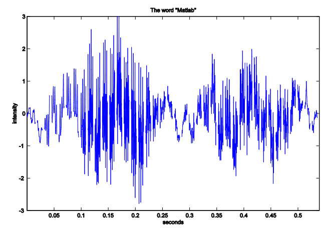
Some Possible Paths Ahead in Estimation, Inference, and Learning

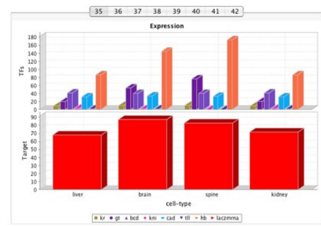
Sanjeev Kulkarni
Department of Electrical Engineering
Princeton University

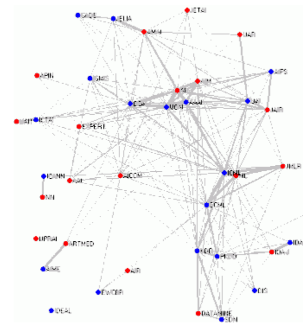
November 12-14, 2009

Some Basic Learning Problems

Some Typical Inputs



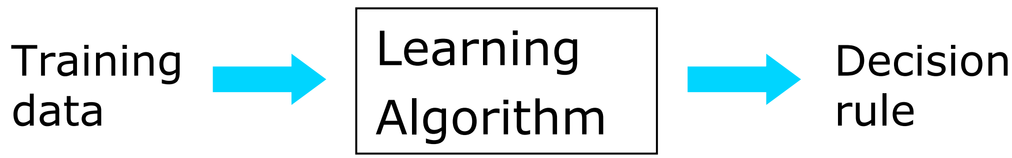




Some Typical Tasks and Applications

- Classification, estimation, adaptation, search, optimization, reinforcement learning, etc.
- Applications such as face/character/target detection and recognition, speech recognition, medical diagnosis, statistical arbitrage, etc.

Classical Paradigm for Supervised Learning
(Nonparametric Estimation/Classification)



- Features X (often in R^d) and label Y (often in R or {0,1}).
- Given training examples $(X_1,Y_1),\ldots,(X_n,Y_n)$, *i.i.d.* $\sim P(X,Y)$.
- Design rule $g{:}X \rightarrow Y$ to predict outputs from observed features that minimizes prediction error $\mathbf{E}\{|g(X)-Y|^2\}$
- Many techniques to choose from, theoretical results to go along, and success in a wide range of applications.
- So, we're all set. Or are we?

We're Not Where We Want to Be

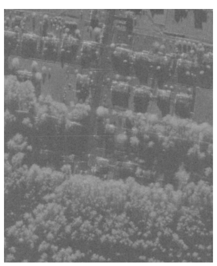
- Standard applications/methods are too contrived/neat/constrained.
- Consider detection of high threats, anomalous behavior, or IEDs:

Video sureveillance

Calendar/Time Information

Context Models

Financial transactions



SAR imagery

Intelligence info

Cell traffic

Learning Algorithm

Detect
- IED
- Anomalous Behavior
- High threat situation

- Wishful thinking for now!
- What are some obstacles (and corresponding opportunities) that lie in the path ahead?

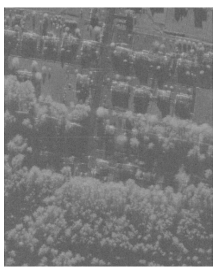
Obstacle/Opportunity 1: Aggregation

Video sureveillance

Calendar/Time Information

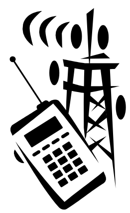
Context Models

Financial transactions



SAR imagery

Intelligence info



Cell traffic

- Data is wildly heterogeneous.
  - Signals, symbols
  - contextual, relational, conceptual
  - raw, processed
  - hard, soft
  - regular, sporadic
  - abundant, scarce
  - local, global, etc.
- Fusion? At what level? How??

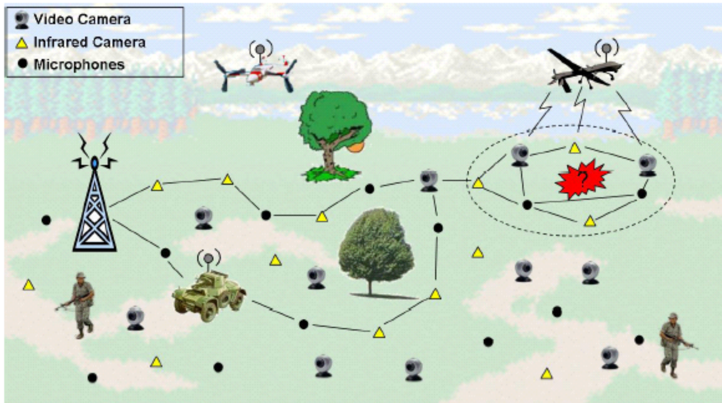
- Need methods for modeling, data representation, aggregation.
- Even simple, canonical problems would be helpful.

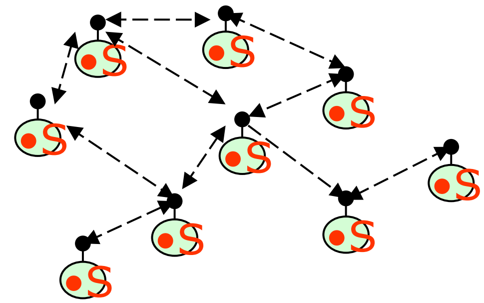
Aggregation continued

Example: Aggregating Probability Forecasts from Multiple Agents
(w/ Predd, Wang, Osherson, Poor)

- Collection of agents/sensors provide probability forecasts $(\phi_1,p_1),\ldots,(\phi_m,p_m)$ where the ϕ are conjunctions, disjunction, negations over a common set of basic events $(X_1,\ldots,X_n)$.
- Can we aggregate these into a single, coherent set of probability forecasts for the events? Provably increases stochastic accuracy.
- Election data: 30,000 individuals provided half million judgments such as P(Obama wins PA or NY), P(McCain wins VA | McCain wins IL).
- Some directions of interest:
  - Repeated trials: adapt to performance of agents
  - Aggregate samples with forecasts
  - Human decision-making

Obstacle/Opportunity 2: Distributed Data





From E. Ekici, http://www.ece.osu.edu/~ekici/images/mwsn.jpg

- Data comes from multiple, distributed sources.
- There may be communication, computation, confidentiality issues.
- Who should send data to whom? What should they send?
- Not just maximizing throughput – joint objective (e.g., consensus, global classifier, field estimation, outlier/anomaly detection, etc.).
- Theory for distributed/networked learning?

Obstacle/Opportunity 3: Scaling Issues

- Standard setting: fix distribution P and then let # of examples $n \rightarrow \infty$.
- Is this the right asymptotic regime?
- Is more data better? Are more features better?
- Are more sensors better? Is more connectivity better?

Other variables that might scale with n:

- Alphabet size A
- Dimension d
- Intrinsic dimension d'
- Number of classes m
- Types of examples $n_1, \ldots, n_s$
- Number of sensors k
- Connectivity

Scaling Issues continued

Example: Natural Language

- 100 characters, 10th-order Markov → 100^{11} transition probabilities.
- Words capture structure, but about 1,000,000 words → causes different problems.
- With small corpus, empirical probabilities give poor estimates.
- Rare-events regime: alphabet $|A_n|$ grows so $|A_n|/n$ → constant.

How many words did Shakespeare know?

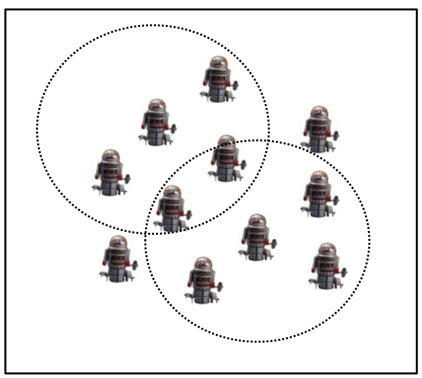
- Corpus: N=884,647 total words; 31,534 distinct words.
- N_1=14,367 words occurred just once.
- Good-Turing estimator: P(unseen words) = N_1/N.

- Further analysis: Shakespeare knew >35,000 more words (Efron & Thisted).
- Modified versions give consistent estimators in rare-events regime (with Wagner, Viswanath).

Obstacle/Opportunity 4: Active Learning

- For better learning, sensing could be (should be?) active.
- Active learning can help address the other issues – right data can change asymptotic scaling, help aggregate, improve coordination
- Active/adaptive/cooperative sampling can significantly improve rates in learning
- Would like joint consideration of learning, control, information, networks

Example: Adaptive field estimation by multiple agents



- Mobile agents with communication constraints.
- Collect noisy samples as they roam.
- Cooperatively control to estimate field.
- Methods for learning, control, communication?
- Fundamental limits?

Summary of Some Paths Ahead: Obstacles and Opportunities

- Scaling issues
- Aggregation
- Networked/Distributed Learning
- Joint consideration of learning, control, information, networks
- Optimistic for significant advances
- Yet, will keep us busy for a long time

Summary of Some Paths Ahead: Obstacles and Opportunities

- Scaling issues
- Aggregation
- Networked/Distributed Learning
- Joint consideration of learning, control, information, networks
- Optimistic for significant advances
- Yet, will keep us busy for a long time