednesday 2:30—3:30 PM
- Email — only for special situations: mariakc@umd.edu
- All course information, all lecture notes, all recording links will be posted on ELMS.

Chapters

- Introduction and review of linear algebra
- Optimization for large-scale machine learning
- Matrix data analysis and latent factor models
- Nonlinear dimensionality reduction. Diffusion maps.
- Graph data analysis