Mathematical Algorithms for Artificial Intelligence and Big Data

Thomas Strohmer
Department of Mathematics
University of California, Davis

Spring 2017

Course Objective

Experiments, observations, and numerical simulations in many areas of science nowadays generate massive amounts of data.

This rapid growth heralds an era of "data-centric science," which requires new paradigms addressing how data are acquired, processed, distributed, and analyzed.

This course covers mathematical concepts and algorithms (many of them very recent) that can deal with some of the challenges posed by Artificial Intelligence and Big Data.

Details about this Big Data course

This course is about mathematical methods for Big Data

Prerequisite:
Linear algebra and a basic experience in programming (preferably Matlab) will be required. Solid basis in undergraduate mathematics is recommended.

What this class is not about:

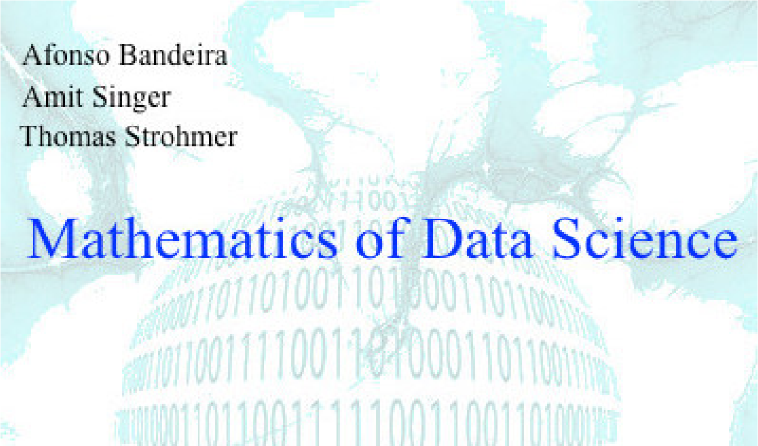
- Formal software development
- Database theory
- Specific applications
- Heuristic methods that lack mathematical foundations (well, except for deep learning ...)

Textbooks

There is no required textbook. The following books contains some material on these topics (but there is no need to buy these books)

- C. Bishop. Pattern Recognition and Machine Learning.
- F. Cucker, D. X. Zho. Learning Theory: an approximation theory viewpoint.
- S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing.
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction.
- Michael W. Mahoney. Randomized Algorithms for Matrices and Data.

Textbook in development



Notes from the book draft will be made available.

Grading Scheme

- 50% Homework: will be assigned about every other week. A subset of these problems will be graded.
- 50% Final Project

Final Project:

Write a 8-page (or so) report on one of the following topics:

- Describe how some of the methods you learned in this course will be used in your research.
- Find a practical application yourself (not copying from papers/books) using the methods you learned in this course; describe how to use them; include numerical demonstrations.
- Find an interesting data set and present a careful numerical comparison of existing algorithms related to one of the topics of this couse.
- If in doubt, please ask me!

Teaching Assistants

Shuyang Ling　　　　　　　　　　　　　　　　　　　　Yang Li

Goal and challenges of Big Data

Goal: The goal is to turn data into information

Challenges: Capture, curation, time-limitations, storage, search, sharing, transfer, analysis, and visualization of the data.

Data can be massive, non-static, multi-modal, incomplete, noisy, non-random, unstructured, dynamic, streaming, ...

“Data is the new (crude) oil for the economy!”

“Data is the new (crude) oil for the economy!”

You are not Google’s customer.

“Data is the new (crude) oil for the economy!”

You are not Google’s customer.

You are Google’s commodity (crude oil)

Big Data Everywhere!

Lots of data is being collected and warehoused

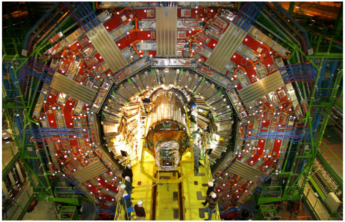
- Web data (often user-provided)
- e-commerce, purchases at stores
- Medical data, health care
- Bank/Credit Card transactions
- Social Network
- Traffic, GPS, ...
- Scientific experiments
- ...

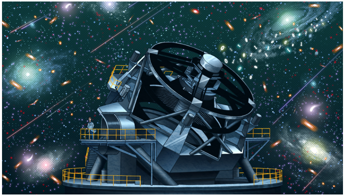
How much data?

- YouTube contains 120 million videos and 72 hours of video uploaded every minute.
- Google processes 3.5 billion requests per day
- There is currently an estimate of 3.8 trillion photographs, 10% of them taken in the last year.
- Facebook has about 140 billion images with about 300 million new images a day.
- 2.5PB are flowing through Walmart's databases
- NYSE collects 1 TB each day.

How much data?

- CERN's Large Hydron Collider generates 15 PB a year
- The BRAIN initiatives produce terabytes of data a day
- The Large Synoptic Survey Telescope in Chile will collect 30TB per night. Headed by Tony Tyson from UC Davis





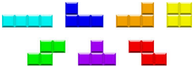
How much data?

Governments (USA, China, Russia, UK, Israel, Germany, ...) collect ??? PB /day

How much data?

Governments (USA, China, Russia, UK, Israel, Germany, ...) collect ??? PB /day

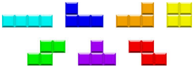
The CIA (via In-Q-Tel) was an early investor in Facebook



How much data?

Governments (USA, China, Russia, UK, Israel, Germany, ...) collect ??? PB /day

The CIA (via In-Q-Tel) was an early investor in Facebook



Somewhere in Nevada is an 8-Football field large storage area that collects all the emails sent in the USA.

More Data ...

Experts now predict that 40 zettabytes of data will be in existence by 2020.

Big Data does not just mean massive amounts of data
Big Data also means complex data

- Heterogeneous data
- Incomplete data
- Unstructured/semi-structured Data
- Graph Data
- Social Network, Semantic Web
- Streaming Data

Big Data is not new

- Seismic data acquisition and processing
- Census
- Wall Street hedge funds (e.g. Renaissance Technologies)
- Governments
- Banks, Insurances
- Scientific Research

Big Data Tasks

- Discovery of useful, possibly unexpected, patterns in data
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Finding outliers (security threat, credit card theft, ...)
- Clustering
- Classification
- Object recognition
- Visualization, dimension reduction
- “Data cleaning”: denoising, smoothing, grouping, ...
- Association Rule Mining (Costumers who buy X often buy Y, Costumer 123 likes product p10)
- Collaborative filtering: users collaborate in filtering information to find information of interest (Amazon, Netflix)

Meta Data Analysis

The idea is 100 years old (see Karl Pearson), but its full potential will be unleashed only now.

Example:
In a recent analysis researchers developed a framework for comparing classiffers common in Machine Learning (Boosted decision trees, Random Forests, SVM, KNN, PAM and DLDA) based on a standard series of datasets.

Result: A simple (but mathematically rigorous) method gave better classification results across the data sets than the “glamorous” methods.

The dawning Age of Big Data will make it not just possible but very common (and perhaps necessary?) to validate methods via such meta data analyses.

Big Data Startups

Crunchbase records more than 2900 Startups and Angellist more than 3500 Startups in "Big Data"

Two random examples (out of 1000+?) of Bay area startups:

- Forensic Logic (Walnut Creek): Crime analysis
- 23andMe (Mountain View): Genomics

Two startups by mathematicians:

- ThetaRay: Cybersecurity (R.R. Coifman, Amir Averbuch)
- Ayasdi: Topological data analysis (Gunnar Carlsson)

Many Data Initiatives Nationwide

Campus-wide initiatives at NYU, Columbia, Michigan, Harvard, MIT, Berkeley, ...

New Master's Degree programs in Data Science, for example at Berkeley, NYU, Stanford, UC Davis, ...

New Alan Turing Institute for Data Sciences in UK

For a long list across the world see
`http://data-science-university-programs.silk.co`

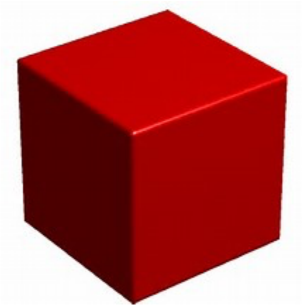
Topic Overview (tentative)

- Basic goals of AI and Machine Learning
- Curses and blessings of dimensionality,
  Surprises in high dimensions
- Singular Value Decomposition,
  Principal Component Analysis
- Data Clustering: k-means, graph Laplacian
- Linear dimension reduction, random projections
- Nonlinear dimension reduction, diffusion maps,
  manifold learning, intrinsic geometry of data,
- Some basics on Deep Learning

Things in high dimension can behave very differently than in low dimension.

Things in high dimension can behave very differently than in low dimension.

A cube in high dimensions does not look like this:

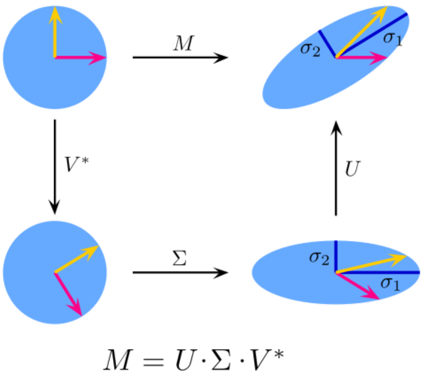


Things in high dimension can behave very differently than in low dimension.

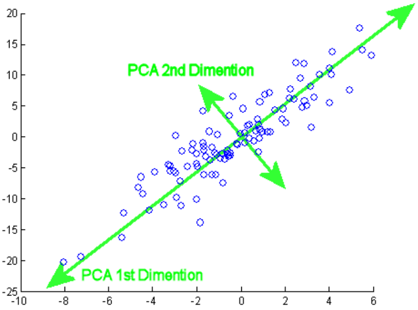
A cube in high dimensions looks like this:

SVD and PCA

Singular Value Decomposition

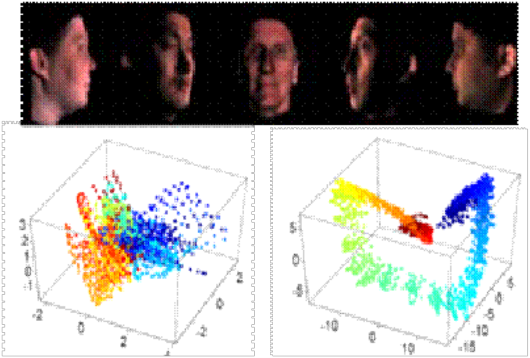


Principal Component Analysis



Dimension reduction

Linear dimension reduction and random projections

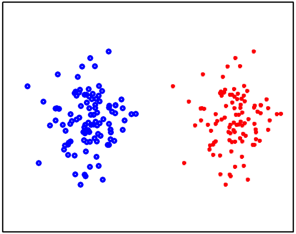


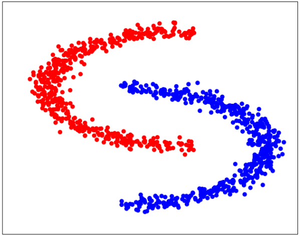
Johnson-Lindenstrauss projections

Clustering

A basic task in data analysis is clustering:

k-means: advantages and limitations





Graph Laplacian, spectral clustering

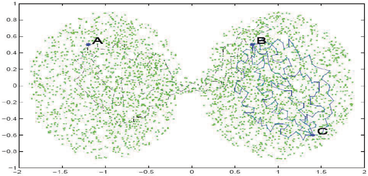
Diffusion maps

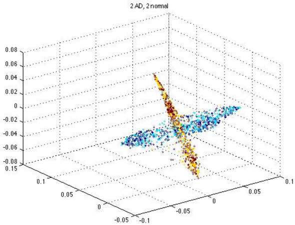
What is a diffusion map?

Manifold learning

Intrinsic geometry of data

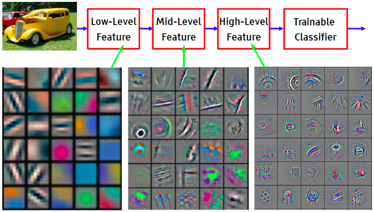
Nonlinear dimension reduction





Deep Learning: neural network with more than one layer

Deep networks achieve state-of-the-art results in several complex object recognition tasks



They learn a huge network of filter banks and non-linearities on large datasets

Heuristic method, a lot of trial-and-error

Almost no mathematical theory (yet)

And last but not least

Algorithms for AI and Big Data are powerful.

Use your power responsibly and carefully.

And last but not least

Algorithms for AI and Big Data are powerful.

Use your power responsibly and carefully.

Einstein: “Not everything that can be counted, counts.
And not everything that counts, can be counted.”