What’s learning?
Point Estimation

Machine Learning – CSE446
Carlos Guestrin
University of Washington
April 1, 2013

What is Machine Learning?
Machine Learning

Study of algorithms that
- improve their performance
- at some task
- with experience

Classification

from data to discrete classes
Spam filtering

Welcome to New Media Installation: Art that Learns

- Carlos Guestrin to E15678root: Teamwork: 9:30 PM (8 hours ago)  
- Welcome to New Media Installation Art that Learns
- This space is for hivemind.
- "Make sure you delete the last email, even if you are the worst Ltd."**
- This team will be held in English only.
- The team will be held 8 to 9 PM EST.
- If you need any assistance, please email hivemindelp@cs.washington.edu.
- You can contact the instructors by emailing: hivemindelp@cs.washington.edu

Natural _LowWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $0.95 shipping & tax

- VitalAid is a natural weight loss product that can help people lose weight and cleanse their bodies faster than most other products on the market.
- Here are some of the benefits of VitalAid: You might not be aware of. These benefits have been shown in people using VitalAid daily to achieve goals and reach new heights in three months of:
- They want to lose weight.
- Fast weight loss.
- Improved digestion - Gut Feel & digestion easily!
- Increased energy - More energy per day!
- More self-confidence - More confidence!
- Detoxify Your Body - Feel better, look better, and feel great again!
- Much More Energy - Feel better, look better, and feel great again!
- A Natural Weight Loss Program

Text classification

Company home page
vs
Personal home page
vs
University home page
vs
...
Object detection

(Prof. H. Schneiderman)

Example training images for each orientation

Reading a noun (vs verb)

[Rustandi et al., 2005]
Weather prediction

The classification pipeline

Training

Testing
Regression

predicting a numeric value

Stock market
Weather prediction revisited

Modeling sensor data

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

[Guestrin et al. '04]
Given image, find similar images
Clustering

discovering structure in data
Clustering Data: Group similar things

Clustering images

Set of Images

[Goldberger et al.]
Clustering web search results

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

[Joseph Turian]
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

[Ng et al. '05]
HURLEY: Uh ... the Chinese people have water. (Sayid and Kate go to check it out.)

[SAYID]

[SUN]

[EXT. BEACH - CRASH SITE]

(Sayid holds the empty bottle in his hand and questions Sun.)

SAYID: (quietly) Where did you get this? (He looks at her.)

[EXT. JUNGLE]

(Sawyer is walking through the jungle. He reaches a spot. He kneels down and looks back to check that no one’s followed him.)

<table>
<thead>
<tr>
<th>Automatically Discovered and Labeled Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>sit down</td>
</tr>
<tr>
<td>smile</td>
</tr>
<tr>
<td>wake</td>
</tr>
<tr>
<td>swim</td>
</tr>
<tr>
<td>follow</td>
</tr>
<tr>
<td>grab</td>
</tr>
<tr>
<td>kiss</td>
</tr>
<tr>
<td>open door</td>
</tr>
<tr>
<td>point</td>
</tr>
</tbody>
</table>
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, especially with **Big Data**
  - 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

One of the most sought for specialties in industry today!!!!

Syllabus

- Covers a wide range of Machine Learning techniques – from basic to state-of-the-art

- You will learn about the methods you heard about:
  - Point estimation, regression, naive Bayes, logistic regression, nearest-neighbor, decision trees, boosting, perceptron, overfitting, regularization, dimensionality reduction, PCA, error bounds, VC dimension, SVMs, kernels, margin bounds, K-means, EM, mixture models, semi-supervised learning, HMMs, graphical models, active learning, reinforcement learning...

- Covers algorithms, theory and applications

- It’s going to be fun and hard work 😊
Prerequisites

- Formally:
  - Either CSE 326 or CSE 332; either STAT 390, STAT 391, or CSE 312
- Probabilities
  - Distributions, densities, marginalization…
- Basic statistics
  - Moments, typical distributions, regression…
- Algorithms
  - Dynamic programming, basic data structures, complexity…
- Programming
  - R will be very useful, but we’ll help you get started
  - We provide some background, but the class will be fast paced
- Ability to deal with “abstract mathematical concepts”

Optional R tutorial

- There are many resources to get started with R online
- We’ll run an optional tutorial:
  - Thursday 4/4 @6pm
  - Location TBD
Staff

- Three Great TAs: Great resource for learning, interact with them!
  - Danielle Bragg
    Office hours: Wednesdays 3:30-5:30pm
  - Daryl Hansen
    Office hours: Thursdays 1:30-3:30pm
  - James McQueen
    Office hours: Tuesdays 9:30-11:30am

- Prof: Carlos Guestrin
  Office hours: Fridays 1:30-2:30pm

Communication Channels

- Only channel for announcements, questions, etc. – Google Group:
  - https://groups.google.com/forum/?fromgroups#!forum/cse446-spr13
  - Subscribe!
  - All non-personal questions should go here
  - Answering your question will help others
  - Feel free to chime in

- For e-mailing instructors about personal issues, use:
  - cse446-staff@cs.washington.edu
Text Books

- Required Textbook:
  - Machine Learning: a Probabilistic Perspective; Kevin Murphy

- Optional Books:
  - Pattern Recognition and Machine Learning; Chris Bishop
  - The Elements of Statistical Learning: Data Mining, Inference, and Prediction; Trevor Hastie, Robert Tibshirani, Jerome Friedman
  - Machine Learning; Tom Mitchell
  - Information Theory, Inference, and Learning Algorithms; David MacKay

Grading

- 4 homeworks (40%)
  - First one goes out 4/4
    - Start early, Start early, Start early, Start early, Start early, Start early, Start early, Start early, Start early, Start early

- Final project (20%)
  - Details out around April 29th
  - Projects done individually, or groups of two students

- Midterm (15%)
  - Wed., May 8th in class

- Final (25%)
  - TBD by registrar, probably 6/12/2013, 8:30-10:20am
Homeworks

- Homeworks are hard, start early 😊
- Due in the beginning of class
- 33% subtracted per late day
- All homeworks **must be handed in**, even for zero credit
- Use Catalyst to submit homeworks

Collaboration
- You may **discuss** the questions
- Each student writes their own answers
- Write on your homework anyone with whom you collaborate
- Each student must write their own code for the programming part
- Please **don't search for answers on the web, Google, previous years' homeworks, etc.**
  - please ask us if you are not sure if you can use a particular reference

Enjoy!

- ML is becoming ubiquitous in science, engineering and beyond
- It’s one of the hottest topics in industry today
- This class should give you the basic foundation for applying ML and developing new methods
- The fun begins…
Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have thumbtack, if I flip it, what’s the probability it will fall with the nail up?
  - You say: Please flip it a few times:

- You say: The probability is:
- **He says: Why???
- You say: Because…

Thumbtack – Binomial Distribution

- \( P(\text{Heads}) = \theta, \ P(\text{Tails}) = 1-\theta \)

- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution

- Sequence \( D \) of \( \alpha_H \) Heads and \( \alpha_T \) Tails

\[
P(D \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}
\]
Maximum Likelihood Estimation

- **Data**: Observed set $D$ of $\alpha_H$ Heads and $\alpha_T$ Tails
- **Hypothesis**: Binomial distribution
- Learning $\theta$ is an optimization problem
  - What’s the objective function?

- MLE: Choose $\theta$ that maximizes the probability of observed data:

$$
\hat{\theta} = \arg \max_\theta P(D \mid \theta) = \arg \max_\theta \ln P(D \mid \theta)
$$

Your first learning algorithm

$$
\hat{\theta} = \arg \max_\theta \ln P(D \mid \theta) = \arg \max_\theta \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}
$$

- Set derivative to zero:

$$
\frac{d}{d\theta} \ln P(D \mid \theta) = 0
$$
How many flips do I need?

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: \( \theta = 3/5 \), I can prove it!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it!

**He says: What’s better?**

- You say: Humm… The more the merrier???
- He says: Is this why I am paying you the big bucks???

Simple bound  
(based on Hoeffding’s inequality)

- For \( N = \alpha_H + \alpha_T \), and \( \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \)

- Let \( \theta^* \) be the true parameter, for any \( \epsilon > 0 \):
  \[ P(\left| \hat{\theta} - \theta^* \right| \geq \epsilon) \leq 2e^{-2N\epsilon^2} \]
PAC Learning

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the thumbtack parameter $\theta$, within $\epsilon = 0.1$, with probability at least $1-\delta = 0.95$. How many flips?

$$P\left(\left|\hat{\theta} - \theta^*\right| \geq \epsilon\right) \leq 2e^{-2N\epsilon^2}$$

What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- You say: Let me tell you about Gaussians...

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
Some properties of Gaussians

- **Affine transformation** (multiplying by scalar and adding a constant)
  - \( X \sim N(\mu, \sigma^2) \)
  - \( Y = aX + b \quad \Rightarrow \quad Y \sim N(a\mu + b, a^2 \sigma^2) \)

- **Sum of Gaussians**
  - \( X \sim N(\mu_X, \sigma^2_X) \)
  - \( Y \sim N(\mu_Y, \sigma^2_Y) \)
  - \( Z = X + Y \quad \Rightarrow \quad Z \sim N(\mu_X + \mu_Y, \sigma^2_X + \sigma^2_Y) \)

Learning a Gaussian

- **Collect a bunch of data**
  - Hopefully, i.i.d. samples
  - e.g., exam scores

- **Learn parameters**
  - Mean
  - Variance

\[
P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]
MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_N\}$:

$$P(D \mid \mu, \sigma) = \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^N \prod_{i=1}^N e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- Log-likelihood of data:

$$\ln P(D \mid \mu, \sigma) = \ln \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^N \prod_{i=1}^N e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2}$$

Your second learning algorithm: MLE for mean of a Gaussian

- What’s MLE for mean?

$$\frac{d}{d\mu} \ln P(D \mid \mu, \sigma) = \frac{d}{d\mu} \left[-N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$
MLE for variance

Again, set derivative to zero:

\[
\frac{d}{d\sigma} \ln P(D | \mu, \sigma) = \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]

\[
= \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^{N} \frac{d}{d\sigma} \left[ \frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]

Learning Gaussian parameters

**MLE:**

\[
\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i
\]

\[
\hat{\sigma}^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]

**BTW.** MLE for the variance of a Gaussian is **biased**

- Expected result of estimation is **not** true parameter!
- Unbiased variance estimator:

\[
\hat{\sigma}^2_{unbiased} = \frac{1}{N - 1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]
What you need to know…

- Learning is…
  - Collect some data
    - E.g., thumbtack flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood
  - Choose an optimization procedure
    - E.g., set derivative to zero to obtain MLE
  - Collect the big bucks

- Like everything in life, there is a lot more to learn…
  - Many more facets… Many more nuances…
  - The fun will continue…