Machine Learning

Michael I. Jordan
University of California, Berkeley
A Machine Learning Syllabus

- Classification
- Regression
- Clustering
- Dimensionality reduction
- Feature selection
- Cross-validation, bootstrap
- Hidden Markov models, graphical models
- Visualization and nonlinear dimensionality reduction
- Collaborative filtering
- Reinforcement learning
- Time series, sequential hypothesis testing, anomaly detection
- Nonparametric Bayesian methods
- Active learning, experimental design
- Multi-class classification, structured classification
Machine Learning
Statistical Inference and Decision Making
Statistical Inference and Decision Making

- information theory
- control theory
- optimization
- algorithms
- databases
- signal processing
- economics
- statistical physics
Statistical Inference and Decision Making

- information theory
- control theory
- optimization
- algorithms
- signal processing
- economics
- statistical physics
- databases
Some Recent Success Stories

- Classification
- Kernel methods and manifold learning
- Topic models
- Graphical models
- Nonparametrics
- Bayesian nonparametrics
- Reinforcement learning
- Applications in computational vision, natural language processing, information retrieval, robotics, computational biology, control of data centers, etc
Current Trends and Issues in Inference and Decision Making

- Nonparametric Bayes
- Massive data sets
- End-to-end objective functions
- Objective Bayes
- Sparsity and beyond
- Connections to control theory
Bayesian Nonparametrics

- Stochastic processes as priors; i.e., prior distributions on objects such as:
  - partitions (*Dirichlet processes*)
  - trees and graphs (*nested and hierarchical DPs*)
  - combinatorial state spaces (*Beta processes*)
  - hazard functions (*Beta processes*)
  - regression functions (*Gaussian processes*)
  - distribution functions (*subordinators*)
  - measures (*completely random measures*)

- Somewhat surprisingly, there are efficient ways to update these priors into posteriors
Bayesian Nonparametrics

- Stochastic processes as priors; i.e., prior distributions on objects such as:
  - partitions (Dirichlet processes)
  - trees and graphs (nested and hierarchical DPs)
  - combinatorial state spaces (Beta processes)
  - hazard functions (Beta processes)
  - regression functions (Gaussian processes)
  - distribution functions (subordinators)
  - measures (completely random measures)

- Somewhat surprisingly, there are efficient ways to update these priors into posteriors
  - but you need to know about sigma algebras to understand how that’s possible
Bayesian Nonparametrics

• Can cope in principle with a number of classical difficulties
  • no more fixed-length feature vectors
  • cardinality of state space can be unknown a priori
  • combinatorial state spaces
  • robustness to distributional assumptions
  • easy to make use of hierarchies (e.g., “transfer learning”)
  • nonstationarity (in space and time)

• Some real success stories
  • protein modeling
  • statistical genetics
  • speech diarization
  • motion capture analysis
Speaker Diarization
Motion Capture Analysis

- Goal: Find coherent “behaviors” in the time series that transfer to other time series (e.g., jumping, reaching)
Completely Random Measures

(Kingman, Pitman, etc)

- Completely random measures are measures on a set $\Omega$ that assign independent mass to nonintersecting subsets of $\Omega$
  - e.g., Brownian motion, gamma processes, beta processes, compound Poisson processes and limits thereof
- (The Dirichlet process is not a completely random measure
  - but it's a normalized gamma process)
- Completely random measures are discrete wp1 (up to a possible deterministic continuous component)
- Completely random measures are random *measures*, not necessarily random *probability measures*
Completely Random Measures

- Consider a non-homogeneous Poisson process on $\Omega \otimes \mathbb{R}$, with rate function obtained from some product measure.
- Sample from this Poisson process and connect the samples vertically to their coordinates in $\Omega$. 
Beta Processes

- The product measure is called a *Levy measure*.
- For the beta process, this measure is defined on the product space $\Omega \otimes (0, 1)$ and is as follows:

$$\nu(d\omega, dp) = cp^{-1}(1 - p)^{c-1}dp B_0(d\omega)$$

- degenerate Beta$(0,c)$ distribution  
- Base measure

- And the resulting random measure can be written simply as:

$$B = \sum_{i} p_i \delta\omega_i$$
\[ B = \sum_{i} p_i \delta \omega_i \]
Beta Process and Bernoulli Process

Concentration $c = 10$  Mass $\gamma = 2$

[Graph showing concentration and mass with data points and a line graph indicating the process behavior.]
• Beta process prior:
  • sparsity
  • encourages sharing
  • allows variability

• Bernoulli process determines which states are used
Massive Data Sets

• A massive *embarassment*
• The classical perspective in machine learning: each year our algorithms get better and better, and we can handle ever larger training sets
• But why can’t we handle arbitrarily large data sets now?
Massive Data Sets

- A massive **embarrassment**
- The classical perspective in machine learning: each year our algorithms get better and better, and we can handle ever larger training sets
- But why can’t we handle arbitrarily large data sets **now**?
  - need general methods (and theory) for backing off to simpler procedures as data accrue, *so that statistical risk decreases under a fixed computational budget*
Massive Data Sets

- A massive *embarassment*
- The classical perspective in machine learning: each year our algorithms get better and better, and we can handle ever larger training sets
- But why can’t we handle arbitrarily large data sets *now*?
  - need general methods (and theory) for backing off to simpler procedures as data accrue, *so that statistical risk decreases under a fixed computational budget*
  - a “simpler procedure” may be a pre-processor that allows us to use more complex procedures cheaply
Massive Data Sets

- A massive *embarassment*
- The classical perspective in machine learning: each year our algorithms get better and better, and we can handle ever larger training sets
- But why can’t we handle arbitrarily large data sets *now*?
  - need general methods (and theory) for backing off to simpler procedures as data accrue, *so that statistical risk decreases under a fixed computational budget*
  - a “simpler procedure” may be a pre-processor that allows us to use more complex procedures cheaply
  - need general methods (and theory) for throwing away data
End-to-End Objective Functions

• A major current direction in machine learning: given a system composed of modules, train the modules so as to minimize an overall loss
• E.g., dimension reduction in regression:
  • old style: compress with the SVD; build a kernel regression on the compressed representation
  • new style: find a surrogate for the regression that allows the compression to be adapted to the regression
End-to-End Objective Functions

• A major current direction in machine learning: given a system composed of modules, train the modules so as to minimize an overall loss

• E.g., dimension reduction in regression:
  • old style: compress with the SVD; build a kernel regression on the compressed representation
  • new style: find a surrogate for the regression that allows the compression to be adapted to the regression

• There is a general problem here that involves finding surrogates for optimizing certain kinds of losses in certain kinds of composite systems
  • can this be a collaborative project with control theory?
Objective Bayes

- Bayesian methods have many favorable properties, but subjective Bayesian methods don’t scale.
- The frequentist dictum: “Let the data speak”
Objective Bayes

- Bayesian methods have many favorable properties, but subjective Bayesian methods don’t scale.
- The frequentist dictum: “Let the data speak”
- *Objective Bayes* is a unifying force in inference that uses frequentist tools in defining priors to achieve these goals.
- Lovely connections to information theory.
- In my view one of the major directions in statistics in the next few decades.
Sparsity and Beyond

- If there exists a sparse representation in some basis, we have an increasingly strong theory that guarantees that certain classes of algorithms can discover that representation
- I’ll let Martin W. elaborate
- It would be desirable to find such bases automatically
- Other concepts that allow us to make progress in the high-dimensional regime?
Connections to Control Theory

- Control theory and statistics are two of the deepest disciplines around
  - they go all the way from theory to practice; they embrace science and engineering
  - they have provided the most useful insights in humankind’s first attempts to understand “intelligence”
Connections to Control Theory

• Control theory and statistics are two of the deepest disciplines around
  • they go all the way from theory to practice; they embrace science and engineering
  • they have provided the most useful insights in humankind’s first attempts to understand “intelligence”

• They are complementary but they have been surprisingly loathe to embrace one another
Connections to Control Theory

• Control theory and statistics are two of the deepest disciplines around
  • they go all the way from theory to practice; they embrace science and engineering
  • they have provided the most useful insights in humankind’s first attempts to understand “intelligence”

• They are complementary but they have been surprisingly loathe to embrace one another

• In the meantime, machine learning and optimization have been having quite a little love affair
Connections to Control Theory

- Control theory and statistics are two of the deepest disciplines around
  - they go all the way from theory to practice; they embrace science and engineering
  - they have provided the most useful insights in humankind’s first attempts to understand “intelligence”
- They are complementary but they have been surprisingly loathe to embrace one another
- In the meantime, machine learning and optimization have been having quite a little love affair
  - damned upstarts…
Connections to Control Theory

• Control theory and statistics are two of the deepest disciplines around
  • they go all the way from theory to practice; they embrace science and engineering
  • they have provided the most useful insights in humankind’s first attempts to understand “intelligence”

• They are complementary but they have been surprisingly loathe to embrace one another

• In the meantime, machine learning and optimization have been having quite a little love affair
  • damned upstarts…

• No, control isn’t just “statistics + optimization”, but that combination is a powerful one that should be a major part of the control landscape