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

What's learning? Point Estimation

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

January 17th, 2007

What is Machine Learning ?

Machine Learning

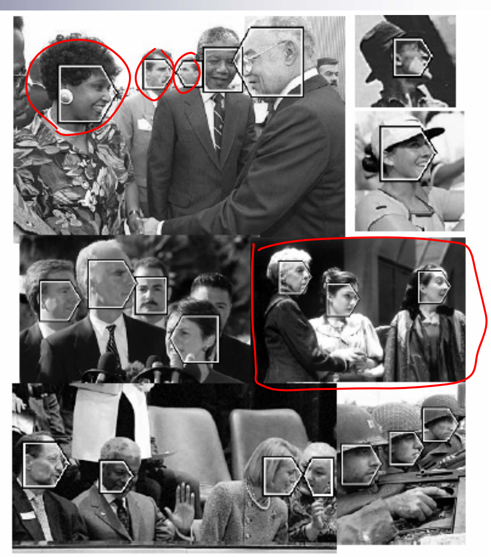
Study of algorithms that

- improve their performance
- at some task
- with experience

Object detection

(Prof. H. Schneiderman)

Example training images for each orientation



Text classification

Company home page

vs

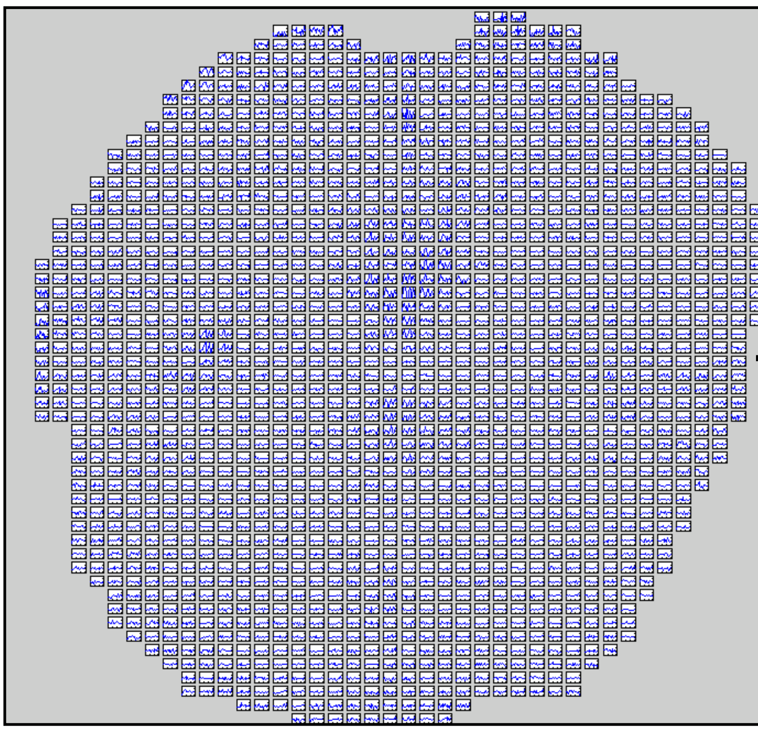
Personal home page

vs

Univeristy home page

vs

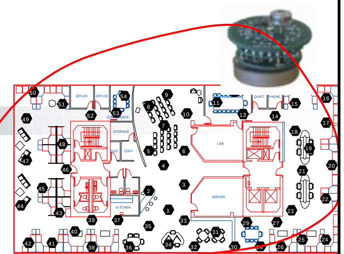
…

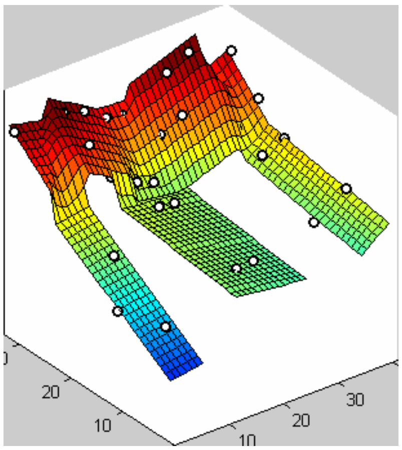


Reading a noun (vs verb)

[Rustandi et al., 2005]

Modeling sensor data





[Guestrin et al. '04]

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

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]

Growth of Machine Learning

- Machine learning is preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - …
- 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

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 ☺**

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”

Review Sessions

- Very useful!
  - Review material
  - Present background
  - Answer questions
- Thursdays, 5:30-6:50 in Wean Hall 5409
- First recitation is **tomorrow**
  - Review of probabilities
- Special recitation on Matlab
  - Jan. 24 Wed. 5:30-6:50pm NSH 1305

Staff

- Four Great TAs: Great resource for learning, interact with them!
  - Andy Carlson, acarlson@cs
  - Jonathan Huang, jch1@cs
  - Purna Sarkar, psarkar@cs
  - Brian Ziebart, bziebart@cs

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

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

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 (30%)
  - First one goes out 1/24
    - 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 Feb 26th
- Midterm (20%)
  - March 7th in class
- Final (30%)
  - May 15th, 1-4 p.m.

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
  - Don't look for answers on the web or from ~~last~~ previous semesters class, etc...

Sitting in & Auditing the Class

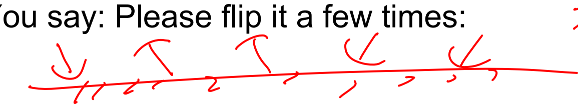
- Due to new departmental rules, every student who wants to sit in the class (not take it for credit), must register officially for auditing
- To satisfy the auditing requirement, you must either:
  - Do *two* homeworks, and get at least 75% of the points in each; or
  - Take the final, and get at least 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: 3/5



  - You say: The probability is: 60%
  - **He says: Why???**
  - You say: Because…

Thumbtack – Binomial Distribution

$\boxed{\text{Model}}$

- P(Heads) = θ, P(Tails) = $1-\theta$

H H T T H $\theta = \frac{3}{5}$

$P(HHTTH) = \theta\theta(1-\theta)(1-\theta)\theta = \theta^3(1-\theta)^2$

- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence *D* of α_H Heads and α_T Tails $\boxed{\text{Data}}$

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

Maximum Likelihood Estimation

- **Data:** Observed set *D* of $\overset{3}{\alpha_H}$ Heads and $\overset{2}{\alpha_T}$ Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
  - What's the objective function? $P(HHTTH|\theta)$

    $P(D|\theta)$
- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\widehat{\theta} &= \arg\max_\theta \; P(\mathcal{D} \mid \theta) \\ &= \arg\max_\theta \; \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

Your first learning algorithm

$\ln a \cdot b = \ln a + \ln b$

$\ln a^b = b \ln a$

$\frac{d}{dx} \ln x = \frac{1}{x}$

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

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

$\frac{d \ln}{d\theta}\left(\theta^{\alpha_H}(1-\theta)^{\alpha_T}\right)$

$= \frac{d}{d\theta}\left[\alpha_H \ln\theta + \alpha_T \ln(1-\theta)\right] = \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0$

$\theta = \frac{\alpha_H}{\alpha_H + \alpha_T}$

if $\alpha_H = 3$ $\alpha_T = 2$

$\theta = \frac{3}{3+2} = \frac{3}{5}$

How many flips do I need?

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: θ = 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)

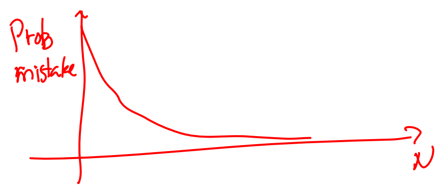
- For $N = \alpha_H + \alpha_T$, and $\hat{\theta}_{MLE} = \dfrac{\alpha_H}{\alpha_H + \alpha_T}$
- Let θ^* be the true parameter, for any $\varepsilon > 0$:

e.g., $\varepsilon = 0.01$

difference

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

$N = \alpha_H + \alpha_T$
$= \#\text{flips}$



PAC Learning

$-\ln\delta = \ln\frac{1}{\delta}$

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

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

$\delta:\ 0.05 \geq 2e^{-2N\varepsilon^2}$

$\ln\delta \geq \ln 2e^{-2N\varepsilon^2} = \ln 2 - 2N\varepsilon^2$

$2N\varepsilon^2 \geq \ln 2 - \ln\delta$

$\#\text{flips} = N \geq \dfrac{\ln 2 + \ln\frac{1}{\delta}}{2\varepsilon^2}$

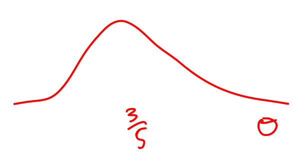
$\delta = 0.05$
$\varepsilon = 0.1$

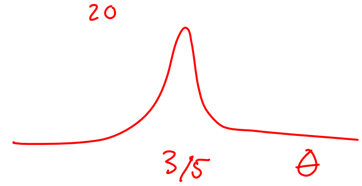
What about prior

- 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

α_H 3
α_T 2

30
20





Bayesian Learning

- Use Bayes rule:

posterior

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

- Or equivalently:

likelihood data   prior

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

$\theta^{\alpha_H} \cdot (1-\theta)^{\alpha_T}$

particular data

Bayesian Learning for Thumbtack

$$\overset{\text{posterior}}{P(\theta \mid \mathcal{D})} \;\propto\; \overset{\text{likelihood}}{P(\mathcal{D} \mid \theta)}\,\overset{\text{prior}}{P(\theta)}$$

beta distribution

- Likelihood function is simply Binomial:

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

- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors:
  - Closed-form representation of posterior
  - **For Binomial, conjugate prior is Beta distribution**

Beta prior distribution – P(θ)

$\theta^a \cdot \theta^b = \theta^{a+b}$

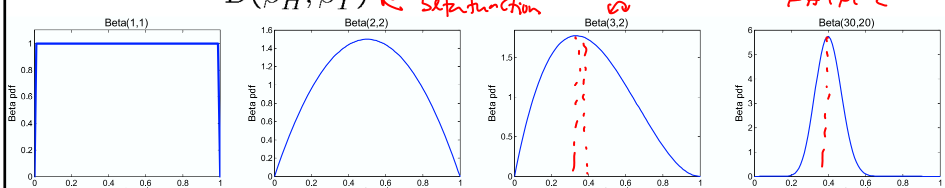
prior

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

beta function

Mean: $\frac{\beta_H}{\beta_H+\beta_T}$

Mode: $\frac{\beta_H-1}{\beta_H+\beta_T-2}$



almost like uniform

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

$$P(\theta|D) \propto \theta^{\alpha_H}(1-\theta)^{\alpha_T}\,\frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H,\beta_T)} \propto \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}$$

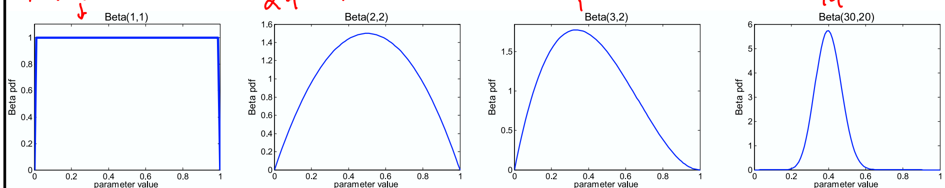
$$\sim Beta(\alpha_H+\beta_H, \alpha_T+\beta_T)$$

Posterior distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H heads and α_T tails (binomial)
- Posterior distribution:

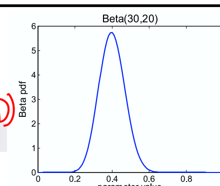
$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$P(\theta) = Beta(1,1)$ α_H 1 α_T 1 2 1 29 19



~~prior~~ $P(\theta \mid D)$ $P(\theta \mid D)$

Using Bayesian posterior



$P(\theta \mid D)$

- Posterior distribution:

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

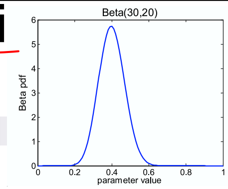
- Bayesian inference:
  - No longer single parameter:

gambling profit

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

  - Integral is often hard to compute

MAP: Maximum a posteriori approximation



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

MLE: $\arg\max_{\theta} P(\mathcal{D}|\theta)$

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

- As more data is observed, Beta is more certain

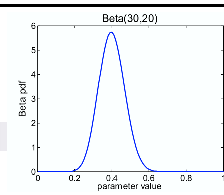
- MAP: use most likely parameter:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) \qquad E[f(\theta)] \approx f(\hat{\theta})$$

$$= \frac{\alpha_H + \beta_H - 1}{\alpha_H + \alpha_T + \beta_H + \beta_T - 2}$$

like MLE,
but "observed"
$(\beta_H - 1, \beta_T - 1)$
extra flips

MAP for Beta distribution



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

- MAP: use most likely parameter:

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \alpha_T + \beta_H + \beta_T - 2}$$

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

What you need to know

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