CSE 546
Machine Learning

Instructor: Luke Zettlemoyer
TA: Lydia Chilton

Slides adapted from Pedro Domingos and Carlos Guestrin
Logistics

• **Instructor:** Luke Zettlemoyer  
  – Email: lsz@cs  
  – Office: CSE 658  
  – Office hours: Tuesdays 11-12

• **TA:** Lydia Chilton  
  – Email: hmslydia@cs  
  – Office hours: TBD

• **Web:** [www.cs.washington.edu/546](http://www.cs.washington.edu/546)
Evaluation

• 3-4 homeworks (40% total)
• Midterm (25%)
  – Actually, 2/3 term
• Final mini-project (30%)
  – Approx. one month’s work. Can incorporate your research! Or, could replicate paper, etc.
• Course participation (5%)
  – includes in class, message board, etc.
Source Materials


• Optional:
  – R. Duda, P. Hart & D. Stork, *Pattern Classification* (2nd ed.), Wiley (Required)
  – Papers
A Few Quotes

• “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Chairman, Microsoft)
• “Machine learning is the next Internet” (Tony Tether, Director, DARPA)
• Machine learning is the hot new thing” (John Hennessy, President, Stanford)
• “Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Dir. Research, Yahoo)
• “Machine learning is going to result in a real revolution” (Greg Papadopoulos, CTO, Sun)
• “Machine learning is today’s discontinuity” (Jerry Yang, CEO, Yahoo)
So What Is Machine Learning?

• Automating automation
• Getting computers to program themselves
• Writing software is the bottleneck
• Let the data do the work instead!
• The future of Computer Science!!!
Traditional Programming

Data → Computer → Output

Program → Computer → Output

Machine Learning

Data → Computer → Program

Output → Computer → Program
Magic?

No, more like gardening

- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs
What We Will Cover

• **Supervised learning**
  – Decision tree induction
  – Linear models for regression and classification
  – Instance-based learning
  – Bayesian learning
  – Neural networks
  – Support vector machines
  – Model ensembles
  – Learning theory

• **Unsupervised learning**
  – Clustering
  – Dimensionality reduction
What is Machine Learning?  
(by examples)
Classification

from data to discrete classes
Spam filtering

Welcome to New Media Installation: Art that Learns

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 Weight Loss Superfood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rik

--- Natural Weight Loss Solution ---
Vital Acai is a natural Weight Loss 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 might 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 Weight Loss
* Increased metabolism - Burn Fat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* Better Sex Life
* A Natural Colon Cleanse
Object detection

(Prof. H. Schneiderman)

Example training images for each orientation
Reading a noun (vs verb)

[Rustandi et al., 2005]
Weather prediction
Regression

predicting a numeric value
Stock market
Weather prediction revisited

Temperature
72° F
Modeling sensor data

- Measure temperatures at some locations
- Predict temperatures throughout the environment

[Guestrin et al. ’04]
Similarity

finding data
Given image, find similar images

http://www.tiltomo.com/
Collaborative Filtering
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 genomics. Historical definitions of race; Race (bearing), the inner and outer rings of a rolling-element bearing. RACE in molecular biology *Raypal ... General - Surnames - Television - Music - Literatur - 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/backgrounder/usarace [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.physsanth.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

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 Siegel of Apogee Software.
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)
Reinforcement Learning

training by feedback
Learning to act

• Reinforcement learning

• An agent
  – Makes sensor observations
  – Must select action
  – Receives rewards
    • positive for “good” states
    • negative for “bad” states

Robot Motor Skill Coordination with EM-based Reinforcement Learning

Petar Kormushev, Sylvain Calinon, and Darwin G. Caldwell

Italian Institute of Technology
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
  – Improved machine learning algorithms
  – Improved data capture, networking, faster computers
  – Software too complex to write by hand
  – New sensors / IO devices
  – Demand for self-customization to user, environment
Supervised Learning: find $f$

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

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

- **Spam Detection**
  - Map email to $\{$Spam, Ham$\}$
- **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)
Example: Spam Filter

- **Input:** email
- **Output:** spam/ham
- **Setup:**
  - Get a large collection of example emails, each labeled “spam” or “ham”
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future emails

- **Features:** The attributes used to make the ham / spam decision
  - Words: FREE!
  - Text Patterns: $dd, CAPS
  - Non-text: SenderInContacts
  - ...

---

**Dear Sir.**
First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidencial and top secret. …

**TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.**

99 MILLION EMAIL ADDRESSES FOR ONLY $99

**Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.**
Example: Digit Recognition

- **Input:** images / pixel grids
- **Output:** a digit 0-9
- **Setup:**
  - Get a large collection of example images, each labeled with a digit
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future digit images

- **Features:** The attributes used to make the digit decision
  - Pixels: (6,8)=ON
  - Shape Patterns: NumComponents,AspectRatio, NumLoops
  - ...
Important Concepts

- **Data**: labeled instances, e.g. emails marked spam/ham
  - Training set
  - Held out set (sometimes call Validation set)
  - Test set

- **Features**: attribute-value pairs which characterize each $x$

- **Experimentation cycle**
  - Select a hypothesis $f$ to best match training set
  - (Tune hyperparameters on held-out set)
  - Compute accuracy of test set
  - Very important: never “peek” at the test set!

- **Evaluation**
  - Accuracy: fraction of instances predicted correctly

- **Overfitting and generalization**
  - Want a classifier which does well on test data
  - Overfitting: fitting the training data very closely, but not generalizing well
  - We’ll investigate overfitting and generalization formally in a few lectures
A Supervised Learning Problem

• Consider a simple, Boolean dataset:
  – \( f : X \rightarrow Y \)
  – \( X = \{0,1\}^4 \)
  – \( Y = \{0,1\} \)

• **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>
Another Sup. Learning Problem

- Consider a simple, regression dataset:
  - \( f : X \rightarrow Y \)
  - \( X = \mathbb{R} \)
  - \( Y = \mathbb{R} \)

- **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: 10 points generated from a sin function, with noise
Hypo. Space: Degree-N Polynomials

- Infinitely many hypotheses
- None / Infinitely many are consistent with our dataset
- How do we choose the best one?
Key Issues in Machine Learning

• What are good hypothesis spaces?
• How to find the best hypothesis? (algorithms / complexity)
• How to optimize for accuracy of unseen testing data? (avoid overfitting, etc.)
• Can we have confidence in results? How much data is needed?
• How to model applications as machine learning problems? (engineering challenge)