

<http://www.cs.cmu.edu/~guestrin/Class/10701/>

What's learning? Point Estimation

Machine Learning – 10701/15781

Carlos Guestrin

Carnegie Mellon University

September 10th, 2007

©2005-2007 Carlos Guestrin

1

What is Machine Learning ?

©2005-2007 Carlos Guestrin

2

Machine Learning

Study of algorithms that

- improve their performance
- at some task
- with experience

©2005-2007 Carlos Guestrin

3

Object detection

(Prof. H. Schneiderman)



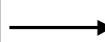
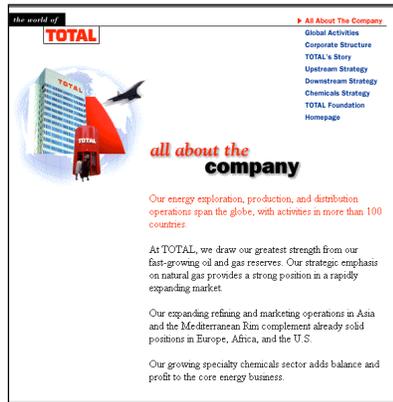
Example training images
for each orientation



©2005-2007 Carlos Guestrin

4

Text classification



Company home page

vs

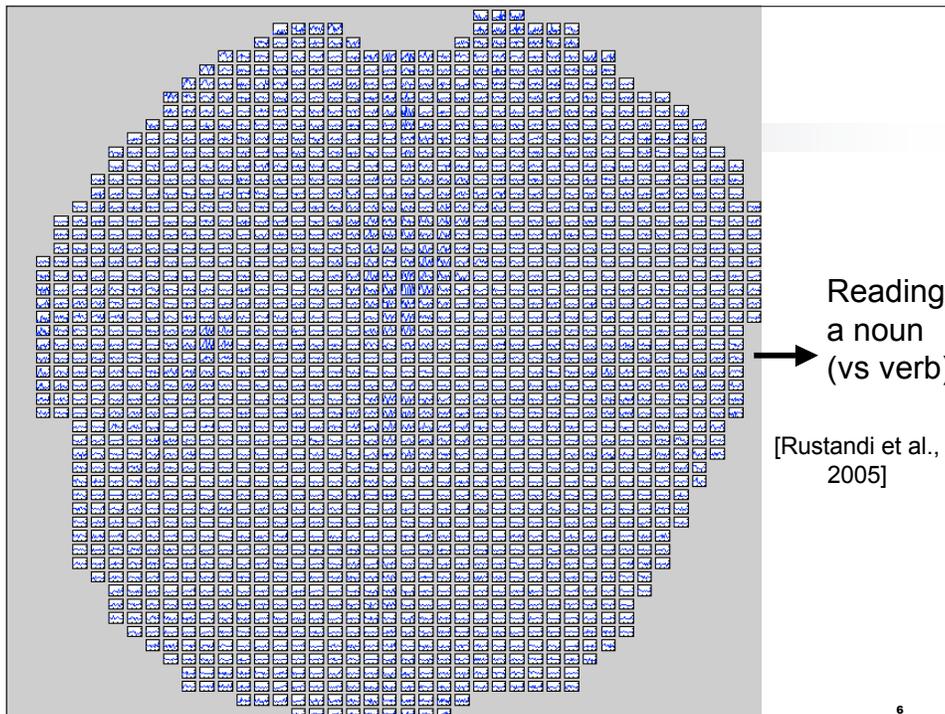
Personal home page

vs

University home page

vs

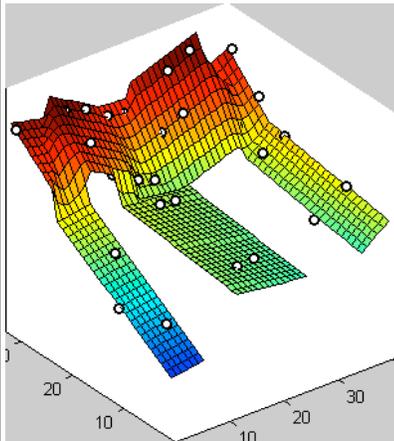
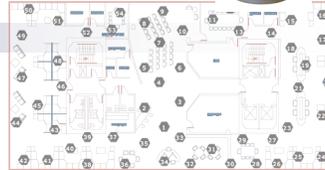
...



Reading
a noun
(vs verb)

[Rustandi et al.,
2005]

Modeling sensor data



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

[Guestrin et al. '04]

©2005-2007 Carlos Guestrin

7

Learning to act

QuickTime™ and a decompressor are needed to see this picture.

[Ng et al. '05]

- Reinforcement learning
- An agent
 - Makes sensor observations
 - Must select action
 - Receives rewards
 - positive for “good” states
 - negative for “bad” states

©2005-2007 Carlos Guestrin

8

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

©2005-2007 Carlos Guestrin

9

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:
 - Naïve Bayes, logistic regression, nearest-neighbor, decision trees, boosting, neural nets, 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 😊**

©2005-2007 Carlos Guestrin

10

Prerequisites

- Probabilities
 - Distributions, densities, marginalization...
- Basic statistics
 - Moments, typical distributions, regression...
- Algorithms
 - Dynamic programming, basic data structures, complexity...
- Programming
 - Mostly your choice of language, but Matlab will be very useful
- We provide some background, but the class will be fast paced

- Ability to deal with “abstract mathematical concepts”

Recitations

- Very useful!
 - Review material
 - Present background
 - Answer questions
- Thursdays, 5:00-6:20 in Wean Hall 5409
- Special recitation 1:
 - **tomorrow, Wean 5409, 5:00-6:20**
 - Review of probabilities
- Special recitation 2 on Matlab
 - Tuesday, Sept. 18th 4:30-5:50pm NSH 3002

Staff

- Four Great TAs: Great resource for learning, interact with them!
 - **Joseph Gonzalez**, Wean 5117, x8-3046, jgonzal@cs, Office hours: Tuesdays 7-9pm
 - **Steve Hanneke**, Doherty 4301H, x8-7375, shanneke@cs, Office hours: Fridays 1-3pm
 - **Jingrui He**, Wean 8102, x8-1299, jingruih@cs, Office hours: Wednesdays 11-1pm
 - **Sue Ann Hong**, Wean 4112, x8-3047, sahong@cs, Office hours: Tuesdays 3-5pm

- Administrative Assistant
 - Monica Hopes, x8-5527, meh@cs

©2005-2007 Carlos Guestrin

13

First Point of Contact for HWs

- To facilitate interaction, a TA will be assigned to each homework question – This will be your “first point of contact” for this question
 - But, you can always ask any of us

- For e-mailing instructors, always use:
 - 10701-instructors@cs.cmu.edu

- For announcements, subscribe to:
 - 10701-announce@cs
 - <https://mailman.srv.cs.cmu.edu/mailman/listinfo/10701-announce>

©2005-2007 Carlos Guestrin

14

Text Books

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

Grading

- **5 homeworks (35%)**
 - First one goes out 9/12
 - Start early, Start early
- **Final project (25%)**
 - Details out around Oct. 1st
 - Projects done individually, or groups of two students
- **Midterm (15%)**
 - Thu., Oct 25 5-6:30pm
 - location: MM A14
- **Final (25%)**
 - TBD by registrar

Homeworks

- Homeworks are hard, start early ☺
- Due in the beginning of class
- 3 late days for the semester
- After late days are used up:
 - Half credit within 48 hours
 - Zero credit after 48 hours
- All homeworks **must be handed in**, even for zero credit
- Late homeworks handed in to Monica Hopes, WEH 4619

- 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

Sitting in & Auditing the Class

- Due to new departmental rules, every student who wants to sit in the class (not take it for credit), must register officially for auditing
- To satisfy the auditing requirement, you must either:
 - Do ***two*** homeworks, and get at least 75% of the points in each;
or
 - Take the final, and get at least 50% of the points; or
 - Do a class project and do ***one*** homework, and get at least 75% of the points in the homework;
 - Only need to submit project proposal and present poster, and get at least 80% points in the poster.
- Please, send us an email saying that you will be auditing the class and what you plan to do.
- If you are not a student and want to sit in the class, please get authorization from the instructor

Enjoy!

- ML is becoming ubiquitous in science, engineering and beyond
- 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 α_H Heads and α_T Tails

$$P(D | \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

©2005-2007 Carlos Guestrin

21

Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
 - What's the objective function?

- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(D | \theta) \\ &= \arg \max_{\theta} \ln P(D | \theta)\end{aligned}$$

©2005-2007 Carlos Guestrin

22

Your first learning algorithm

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} \ln P(\mathcal{D} | \theta) \\ &= \arg \max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}\end{aligned}$$

- Set derivative to zero: $\frac{d}{d\theta} \ln P(\mathcal{D} | \theta) = 0$

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 θ^* be the true parameter, for any $\epsilon > 0$:

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

PAC Learning

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

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

What about prior

- Billionaire says: Wait, I know that the thumbtack is “close” to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way...**
- Rather than estimating a single θ , we obtain a distribution over possible values of θ

Bayesian Learning

- Use Bayes rule:

$$P(\theta | \mathcal{D}) = \frac{P(\mathcal{D} | \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta)$$

Bayesian Learning for Thumbtack

$$P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta)$$

- Likelihood function is simply Binomial:

$$P(\mathcal{D} | \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

- What about prior?

- Represent expert knowledge
- Simple posterior form

- Conjugate priors:

- Closed-form representation of posterior
- For Binomial, conjugate prior is Beta distribution**

©2005-2007 Carlos Guestrin

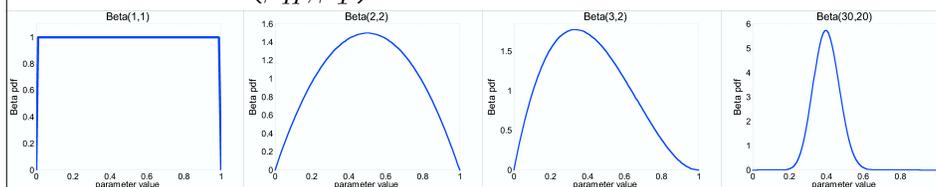
29

Beta prior distribution – $P(\theta)$

$$P(\theta) = \frac{\theta^{\beta_H - 1} (1 - \theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

Mean:

Mode:



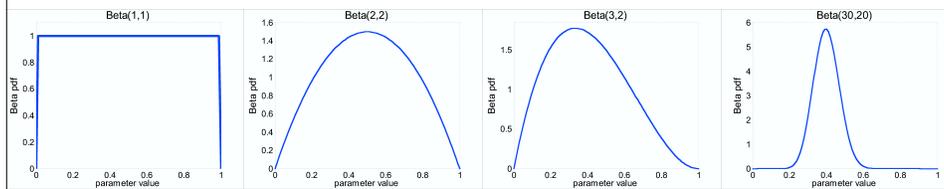
- Likelihood function: $P(\mathcal{D} | \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$
- Posterior: $P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta)$

©2005-2007 Carlos Guestrin

30

Posterior distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails
- Posterior distribution:
$$P(\theta | \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

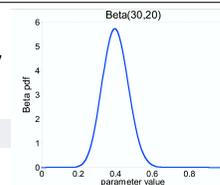


©2005-2007 Carlos Guestrin

31

Using Bayesian posterior

- Posterior distribution:
$$P(\theta | \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



- Bayesian inference:
 - No longer single parameter:
$$E[f(\theta)] = \int_0^1 f(\theta)P(\theta | \mathcal{D})d\theta$$
 - Integral is often hard to compute

©2005-2007 Carlos Guestrin

32

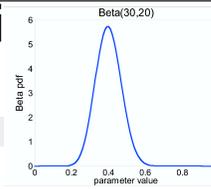
MAP: Maximum a posteriori approximation

$$P(\theta | \mathcal{D}) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta | \mathcal{D}) d\theta$$

- As more data is observed, Beta is more certain
- MAP: use most likely parameter:

$$\hat{\theta} = \arg \max_{\theta} P(\theta | \mathcal{D}) \quad E[f(\theta)] \approx f(\hat{\theta})$$



©2005-2007 Carlos Guestrin

33

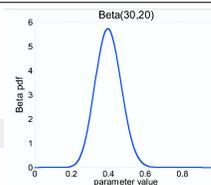
MAP for Beta distribution

$$P(\theta | \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1} (1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

- MAP: use most likely parameter:

$$\hat{\theta} = \arg \max_{\theta} P(\theta | \mathcal{D}) =$$

- Beta prior equivalent to extra thumbtack flips
- As $N \rightarrow \infty$, prior is “forgotten”
- **But, for small sample size, prior is important!**



©2005-2007 Carlos Guestrin

34

What you need to know

- Go to the recitation on intro to probabilities
 - And, other recitations too
- Point estimation:
 - MLE
 - Bayesian learning
 - MAP

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...**

$$P(x | \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Some properties of Gaussians

- affine transformation (multiplying by scalar and adding a constant)
 - $X \sim N(\mu, \sigma^2)$
 - $Y = aX + b \rightarrow Y \sim N(a\mu + b, a^2\sigma^2)$
- Sum of Gaussians
 - $X \sim N(\mu_X, \sigma_X^2)$
 - $Y \sim N(\mu_Y, \sigma_Y^2)$
 - $Z = X + Y \rightarrow Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

©2005-2007 Carlos Guestrin

37

Learning a Gaussian

- Collect a bunch of data
 - Hopefully, i.i.d. samples
 - e.g., exam scores
- Learn parameters
 - Mean
 - Variance

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

©2005-2007 Carlos Guestrin

38

MLE for Gaussian

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

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

- Log-likelihood of data:

$$\begin{aligned} \ln P(\mathcal{D} \mid \mu, \sigma) &= \ln \left[\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^N \prod_{i=1}^N e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \right] \\ &= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2} \end{aligned}$$

©2005-2007 Carlos Guestrin

39

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

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

©2005-2007 Carlos Guestrin

40

MLE for variance

- Again, set derivative to zero:

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

©2005-2007 Carlos Guestrin

41

Learning Gaussian parameters

- MLE:

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

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

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

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

$$\hat{\sigma}_{unbiased}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

©2005-2007 Carlos Guestrin

42

Bayesian learning of Gaussian parameters

- Conjugate priors
 - Mean: Gaussian prior
 - Variance: Wishart Distribution

- Prior for mean:

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{-\frac{(\mu-\eta)^2}{2\lambda^2}}$$

MAP for mean of Gaussian

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

$$\frac{d}{d\mu} [\ln P(\mathcal{D} \mid \mu) P(\mu)] = \frac{d}{d\mu} [\ln P(\mathcal{D} \mid \mu) + \ln P(\mu)]$$