Introduction to Machine Learning (CSCI-UA.0480-007)

David Sontag
New York University

Slides adapted from Luke Zettlemoyer, Pedro Domingos, and Carlos Guestrin
Logistics

• **Class webpage:**
  – http://cs.nyu.edu/~dsontag/courses/ml16/
  – Sign up for Piazza!

• **Office hours:** TBD

• **Teaching assistant:**
  Kevin Jiao <jjiao@stern.nyu.edu>

• **Graders:**
  – Yijun Xiao <ryjxiao@nyu.edu>
  – Alexandre Sablayrolles
    <alexandre.sablayrolles@gmail.com>
Evaluation

• 6-7 homeworks (50%)
  – Both theory and programming
  – Collaboration policy:
    • First try to solve the problems on your own
    • Then, can discuss with other classmates
    • Write-up solutions on your own
    • List names of anyone you talked to

• Midterm exam (25%)
• Project (20%)
• Course participation (5%)
Projects

• Be creative – think of new problems that you can tackle using machine learning
  – Scope: ~40 hours/person

• Logistics:
  – 2-3 students per group
  – Begins mid-March. Project proposal due week after midterm exam
  – Will still be problem sets during this period!
Prerequisites

REQUIRED:
• Basic algorithms (CS 310)
  – Dynamic programming, algorithmic analysis
  – *Can be taken concurrently*

STRONGLY RECOMMENDED:
• Linear algebra (Math 140)
  – Matrices, vectors, systems of linear equations
  – Eigenvectors, matrix rank
  – Singular value decomposition
• Multivariable calculus (Math 123)
  – Derivatives, integration, tangent planes
  – Optimization, Lagrange multipliers
• Good programming skills: Python highly recommended
Source Materials

No textbook required. Readings will come from freely available online material.

If you really want a book for an additional reference, these are OK options:
• … may update this list throughout semester. I wouldn’t buy anything yet.
What is Machine Learning?
(by examples)
Classification
from data to discrete classes
Spam filtering

data

Welcome to New Media Installation: Art that Learns

Carlos Guestrin to 10615-announce, Osman, Miche | show details 3:15 PM (8 hours ago) | Reply |
Hi everyone,

Welcome to New Media Installation: Art that Learns

The class will start tomorrow.

***Make sure you attend the first class, even if you are on the Wait List.***
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rIk | spam |

Jaquelyn Halley to nherlein, bcc: thehorney, bcc: ang | show details 9:52 PM (1 hour ago) | Reply |

=== Natural WeightLOSS Solution ===

Vital Acai is a natural WeightLOSS product that enables people to lose weight and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that you may not be aware of. These benefits have helped people who have been using Vital Acai daily to achieve goals and reach new heights in their dieting that they never thought they could.

* Rapid WeightLOSS
* Increased metabolism - Burn Fat & Calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* BetterSexLife
* A Natural Colon Cleanse

Spam vs. Not Spam
Face recognition

Example training images for each orientation
Weather prediction
Regression

predicting a numeric value
Stock market
Weather prediction revisited

Temperature

72° F
Ranking

comparing items
Web search

Learning to rank - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Learning_to_rank
Learning to rank or machine-learned ranking (MLR) is a type of supervised or semi-supervised machine learning problem in which the goal is to automatically ...

Applications Feature vectors Evaluation measures Approaches

Yahoo! Learning to Rank Challenge
learneratorankchallenge.yahoo.com/
Learning to Rank Challenge is closed! Close competition, innovative ideas, and fierce determination were some of the highlights of the first ever Yahoo!

[PDF] Large Scale Learning to Rank
www.eecs.tufts.edu/~dsculley/papers/large-scale-rank.pdf
File Format: PDF/Adobe Acrobat - Quick View
by D Sculley - Cited by 24 - Related articles
Pairwise learning to rank methods such as RankSVM give good performance, ... In this paper, we are concerned with learning to rank methods that can learn on ...

Microsoft Learning to Rank Datasets - Microsoft Research
research.microsoft.com/en-us/projects/mslr/
We release two large scale datasets for research on learning to rank: L2R-WEB30k with more than 30000 queries and a random sampling of it L2R-WEB10K ...

LETOR: A Benchmark Collection for Research on Learning to Rank ...
research.microsoft.com/~letor/
This website is designed to facilitate research in Learning TO Rank (LETOR). Much information about learning to rank can be found in the website, including ...
Given image, find similar images

http://www.tiltomo.com/
Collaborative Filtering
Recommendation systems
Recommendation systems

Machine learning competition with a $1 million prize

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>Best Score</th>
<th>% Improvement</th>
<th>Last Submit Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The Ensemble</td>
<td>0.8553</td>
<td>10.10</td>
<td>2009-07-26 18:38:22</td>
</tr>
<tr>
<td>2</td>
<td>BellKor’s Pragmatic Chaos</td>
<td>0.8554</td>
<td>10.09</td>
<td>2009-07-26 18:18:28</td>
</tr>
</tbody>
</table>

**Grand Prize - RMSE <= 0.0563**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>Best Score</th>
<th>% Improvement</th>
<th>Last Submit Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Grand Prize Team</td>
<td>0.8571</td>
<td>9.91</td>
<td>2009-07-24 13:07:49</td>
</tr>
<tr>
<td>4</td>
<td>Opera Solutions and Vandelay United</td>
<td>0.8573</td>
<td>9.89</td>
<td>2009-07-25 20:05:52</td>
</tr>
<tr>
<td>5</td>
<td>Vandelay Industries !</td>
<td>0.8579</td>
<td>9.83</td>
<td>2009-07-26 02:49:53</td>
</tr>
<tr>
<td>6</td>
<td>PragmaticTheory</td>
<td>0.8582</td>
<td>9.80</td>
<td>2009-07-12 16:09:53</td>
</tr>
<tr>
<td>7</td>
<td>BellKor in BigChaos</td>
<td>0.8590</td>
<td>9.71</td>
<td>2009-07-26 12:57:25</td>
</tr>
<tr>
<td>8</td>
<td>Dace</td>
<td>0.8603</td>
<td>9.58</td>
<td>2009-07-24 17:18:43</td>
</tr>
<tr>
<td>9</td>
<td>Opera Solutions</td>
<td>0.8611</td>
<td>9.49</td>
<td>2009-07-26 18:02:08</td>
</tr>
<tr>
<td>10</td>
<td>BellKor</td>
<td>0.8612</td>
<td>9.48</td>
<td>2009-07-26 18:19:11</td>
</tr>
<tr>
<td>11</td>
<td>BigChaos</td>
<td>0.8613</td>
<td>9.47</td>
<td>2009-06-23 23:06:52</td>
</tr>
<tr>
<td>12</td>
<td>Feeds2</td>
<td>0.8613</td>
<td>9.47</td>
<td>2009-07-24 20:06:46</td>
</tr>
</tbody>
</table>

**Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>Best Score</th>
<th>% Improvement</th>
<th>Last Submit Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>xiangliang</td>
<td>0.8633</td>
<td>9.26</td>
<td>2009-07-21 02:04:40</td>
</tr>
<tr>
<td>14</td>
<td>Gravty</td>
<td>0.8634</td>
<td>9.25</td>
<td>2009-07-26 16:58:34</td>
</tr>
<tr>
<td>15</td>
<td>Ces</td>
<td>0.8642</td>
<td>9.17</td>
<td>2009-07-25 17:42:38</td>
</tr>
<tr>
<td>16</td>
<td>Invisible Ideas</td>
<td>0.8644</td>
<td>9.14</td>
<td>2009-07-20 03:26:12</td>
</tr>
<tr>
<td>17</td>
<td>Just a guy in a garage</td>
<td>0.8650</td>
<td>9.08</td>
<td>2009-07-22 14:10:42</td>
</tr>
<tr>
<td>18</td>
<td>Craig Carmichael</td>
<td>0.8656</td>
<td>9.02</td>
<td>2009-07-25 16:00:54</td>
</tr>
<tr>
<td>19</td>
<td>J Dennis Su</td>
<td>0.8659</td>
<td>9.00</td>
<td>2009-03-11 09:41:54</td>
</tr>
<tr>
<td>20</td>
<td>acehill</td>
<td>0.8659</td>
<td>8.99</td>
<td>2009-04-06:29:35</td>
</tr>
</tbody>
</table>

**Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell**

**Cinematch score on quiz subset - RMSE = 0.9514**
Clustering
discovering structure in data
Clustering Data: Group similar things
Clustering images

Set of Images

[Goldberger et al.]
Clustering web search results

1. **Race (classification of human beings) - Wikipedia, the free encyclopedia**
   - The term race or racial group usually refers to the concept of dividing humans into populations or groups on the basis of various sets of characteristics. The most widely used human racial categories are based on visible traits (especially skin color, cranial or facial features and hair texture), and self-identification. Conceptions of race, as well as specific ways of grouping races, vary by culture and over time, and are often controversial for scientific as well as social and political reasons. History - Modern debates - Political and ...

2. **Race - Wikipedia, the free encyclopedia**
   - General. Racing competitions The Race (yachting race), or La course du millénaire, a no-rules round-the-world sailing event. Race (biology), classification of flora and fauna; Race (classification of human beings) Race and ethnicity in the United States Census, official definitions of "race" used by the US Census Bureau; Race and genetics, notion of racial classifications based on genetics; Historical definitions of race; Race (bearing), the inner and outer rings of a rolling-element bearing. RACE in molecular biology — Rad ... General - Surnames - Television - Music - Literature - Video games

3. **Publications | Human Rights Watch**
   - The use of torture, unlawful rendition, secret prisons, unfair trials, ... Risks to Migrants, Refugees, and Asylum Seekers in Egypt and Israel ... In the run-up to the Beijing Olympics in August 2008, ...
   - [www.hrw.org/background/usa/race](www.hrw.org/background/usa/race) - [cache] - Ask

4. **Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich ...**

5. **AAPA Statement on Biological Aspects of Race**
   - AAPA Statement on Biological Aspects of Race ... Published in the American Journal of Physical Anthropology, vol. 101, pp 569-570, 1996 ... PREAMBLE As scientists who study human evolution and variation, ...
   - [www.physanth.org/positions/race.html](www.physanth.org/positions/race.html) - [cache] - Ask

6. **race: Definition from Answers.com**
   - race n. A local geographic or global human population distinguished as a more or less distinct group by genetically transmitted physical ...
   - [www.answers.com/topic/race-1](www.answers.com/topic/race-1) - [cache] - Live

7. **Dopefish.com**
   - Site for newbies as well as experienced Dopefish followers, chronicling the birth of the Dopefish, its numerous appearances in several computer games, and its eventual take-over of the human race. Maintained by Mr. Dopefish himself, Joe Stiegler of Apegone Software.
   - [www.dopefish.com](www.dopefish.com) - [cache] - Open Directory
Embedding
visualizing data
Embedding images

• Images have thousands or millions of pixels.

• Can we give each image a coordinate, such that similar images are near each other?

[Saul & Roweis ’03]
Embedding words
Embedding words (zoom in)
Structured prediction
from data to discrete classes
Speech recognition

“ I need to hide a body ”

What kind of place are you looking for?

- reservoirs
- metal foundries
- mines
- dumps
- swamps
Natural language processing

I need to hide a body

noun, verb, preposition, ...

“ I need to hide a body ”

What kind of place are you looking for?

reservoirs
metal foundries
mines
dumps
swamps
Growth of Machine Learning

- Machine learning is preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology
  - Sensor networks
  - ...

- This trend is accelerating
  - Big data
  - Improved machine learning algorithms
  - Faster computers
  - Good open-source software
Course roadmap

• First half of course: supervised learning
  – SVMs, kernel methods
  – Learning theory
  – Decision trees, boosting, deep learning

• Second half of course: data science
  – Unsupervised learning, EM algorithm
  – Dimensionality reduction
  – Topic models
Supervised Learning: find $f$

- **Given:** Training set $\{(x_i, y_i) \mid i = 1 \ldots N\}$
- **Find:** A good approximation to $f : X \rightarrow Y$

**Examples:** what are $X$ and $Y$?

- **Spam Detection**
  - Map email to \{Spam, Not Spam\}
- **Digit recognition**
  - Map pixels to $\{0,1,2,3,4,5,6,7,8,9\}$
- **Stock Prediction**
  - Map new, historic prices, etc. to $\mathbb{R}$ (the real numbers)
A Supervised Learning Problem

Our goal is to find a function \( f : X \rightarrow Y \)
- \( X = \{0,1\}^4 \)
- \( Y = \{0,1\} \)

Question 1: How should we pick the hypothesis space, the set of possible functions \( f \)?

Question 2: How do we find the best \( f \) in the hypothesis space?

Dataset:

<table>
<thead>
<tr>
<th>Example</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Most General Hypothesis Space

Consider all possible boolean functions over four input features!

- $2^{16}$ possible hypotheses
- $2^9$ are consistent with our dataset
- How do we choose the best one?

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>

Dataset:

<table>
<thead>
<tr>
<th>Example</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
A Restricted Hypothesis Space

Consider all conjunctive boolean functions.

- 16 possible hypotheses
- None are consistent with our dataset
- How do we choose the best one?

<table>
<thead>
<tr>
<th>Rule</th>
<th>Counterexample</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>1</td>
</tr>
<tr>
<td>( x_1 \Rightarrow y )</td>
<td>3</td>
</tr>
<tr>
<td>( x_2 \Rightarrow y )</td>
<td>2</td>
</tr>
<tr>
<td>( x_3 \Rightarrow y )</td>
<td>1</td>
</tr>
<tr>
<td>( x_4 \Rightarrow y )</td>
<td>7</td>
</tr>
<tr>
<td>( x_1 \land x_2 \Rightarrow y )</td>
<td>3</td>
</tr>
<tr>
<td>( x_1 \land x_3 \Rightarrow y )</td>
<td>3</td>
</tr>
<tr>
<td>( x_1 \land x_4 \Rightarrow y )</td>
<td>3</td>
</tr>
<tr>
<td>( x_2 \land x_3 \Rightarrow y )</td>
<td>3</td>
</tr>
<tr>
<td>( x_2 \land x_4 \Rightarrow y )</td>
<td>3</td>
</tr>
<tr>
<td>( x_3 \land x_4 \Rightarrow y )</td>
<td>4</td>
</tr>
<tr>
<td>( x_1 \land x_2 \land x_3 \Rightarrow y )</td>
<td>3</td>
</tr>
<tr>
<td>( x_1 \land x_2 \land x_4 \Rightarrow y )</td>
<td>3</td>
</tr>
<tr>
<td>( x_1 \land x_3 \land x_4 \Rightarrow y )</td>
<td>3</td>
</tr>
<tr>
<td>( x_2 \land x_3 \land x_4 \Rightarrow y )</td>
<td>3</td>
</tr>
<tr>
<td>( x_1 \land x_2 \land x_3 \land x_4 \Rightarrow y )</td>
<td>3</td>
</tr>
</tbody>
</table>

Dataset:

<table>
<thead>
<tr>
<th>Example</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Occam’s Razor Principle

• William of Occam: Monk living in the 14th century
• Principle of parsimony:

“One should not increase, beyond what is necessary, the number of entities required to explain anything”

• When many solutions are available for a given problem, we should select the simplest one
• But what do we mean by simple?
• We will use prior knowledge of the problem to solve to define what is a simple solution

*Example of a prior: smoothness*

[Samy Bengio]
Key Issues in Machine Learning

• How do we choose a hypothesis space?
  – Often we use prior knowledge to guide this choice
• How can we gauge the accuracy of a hypothesis on unseen data?
  – Occam’s razor: use the simplest hypothesis consistent with data! This will help us avoid overfitting.
  – Learning theory will help us quantify our ability to generalize as a function of the amount of training data and the hypothesis space
• How do we find the best hypothesis?
  – This is an algorithmic question, the main topic of computer science
• How to model applications as machine learning problems? (engineering challenge)