Introduction

Le Song

Machine Learning
CS 7641, CSE/ISYE 6740, Fall 2015
What is machine learning (ML)

- Study of algorithms that improve their **performance** at some task with **experience**
Common to industrial scale problems

- 13 million wikipedia pages
- 800 million users
- 6 billion photos
- 340 million tweets per day
- 24 hours video uploaded per minutes
- > 1 trillion webpages

Data from 2014
Increasingly relevant to science problems
Syllabus

Cover a number of most commonly used machine learning algorithms in sufficient amount of details on their mechanisms.

Organization

Unsupervised learning (data exploration)
- Learning without labels or without optimizing for predictive task

Supervised learning (predictive models)
- Learning with labels, focusing on predictive performance

Complex models (dealing with nonlinearity, combine models etc)
- Nonlinearity, complex dependency, real world applications

Basics and Breadth (guest lecture and seminars)
Syllabus: unsupervised learning

- Learning without labels or without optimizing for predictive task
  - Clustering vectorial data
    - Kmeans
    - Hierarchical clustering
  - Clustering networks
    - Spectral algorithm
  - Dimensionality reduction,
    - Principal component analysis
  - Dimensionality for manifold
    - Locally linear embedding
  - Density estimation
    - Feature selection
    - Novelty/abnormality detection
Organizing Images

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
Find community in social networks

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
Each image has thousands or millions of pixels.

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
Shape of data

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
Feature selection

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Find abnormal object
Syllabus: supervised learning

- Learning with labels, focusing on predictive performance
  - Classifications
    - Nearest neighbor classifier
    - Naïve Bayes classifier
    - Logistic regression
    - Support vector machine
  - Combined classifiers
    - Boosting
  - Regressions
    - Ridge regression
    - Cross-validation
Image classification

What are the desired outcomes?
What are the inputs (data)?
What are the learning paradigms?
Face Detection

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
Weather Prediction

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

Predict

Numeric values:
40 F
Wind: NE at 14 km/h
Humidity: 83%
Understanding brain activity

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
Nonlinear classifier

Nonlinear Decision Boundaries

Kernel methods

Neural Networks

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
Syllabus: advanced topics, complex models

- Nonlinearity, complex dependency, real world applications
  - Kernel methods
- Hidden Markov models
- Graphical models
- Topic modeling
- Social network analysis
- Collaborative filtering
Handwritten digit recognition/text annotation

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
“Machine Learning is the preferred method for speech recognition ...”
Spam Filtering

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
Webpage classification

Company homepage vs. University homepage

Le Song
CSE, College of Computing
Georgia Institute of Technology
1340 Klaus Building
266 Ferst Drive
Atlanta, GA 30332, USA
email: le.song@ccc.gatech.edu

Research Interests
I conduct research in machine learning, with primary interests in non-parametric methods, probabilistic graphical models, time series and network analysis. I am also interested in large-scale and distributed learning problems, and machine learning applications in texts, images, networks, computational biology and information rich social media.

Selected Recent Publications

Mariya Ishteva
Address:
School of Computational Science and Engineering
College of Computing, 266 Ferst Drive
Georgia Institute of Technology
Atlanta, GA 30332, USA
Office: 1115, Center for Computational Engineering Science
E-mail: mariya@gatech.edu

Contact Information
Curriculum Vitae
Publication List

Updated 9 Oct 2011
Organizing documents

- Reading, digesting, and categorizing a vast text database is too much for human!

We want:

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANAGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
Robot Control

- Now cars can find their own ways!

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?
Basics/Prerequisites

- Probabilities
  - Distributions, densities, marginalization, conditioning

- Statistics
  - Mean, variance, maximum likelihood estimation

- Linear algebra
  - Vector, matrix, multiplication, inversion, eigen-decomposition

- Algorithms and Programming
  - Matlab, Basic data structures, computational complexity

- Convex optimization
  - Basics will be covered during lecture
Machine learning for apartment hunting

- Suppose you are to move to Atlanta
- And you want to find the **most reasonably priced** apartment satisfying your needs:
  - square-ft., # of bedroom, distance to campus ...

<table>
<thead>
<tr>
<th>Living area (ft$^2$)</th>
<th># bedroom</th>
<th>Rent ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>230</td>
<td>1</td>
<td>600</td>
</tr>
<tr>
<td>506</td>
<td>2</td>
<td>1000</td>
</tr>
<tr>
<td>433</td>
<td>2</td>
<td>1100</td>
</tr>
<tr>
<td>109</td>
<td>1</td>
<td>500</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>270</td>
<td>1.5</td>
<td>?</td>
</tr>
</tbody>
</table>
Assume $y$ is a linear function of $x$ (features) plus noise $\epsilon$

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n + \epsilon$$

where $\epsilon$ is an error model as Gaussian $N(0, \sigma^2)$

Let $\theta = (\theta_0, \theta_1, \ldots, \theta_n)^T$, and augment data by one dimension

$$x \leftarrow (1, x)^T$$

Then $y = \theta^T x + \epsilon$
Least mean square method

- Given m data points, find $\theta$ that minimizes the mean square error

$$\hat{\theta} = \arg\min_{\theta} L(\theta) = \frac{1}{m} \sum_{i} (y^i - \theta^T x^i)^2$$

- Set gradient to 0 and find parameter

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{m} \sum_{i} (y^i - \theta^T x^i)x^i = 0$$

$$\Leftrightarrow -\frac{2}{m} \sum_{i} y^i x^i + \frac{2}{m} \sum_{i} x^i x^i^T \theta = 0$$
Matrix version of the gradient

- Define \( X = (x^1, x^2, \ldots, x^m), y = (y^1, y^2, \ldots, y^m)^\top \), gradient becomes

\[
\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{m} X y + \frac{2}{m} X X^\top \theta
\]

\[\Rightarrow \hat{\theta} = (XX^\top)^{-1}Xy\]

- Matrix inversion in \( \hat{\theta} = (XX^\top)^{-1}Xy \) expensive to compute

- Gradient descent

\[\hat{\theta}^{t+1} \leftarrow \hat{\theta}^t + \frac{\alpha}{m} \sum_{i}^{m} \left(y^i - \hat{\theta}^t x^i\right) x^i\]
Probabilistic Interpretation of LMS

- Assume $y$ is a linear in $x$ plus noise $\epsilon$
  \[ y = \theta^\top x + \epsilon \]

- Assume $\epsilon$ follows a Gaussian $N(0,\sigma)$
  \[ p(y^i|x^i;\theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y^i - \theta^\top x^i)^2}{2\sigma^2}\right) \]

- By independence assumption, likelihood is
  \[ L(\theta) = \prod_{i}^{m} p(y^i|x^i;\theta) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^m \exp\left(-\frac{\sum_{i}^{m}(y^i - \theta^\top x^i)^2}{2\sigma^2}\right) \]
Probabilistic Interpretation of LMS, cont.

- Hence the log-likelihood is:

\[
\log L(\theta) = m \log \frac{1}{\sqrt{2\pi\sigma}} - \frac{1}{2\sigma^2} \sum_{i}^{m} (y^i - \theta^\top x^i)^2
\]

- LMS is equivalent to MLE of \( \theta \)!

\[
LMS: \quad \frac{1}{m} \sum_{i}^{m} (y^i - \theta^\top x^i)^2
\]

- How to make it work in real data?

**Statistics**

**Algorithms**

**Programming**
Textbooks

- No a single book which covers everything machine learning

- Pattern Recognition and Machine Learning, Chris Bishop
  - Written by leading industrial researcher
  - Presented from probabilistic and graphical model view

- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman
  - Written by leading statisticians
  - Presented from statistical view

- Machine Learning: a Probabilistic Perspective, Kevin Murphy
  - Written by previous university professor and now googler
  - More details on derivation and good for self studying
Grading

- 4 assignments (50%)
  - Approximately 1 assignment every 5 lectures
  - More next slide

- Midterm exam (20%)

- Final exam (20%)

- Background test (5%)
  - More in 2 slides

- Participation (5%)
  - guest lectures, seminar attendance, and class feedback.
Assignments

- Homework should be submitted before the deadline set in T-Square.

- No late submission will be accepted through email.

- We strongly encourage to use LaTeX for your submission.

- Any kind of academic misconduct is subject to F grade as well as reporting to the Dean of students.
  - All answers and codes should be prepared by yourself.
  - If you refer to any material, it should be properly cited.
  - Write on your homework anyone with whom you collaborate
Background test

- Background test in the second lecture
  - check your basic knowledge of probability and statistics, linear algebra, and matlab.

- Aim to check your readiness for the class
  - < 40%, the class may be too difficult, consider taking it next time
  - 40-79%, brush up your basic knowledge

- In both cases, need to attend a mandatory review section.
The team

- **Instructor**: Le Song, Klaus 1340

- **TA**: Bo Xie, Hanjun Dai, Ashwin Shenoi, Rashit Tevali

- **Guest Lecturer**: TBD

**Links:**
- TA contacts: cdaml2015@gmail.com
- Discussion group: http://piazza.com/gatech/fall2015/cse6740cs7641isye6740/home
- More information: http://www.cc.gatech.edu/~lsong/teaching/CSE6740fall15.html