ECE6980

An Algorithmic and Information Theoretic Toolbox for Massive Data

Logistics

Instructor: Jayadev Acharya

Email: acharya@cornell.edu

Lectures: TuTh 1.25-2.40, 203 Phillips

Office Hours: MoTh 3-4, 304 Rhodes

Website:
http://people.csail.mit.edu/jayadev/ece6980

Grading

- Scribe a lecture: 10%
  - Encouraged to fill in the details, provide examples
- Assignments 30-60%
  - 2-3 assignments
  - Typeset?
- Project report and presentation: 40-60%
  - Read a new related paper
  - Present a summary in your own words
  - Can choose from a list
- Interruptions: 5%

Lectures

- Lectures primarily on the board
- Derive things (mostly from scratch)

Course overview

- Lot of interest in data science
  - Number of courses on offer
  - Many aspects can be covered
- This course:
  - Core primitives
  - Efficient algorithms
  - Fundamental limits
  - Mostly theoretical, encourage implementation

Prerequisites

- Undergraduate probability/random processes
- Basic combinatorics

- What is the variance of a random variable?
- What is a binomial distribution?
- When are two random variables independent?

What you should learn?

- Fast algorithms for statistical problems
  - Learning discrete distributions
  - Finite sample hypothesis testing
- How to prove information theoretic lower bounds

Probabilistic thinking

Old questions, new issues

Classical

Domain: 

$n = 1000$ tosses

Small domain D

$n\ large, |D|\ small$

Asymptotic analysis
Computation **not crucial**

Modern

Domain: 

One human genome

Large domain D

$n\ small, |D|\ large$

New challenges

Resources

Samples

- How much data needed?
- Inference when data is **scarce**

Computation

- How does run-time scale with data and domain size?
- Even quadratic might be **prohibitive**

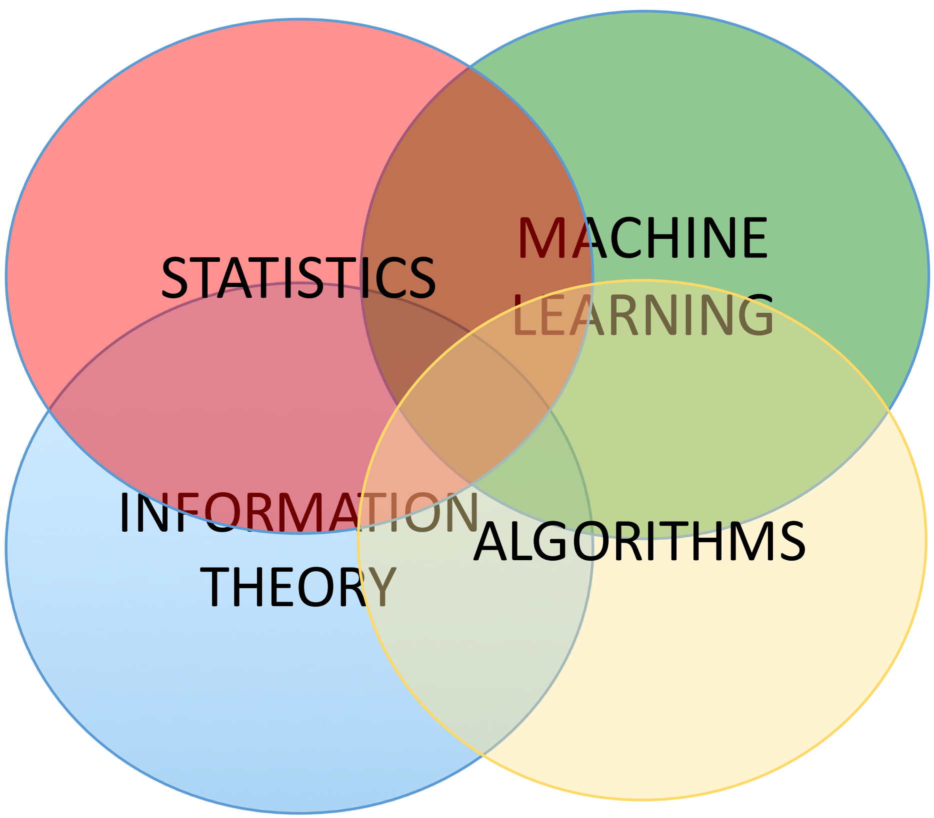
Other resources

- Storage: Not enough space to store all data
- Communication: Distributed data across servers

Goals

For statistical inference

- **Design** efficient algorithms
- **Understand** fundamental limits



Distribution learning

A simple setting

- Support set $\mathcal{X}$
- Distribution $p: \mathcal{X} \rightarrow \mathbb{R}_{\geq 0,}$ such that $\sum_{x \in \mathcal{X}} p(x) = 1$
- Samples $x^n = x_1 x_2 \ldots x_n$ drawn from p
- Output a distribution $q(x^n)$ after observing x^n

Toss a coin: *H T T T H T T H*

Throw a die: 3 1 3 4 4 5 3 6

What is a good estimator

- Would like q to be close to p
- $L(p, q)$: Loss for estimating p with q
  - Total variation distance, KL divergence, …
- Find an estimator with small L
- For a given loss function, how many samples needed?

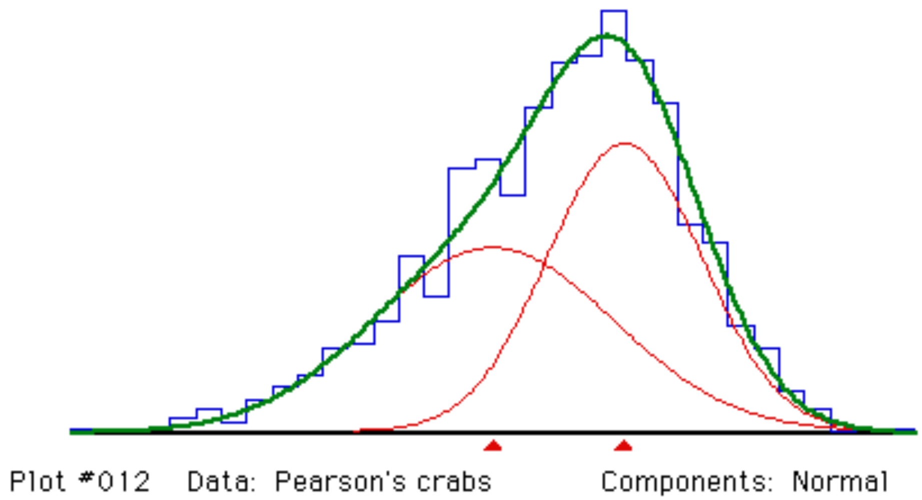
- Empirical estimators:
  - $q(H) = \frac{3}{8}$
- Analyze the performance of empirical estimators

Learning

- Given samples from a Gaussian distribution $N(\mu, \sigma^2)$
- Learn with a Gaussian distribution

- Relatively simple

Learning



Ratio of breadth to height of 1000 crabs by W. Weldon

Not normally distributed, more than one species?

Karl Pearson: Mixtures of Gaussians (much harder!!)

Distribution testing

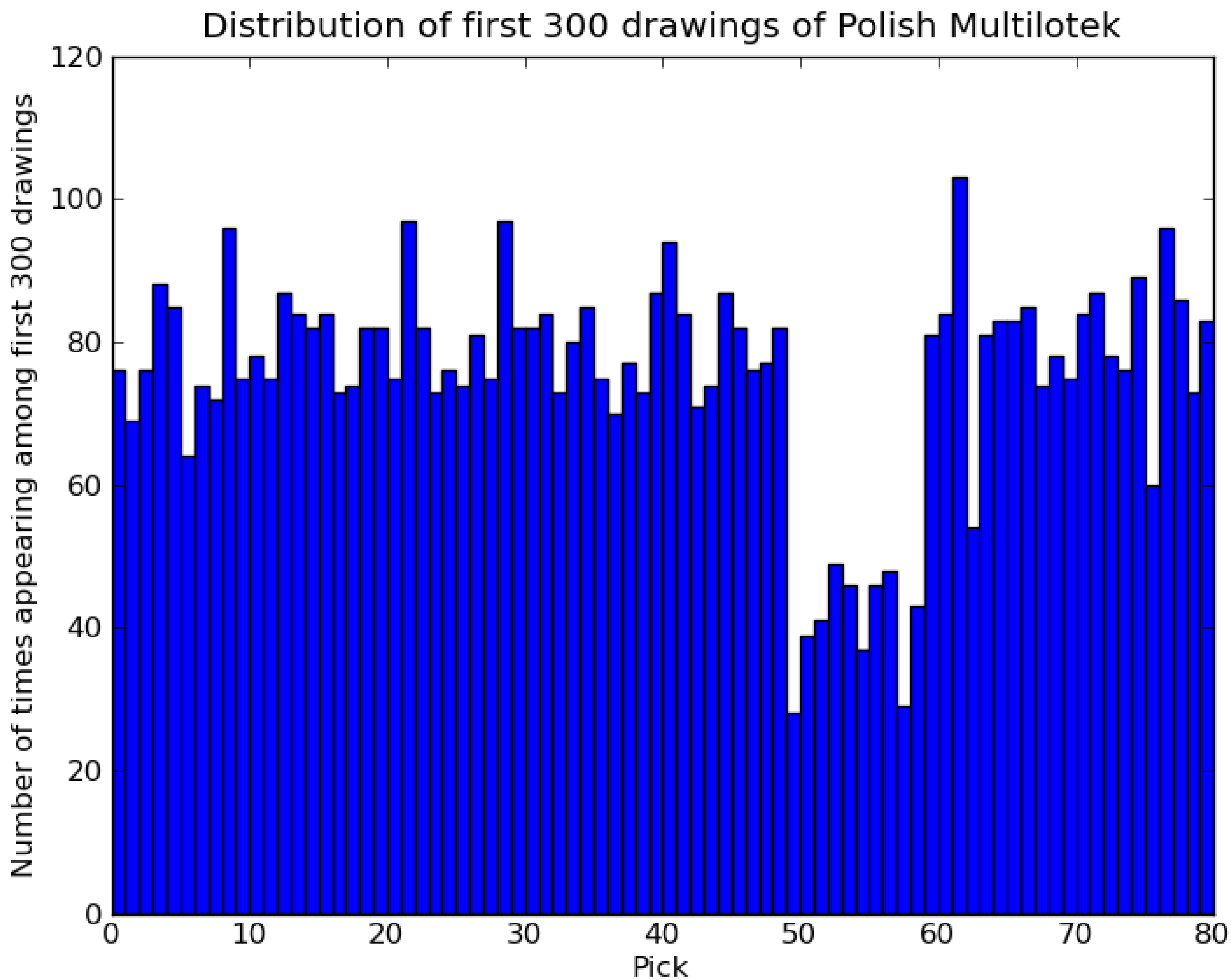
Testing uniformity

Polish Multilotek:

- Picks 20 numbers between 1,…,80

Is it fair?

Testing uniformity (contd)



(Figure by Onak, Price, Rubinfeld)

A simple setting

- $|\mathcal{X}| = k$
- u: uniform distribution over $\mathcal{X}$
- x^n: n samples from a distribution p

Question: Is $p = u$ OR $|p - u|_1 \geq \varepsilon$?

How many samples do we need? Take a guess …

A simple classification problem

X^n: a sports article

Y^n: a religious article

Z: one word

Q: Is Z more likely to appear in sports or religion?

Necessarily assign Z to where it appears **more often**?

Property estimation

Predicting new elements

. . .

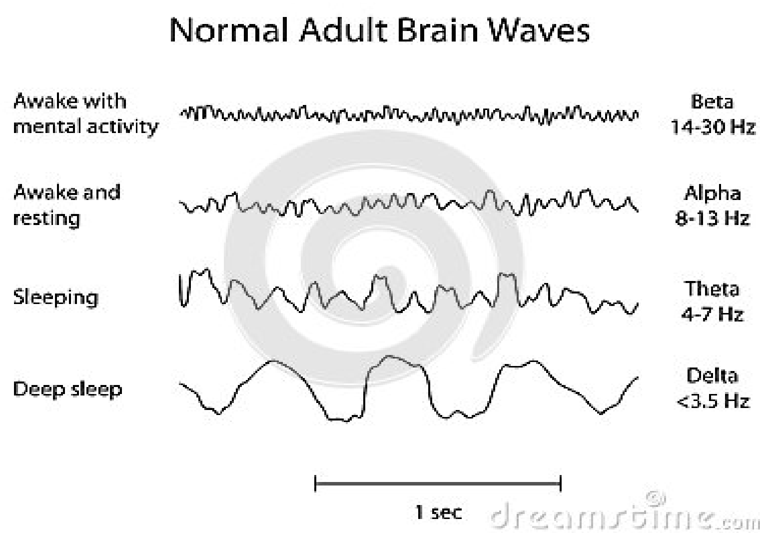
How many **new species**?

Corbet collected butterflies in Malaya for one year

Frequency	1	2	3	4	5	6	7	..
Species	118	74	44	24	29	22	20	..

How many new species if he goes for one more year?

Entropy estimation



How much **randomness** in neural spikes?

How to estimate **entropy** from observations?

Entropy estimation

- $|\mathcal{X}| = k$
- x^n: n samples from a distribution p

Question: Estimate $H(p)$

Resource constraints

sample stream

decision

limited memory

distributed data

limited communication

Data too big to be stored in a single machine
Lot of recent interest