Lecture 11
Closing Remarks

(March 27, 2015)

Mu Zhu
University of Waterloo
1/09: Basics

bias-variance trade-off; curse of dimensionality

1/16: Optimization

\( \ell_1 \) norm; nuclear norm; convex relaxation; coordinate descent

1/23: Unsupervised Learning

latent variables; EM algorithm; “borrow strength”

1/30: Towards Deep Learning

restricted Boltzmann machine

Gibbs sampler; gradient descent; quasi-Newton

2/06: Ensembles

strength-diversity trade-off; functional gradient descent

2/13: Kernel Machines

VC-theory; KKT conditions; RKHS; “kernel trick”
Remarks on “Basics”

• important ideas that I didn’t cover specifically
  – over fitting
  – cross validation

• talk by Hugh Chipman @ Fields-OCBC

OCBC = Opening Conference and Boot Camp
Remarks on “Basics”

• machine learning
  – given some data, try to learn something from them

• statistics
  – given a question, try to find an answer to it
Remarks on “Basics”

• if goal = answer an underlying question, then, most effective to think about the “whole package”:

  data collection
  (sampling, design)

  ↓

  modeling

  ↓

  estimation, inference

• shouldn’t forget what makes statistics a unique discipline and what it does best
Remarks on “Basics”

- but this doesn’t mean we shouldn’t try to discover potentially interesting information from existing (observational) data

- actually, seems awfully wasteful if we don’t

- just have to be VERY careful with what we can conclude, and not confuse the two very different types of objectives
Ex I: Replication Crisis in Science


- 12 clinical trials between 1990 and 2010
- tested 52 scientific claims about the health benefits (or hazards) of vitamin E, vitamin D, calcium, selenium, hormone replacement therapy, folic acid, beta-carotene, and so on
- unable to replicate ANY of the 52 claims
Ex I: Replication Crisis in Science

- e.g., one study followed ∼ 87,000 women for ∼ 8 years

- found ∼ 11,500 who took vitamin E supplement regularly (not a randomized assignment) had ∼ 31% reduction in relative risk for nonfatal myocardial infarction and death from cardiovascular disease

- later found reduction in risk had nothing to do with taking vitamin E supplements

  ⇓  ⇓  ⇓

- multiple testing, another “Big Data” problem that I would have enjoyed discussing but didn’t have time for
Large-Scale Inference


- testing thousands of hypotheses at the same time
- e.g., 1000 hypotheses, each tested at significance level of 0.05 ⇒ expect to find 50 “significant” hypotheses just by chance, even if none of them is
- not so much machine learning, but definitely big data
Large-Scale Inference

- **false discovery rate (FDR)** rather than **type I error**, i.e.,
  \[ P(\text{null} | \text{significant}) \text{ rather than } P(\text{significant} | \text{null}) \]

- to control FDR @ level \( \alpha \) when testing \( m \) hypotheses simultaneously, cutoff @
  \[ [\text{p-value}]_{(k)} \leq (k \alpha)/m, \]
  *if the \( m \) tests are independent* (Benjamini-Hochberg)

- contrast with **Bonferroni** [focusing on \( P(\text{significant} | \text{null}) \)], which uses
  \[ [\text{p-value}]_{k} \leq \alpha/m \]
  *for all* \( 1 \leq k \leq m \)
Ex II: A More Positive Story

- Computer algorithm to assess heart-attack risk (essentially a decision tree based on a few thousand training samples), developed in the 1980s
- Doctors refused to believe it
- When used ~ 20 years later, it made better assessments than MDs in the ER
Remarks on “Optimization”

• data w/ noise ⇒ pointless to optimize “too well”
  – e.g., stop early; take just one gradient/Newton step; ...

• talk by Martin Wainwright @ Fields-OCBC:

\[
\| \theta^{(t)} - \hat{\theta} \| \quad \text{versus} \quad \| \theta^{(t)} - \theta^* \|
\]

where

\[
\theta^* \equiv \arg \max_{\theta} \mathbb{E}[\ell(\theta; x, y)]
\]

\[
\hat{\theta} \equiv \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; x_i, y_i) - J_\lambda(\theta)
\]

\[
\theta^{(t)} \equiv \text{estimate at iteration } t
\]
Remarks on “Unsupervised Learning”

• topic models for text data: a big field in machine learning

• mostly Bayesian

• seminal work: latent Dirichlet allocation

• viewpoint adopted in this course:
  – a particular mixture of multinomials
  – instead of EM, use Bayesian model fitting with priors
  – use of priors allows us to “borrow strength”
Remarks on “Unsupervised Learning”

Exercise

(a) Implement an EM algorithm to fit the mixture model used in latent Dirichlet allocation (LDA), and compare EM with LDA.

(b) Experiment with some modifications to your EM algorithm — e.g., by adding penalties to $(\pi_{i1}, \pi_{i2}, ..., \pi_{iK})$ and/or $(\theta_{k1}, \theta_{k2}, ..., \theta_{kd})$ — to see if you can improve your results.

Suggestion  Not unreasonable for a course project.
Reminder

• collection of text documents, \( \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \)

• \( \mathbf{x}_i: \ (x_{i1}, \ldots, x_{im_i})^T \)

• \( m_i: \) number of words in document \( i \)

• \( x_{it} \in \{1, 2, \ldots, d\}: \) word \( t \) in document \( i \) ...

• mixture model

\[
\mathbf{x}_i \sim \prod_{t=1}^{m_i} \left[ \sum_{k=1}^{K} \pi_{ik} p(x_{it}; \boldsymbol{\theta}_k) \right], \quad p(x_{it}; \boldsymbol{\theta}_k) = \prod_{j=1}^{d} \theta_{kj} I(x_{it} = j),
\]

where each \( \theta_{kj} \) is the group-specific probability for word \( j \)

• \( \pi_{ik} \geq 0; \pi_{i1} + \ldots + \pi_{iK} = 1 \ \forall \ i; \) likewise for \( \theta_{kj} \)
Remarks on “Towards Deep Learning”

- deep learning: another big field in machine learning

- important role played by latent variables (hidden nodes) and unsupervised learning [betting “in the money” because most data are unlabelled]

- “obvious” connection to data visualization:
  - PCA arguably the most widely used tool
  - doesn’t really work for mixed data (continuous + discrete)
  - fit RBM instead (mix of binary + Gaussian nodes)

Suggestion Not unreasonable for a course project.
Remarks on “Ensembles”

- noticeable contrast:
  - practical impact ... huge
  - literature generated ... not as much

- not all researchers as excited about it as I am

- for some, lesson from Netflix contest is “disappointing”

- people are uncomfortable with certain implications

History Lessons  Why were we so shaken by the Copernican and Darwinian revolutions? What were the forces behind the social Darwinism movement?
Ex III: An Interesting Debate


Remarks on “Kernel Machines”

• opposite contrast:
  
  practical impact ... not as much
  literature generated ... huge

• the “fashion” is waning
  
  – often, easier to think in terms of features (variables)
  – e.g., proteins (sequence data)

• still very well suited for SOME problems
  
  – sometimes, easier to think in terms of similarities (kernels)
  – e.g., proteins (structural data, angular); networks
Some Nice Books to Read

- Past, Present, and Future of Statistical Science
- Statistics in Action: A Canadian Outlook
- Statistics in the 21st Century
My Experience

• learned something new [although didn’t achieve deep learning]
  – e.g., latent Dirichlet allocation; RBM

• renewed faith in fundamental statistical ideas
  – e.g., sampling, design

• improved understanding of what I thought I already knew well
  – e.g., primal-dual; KKT conditions; Breiman’s theorem
Your Experience

- most important technical know-how
- most impressive topic
- idea that most profoundly influenced your thinking

tell us @ https://www.surveymonkey.com/r/YDGVSWL
a good way to review the materials and prepare for the test
Sample Test Question I

The lecture on “unsupervised learning” (January 23, 2015) touched upon all of the following topics EXCEPT

A. principal component analysis (PCA).
B. the $K$-means algorithm.
C. mixture models.
D. latent Dirichlet allocation.
Sample Test Question II

Consider the penalized regression problem,

$$\min_{\beta_1, \ldots, \beta_d} \|y - (\beta_1 x_1 + \ldots + \beta_d x_d)\|^2 + \lambda \sum_{j=1}^{d} |\beta_j|^\alpha,$$

where $\lambda > 0$ is a fixed constant, and $y, x_1, \ldots, x_d \in \mathbb{R}^n$ are all “properly” standardized. This problem is NP-hard for

A. $\alpha = 2$.
B. $\alpha = 1$.
C. $\alpha = 0$.
D. any $\alpha$. 