High-dimensional probability and statistics for the data sciences

Larry Goldstein and Mahdi Soltanolkotabi

Motivation
August 21, 2017

Ming Hsieh Department of Electrical Engineering

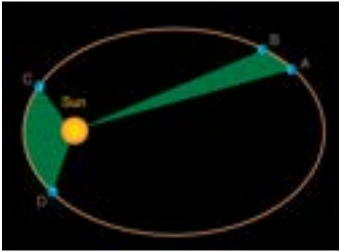
USC University of Southern California

Please ask questions!

Traditionally mathematics driven by physics...

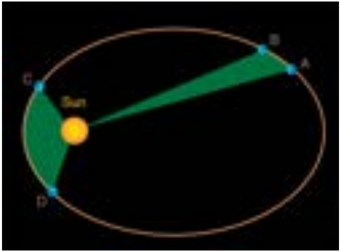
Traditionally mathematics driven by physics...

- Kepler's law of planetary motion

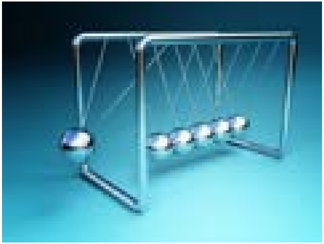


Traditionally mathematics driven by physics...

- Kepler's law of planetary motion

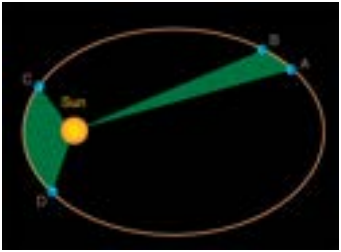


- Newtonian Mechanics

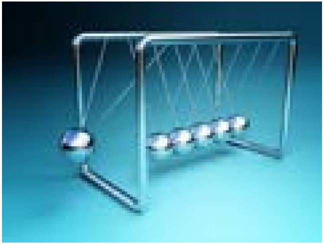


Traditionally mathematics driven by physics...

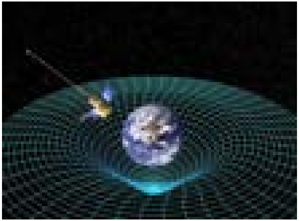
- Kepler's law of planetary motion



- Newtonian Mechanics



General relativity



Todays mathematics is driven by something else...

Todays mathematics is driven by something else...

Can you guess?

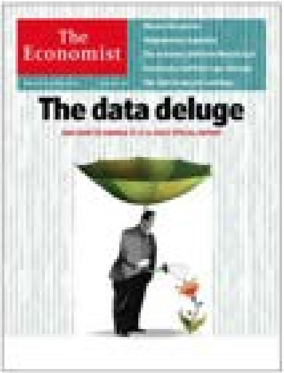
Todays mathematics is driven by something else...

Can you guess?

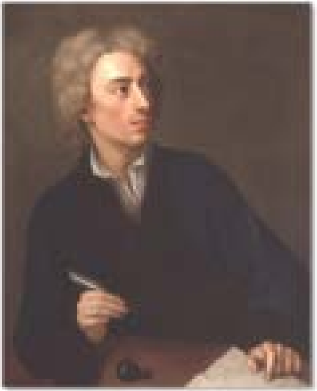
What is different about the data we have today?

What is different about the data we have today?

Size? Data deluge?



Ye Olde Data Deluge



> "Paper became so cheap, and printers so numerous, that a deluge of authors covered the land"
>
> Alexander Pope, 1728

What is different about the data we have today?

variety and complexity

What is different about the data we have today?

variety and complexity

Web Data

What is different about the data we have today?

variety and complexity

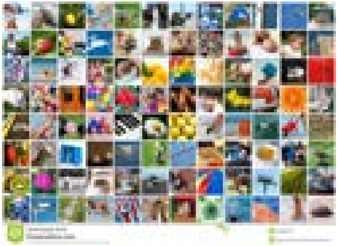
Web Data

- Text Data

- Social network data

- Video Data

- Image data



What is different about the data we have today?

variety and complexity

What is different about the data we have today?

variety and complexity

Scientific Data

What is different about the data we have today?

variety and complexity

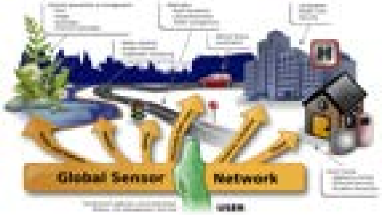
Scientific Data

- Remote Sensing Data

- Genomic Data

- Brain Data

- Sensor Network Data



What is different about the data we have today?

variety and complexity

What is different about the data we have today?

variety and complexity

variety of platforms

Challenge


WHAT WOULD YOU DO WITH ALL THIS DATA?
Pictures
Website
Tweets
Facebook

Challenge

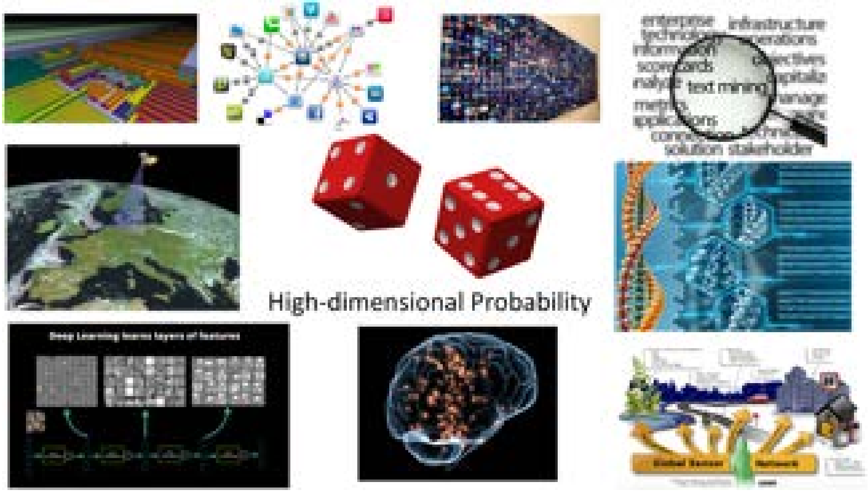
Conclusion

"Scientific Data Has Become So Complex, We Have to Invent New Math to Deal With It"

Need new Math...

Mathematics Driven by Data

Probability at the heart of data science



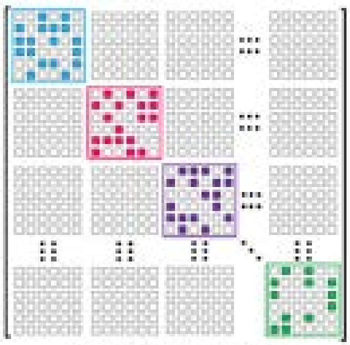
Example I:
Community detection and stochastic block models

Standard Machine

1 Construct affinity matrix $\boldsymbol{W}$ between samples $\rightarrow$ weighted graph

$$\boldsymbol{W}_{i,j} = \exp\left(-\frac{\|\boldsymbol{x_i} - \boldsymbol{x}_j\|_{\ell_2}^2}{2\sigma^2}\right).$$

2 Construct clusters by applying spectral clustering to $\boldsymbol{W}$



Ideal affinity matrix

Spectral clustering

- Affinity matrix $\boldsymbol{W}$ $(N \times N)$
- Degree matrix $\boldsymbol{D}$=diag(d_i)

$$d_i = \sum_j W_{ij}$$

- Normalized graph Laplacian (symmetric form)

$$\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{1/2} \quad (N \times N)$$

Spectral clustering

- Affinity matrix $\boldsymbol{W}$ $(N \times N)$
- Degree matrix $\boldsymbol{D}$=diag(d_i)

$$d_i = \sum_j W_{ij}$$

- Normalized graph Laplacian (symmetric form)

$$\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{1/2} \quad (N \times N)$$

	$\boldsymbol{v}_1$	$\boldsymbol{v}_2$	$\boldsymbol{v}_3$
$\boldsymbol{r}_1$	v_{11}	v_{12}	v_{13}
$\boldsymbol{r}_2$	v_{21}	v_{22}	v_{23}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{r}_N$	v_{N1}	v_{N2}	v_{N3}

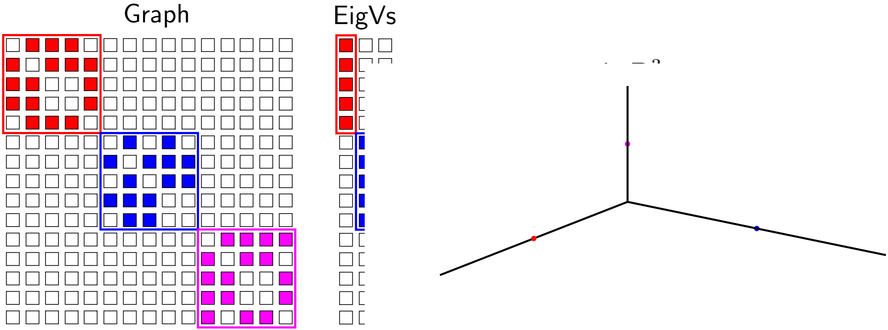
Dim. reduction: $N \longrightarrow k$

Spectral clustering

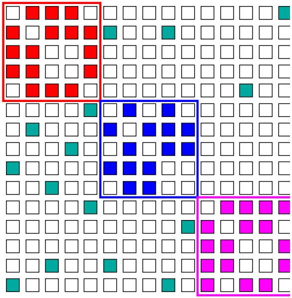
(1) Build $\boldsymbol{V} \in \mathbb{R}^{N \times k}$ with first k lowest eigenvectors of $\boldsymbol{L}$ as columns
(2) Interpret ith row of $\boldsymbol{V}$ as new data point $\boldsymbol{r}_i$ in $\mathbb{R}^k$ representing observation i
(3) Apply k-means clustering to the points $\{\boldsymbol{r}_i\}$

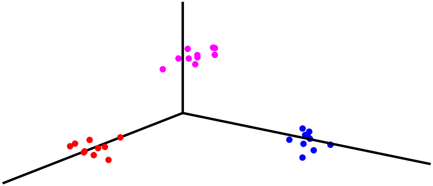
Fantastic tutorial: U. von Luxburg

Spectral Clustering: Ideal case



Spectral Clustering: non-Ideal case





Adjacency matrix of random graphs

- Adjacency matrix of Graph

$$\boldsymbol{A}_{ij} = \begin{cases} 1 & \text{with prob. } P_{ij} \\ 0 & \text{otherwise} \end{cases}$$

Adjacency matrix of random graphs

- Adjacency matrix of Graph

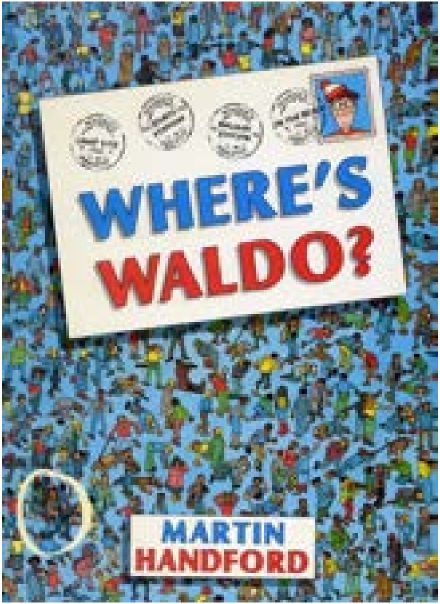
$$\boldsymbol{A}_{ij} = \begin{cases} 1 & \text{with prob. } P_{ij} \\ 0 & \text{otherwise} \end{cases}$$

- Probability matrix for stochastic block model

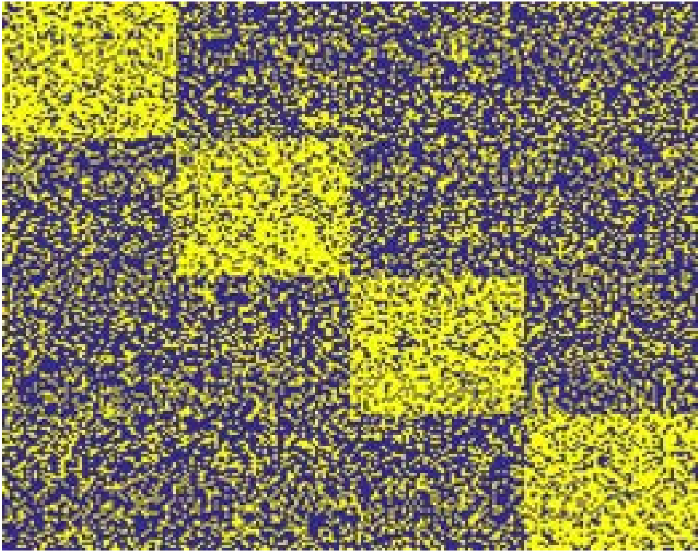
$$\boldsymbol{P} = \begin{bmatrix}
p & p & p & q & q & q & q & q & q & q & q & q \\
p & p & p & q & q & q & q & q & q & q & q & q \\
p & p & p & q & q & q & q & q & q & q & q & q \\
q & q & q & p & p & p & q & q & q & q & q & q \\
q & q & q & p & p & p & q & q & q & q & q & q \\
q & q & q & p & p & p & q & q & q & q & q & q \\
q & q & q & q & q & q & p & p & p & q & q & q \\
q & q & q & q & q & q & p & p & p & q & q & q \\
q & q & q & q & q & q & p & p & p & q & q & q \\
q & q & q & q & q & q & q & q & q & p & p & p \\
q & q & q & q & q & q & q & q & q & p & p & p \\
q & q & q & q & q & q & q & q & q & p & p & p
\end{bmatrix}$$

k clusters of size n/k $\quad 0 \leq q \leq p \leq 1.$

Where's Waldo?

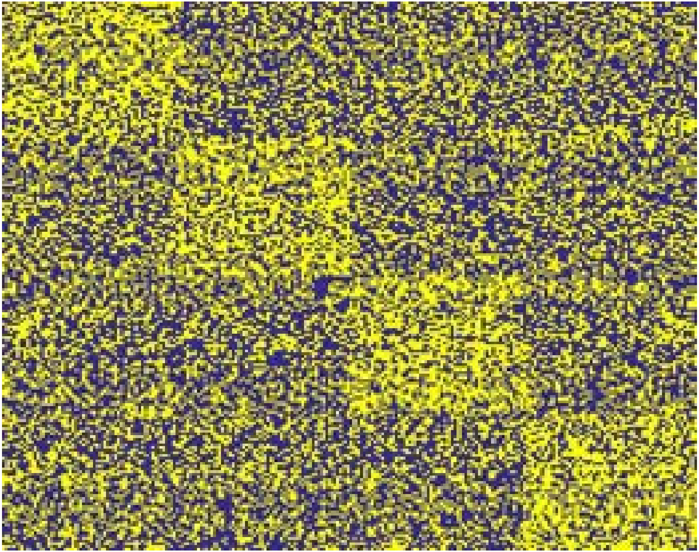


Where are the clusters?



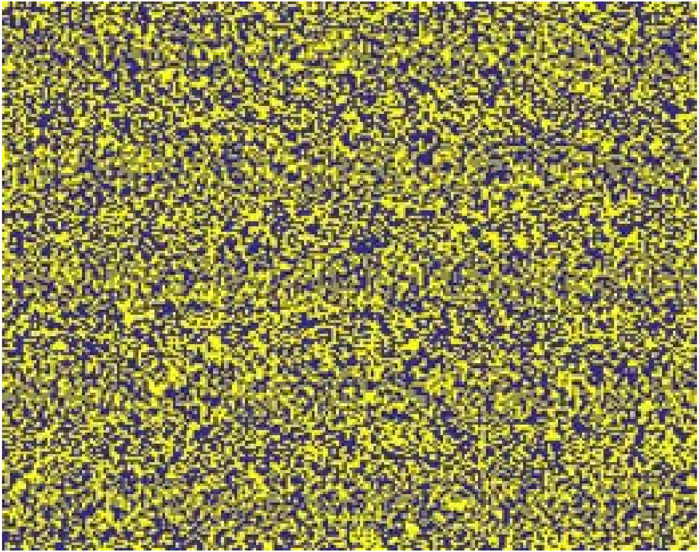
$n = 200,\ k = 4,\ p = 0.7,\ q = 0.3.$

Where are the clusters?



$n = 200$, $k = 4$, $p = 0.6$, $q = 0.4$.

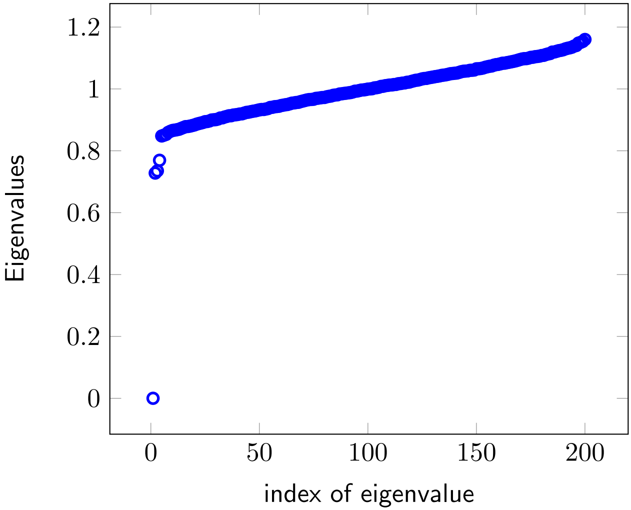
Where are the clusters?



$n = 200$, $k = 4$, $p = 0.6$, $q = 0.4$.

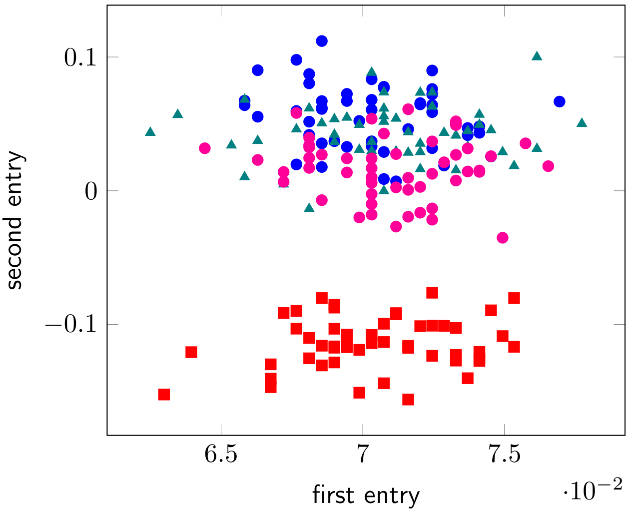
Will it work?

Eigenvalues of the normalized Laplacian with $n = 200$, $k = 4$, $p = 0.6$, $q = 0.4$.



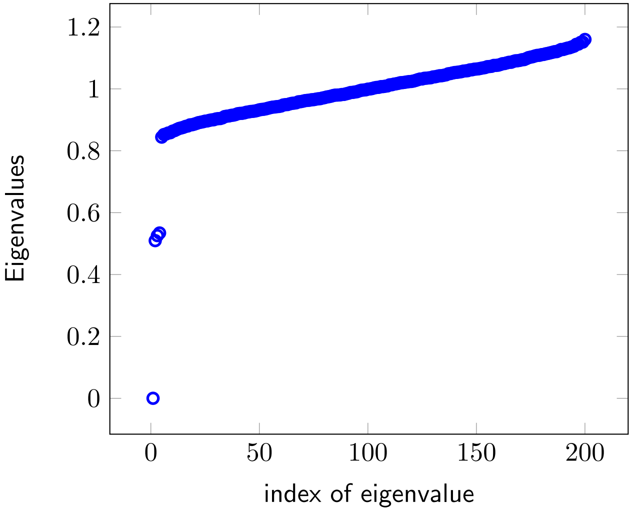
Eigen vectors

Top two eigenvectors of the normalized Laplacian $n = 200$, $k = 4$, $p = 0.6$, $q = 0.4$.



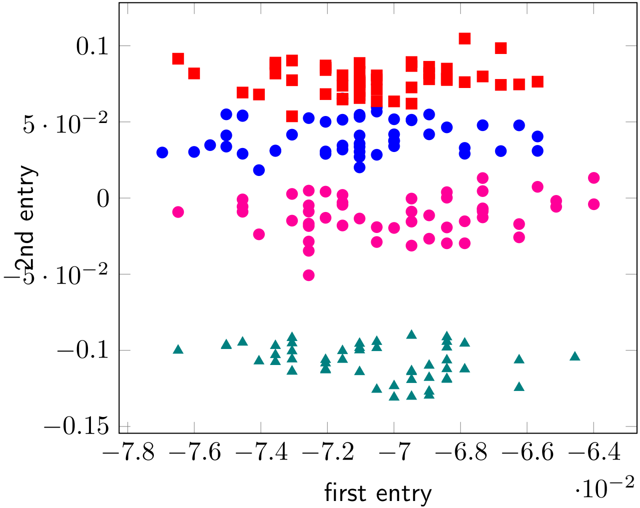
Will it work?

Eigenvalues of the normalized Laplacian with $n = 200$, $k = 4$, $p = 0.7$, $q = 0.3$.



Eigen vectors

Top two eigenvectors of the normalized Laplacian $n = 200$, $k = 4$, $p = 0.7$, $q = 0.3$.

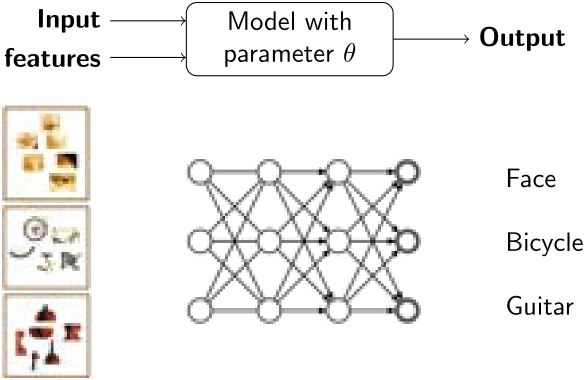


Example II:
Learning models from data and uniform concentration results

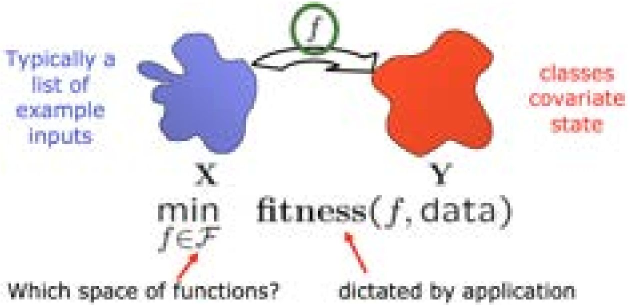
Learn model from training data

Main abstraction of machine learning: parameter estimation

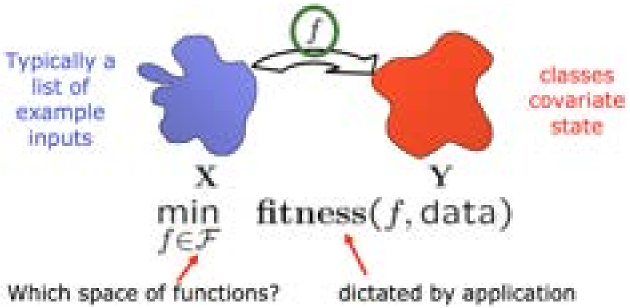
- Data (n samples)
  - Features: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$.
  - Response/class: $y_1, y_2, \ldots, y_n$.



Mathematical abstraction: Empirical Risk Minimization



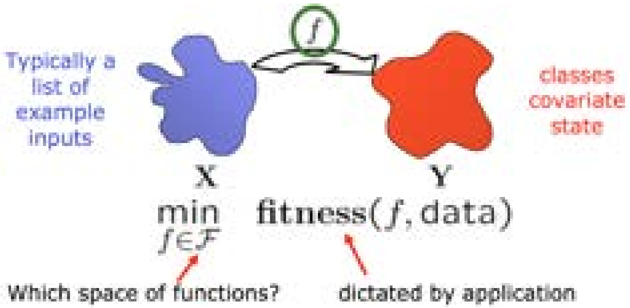
Mathematical abstraction: Empirical Risk Minimization



If we had infinite input/output data according to a distribution $\mathcal{D}$

$$\boldsymbol{\theta}^* = \arg\min_{\theta \in \Theta} \bar{\mathcal{L}}(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)]$$

Mathematical abstraction: Empirical Risk Minimization



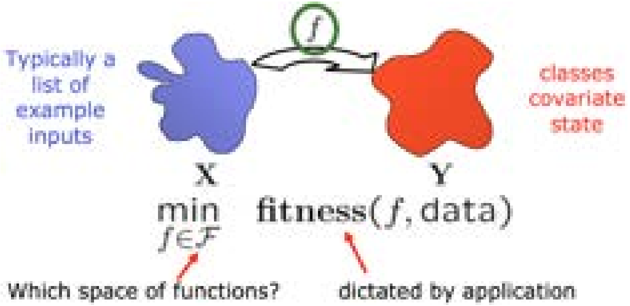
If we had infinite input/output data according to a distribution $\mathcal{D}$

$$\boldsymbol{\theta}^* = \underset{\theta \in \Theta}{\arg\min}\ \bar{\mathcal{L}}(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)]$$

Given a data set $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ find the best function that fits this data

$$\hat{\boldsymbol{\theta}} = \underset{\theta \in \Theta}{\arg\min}\ \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)$$

Mathematical abstraction: Empirical Risk Minimization



If we had infinite input/output data according to a distribution $\mathcal{D}$

$$\boldsymbol{\theta}^* = \arg\min_{\theta \in \Theta} \bar{\mathcal{L}}(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)]$$

Given a data set $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ find the best function that fits this data

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta \in \Theta} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)$$

How good is $f_{\hat{\theta}}$? That is a new sample $\boldsymbol{x}$ how well can $f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ estimate y?

Uniform concentration

For a fixed $\boldsymbol{\theta}$ we know

$$\frac{1}{n}\sum_{i=1}^{n}\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i) \approx \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)].$$

Uniform concentration

For a fixed $\boldsymbol{\theta}$ we know

$$\frac{1}{n}\sum_{i=1}^{n}\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i) \approx \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)].$$

Questions?

- Can we guarantee this for all $\boldsymbol{\theta} \in \Theta$?

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i) - \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)]\right| \leq \delta$$

Uniform concentration

For a fixed $\boldsymbol{\theta}$ we know

$$\frac{1}{n}\sum_{i=1}^{n}\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i) \approx \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)].$$

Questions?

- Can we guarantee this for all $\boldsymbol{\theta} \in \Theta$?

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i) - \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)]\right| \leq \delta$$

much more difficult that proving this for one point asymptotically (law of large numbers)

Uniform concentration

For a fixed $\boldsymbol{\theta}$ we know

$$\frac{1}{n}\sum_{i=1}^{n}\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i) \approx \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)].$$

Questions?

- Can we guarantee this for all $\boldsymbol{\theta} \in \Theta$?

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i) - \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)]\right| \leq \delta$$

  much more difficult that proving this for one point asymptotically (law of large numbers)

- How many data samples do we need as a function of Θ, f and δ.

Course Logistics

Goals

- Learn modern techniques in probability
- Concentration in high-dimensions
- geared towards applications for data sciences, statistics and machine learning

Why is this course needed? Distinctions with other courses?

- Advanced probability courses (while very technical e.g. cover measure theory) do not cover some of the most useful techniques in modern probability
- Discuss analogous results to low-dimensions in high-dimensions (law of large numbers, concentration, etc.)
- Over the past 5-10 years there has been tremendous progress simplifying many proofs
- Focus on the most useful techniques

Background and disclaimer

- Prerequisites:
  - EE 599 enrollees (EE 441 and EE 503)
  - MATH 605 enrollees (MATH 505a or MATH 507a)

Background and disclaimer

- Prerequisites:
  - EE 599 enrollees (EE 441 and EE 503)
  - MATH 605 enrollees (MATH 505a or MATH 507a)
- More mathy than EE 441 and EE 503 (but not abstract e.g. don't care about measure theory, Hilbert spaces, and Banach spaces)
- We cover a lot of material and many applications

Background and disclaimer

- Prerequisites:
  - EE 599 enrollees (EE 441 and EE 503)
  - MATH 605 enrollees (MATH 505a or MATH 507a)
- More mathy than EE 441 and EE 503 (but not abstract e.g. don't care about measure theory, Hilbert spaces, and Banach spaces)
- We cover a lot of material and many applications
- Do I need to know these applications? Do I need to know measure theory or Morse theory? Do I need to be a math graduate student? Do I need to be an electrical engineering student?

Background and disclaimer

- Prerequisites:
  - EE 599 enrollees (EE 441 and EE 503)
  - MATH 605 enrollees (MATH 505a or MATH 507a)
- More mathy than EE 441 and EE 503 (but not abstract e.g. don't care about measure theory, Hilbert spaces, and Banach spaces)
- We cover a lot of material and many applications
- Do I need to know these applications? Do I need to know measure theory or Morse theory? Do I need to be a math graduate student? Do I need to be an electrical engineering student?
- **Answer:** Absolutely not.

Logistics

- Class: Mon, Wed 10:30-11:50 PM, VKC 256.
- Instructor office hours:
  - Larry: Monday 12-1:30, Wednesday 3:30-5, KAP 406D
  - Mahdi: Monday and Wednesday 5:30-7 PM EEB 422
- Course website: blackboard
- Grading
  - % 10 participation
  - % 90 Homework

Logistics

- Class: Mon, Wed 10:30-11:50 PM, VKC 256.
- Instructor office hours:
  - Larry: Monday 12-1:30, Wednesday 3:30-5, KAP 406D
  - Mahdi: Monday and Wednesday 5:30-7 PM EEB 422
- Course website: blackboard
- Grading
  - % 10 participation
  - % 90 Homework
- Lowest homework will be dropped

Logistics

- Class: Mon, Wed 10:30-11:50 PM, VKC 256.
- Instructor office hours:
  - Larry: Monday 12-1:30, Wednesday 3:30-5, KAP 406D
  - Mahdi: Monday and Wednesday 5:30-7 PM EEB 422
- Course website: blackboard
- Grading
  - % 10 participation
  - % 90 Homework
- Lowest homework will be dropped
- We don't care where you find the solution just write the proof in your own language (no plagarism)
- Course Policy: Use of sources (people, books, internet, etc.) without citation results in failing grade.

Textbook

- Required textbook
  - High-Dimensional Probability: An Introduction with Applications in Data Science. Roman Vershynin.
- Additional textbooks
  - Concentration Inequalities: A Non-asymptotic Theory of Independence. Stephane Boucheron, Gabor Lugosi, Pascal Massart
  - The Concentration of measure phenomenon. Michel Ledoux

Why you should not take this class

- Probability is not your thing.
- Eclectic topics.

Why you should not take this class

- Probability is not your thing.
- Eclectic topics.
- This class is rated P ...


START