What’s learning?
Point Estimation

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University
January 18th, 2005
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, 5:00-6:30 in Wean Hall 5409
- First recitation is tomorrow
  - Review of probabilities
- Special recitation on Matlab
  - Jan. 25 Wed. 5:00-7:00pm, NSH 3305
Staff

- Four Great TAs: Great resource for learning, interact with them!
  - Anton Chechetka, antonc@cs
  - Stanislav Funiak, sfuniak@cs
  - Andreas Krause, krausea@cs
  - Jure Leskovec, jure@cs

- Course General Czar
  - Terrill L. Frantz, TerrillFrantz@cmu

- 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
All Text Books are Optional, but very useful

- *Machine Learning*, Tom Mitchell
- *Pattern Classification (2nd Edition)*, Duda, Hart and Stork
- *Neural Networks for Pattern Recognition*, Chris Bishop
Grading

- 5 homeworks (30%)
  - First one goes out 1/23

- Final project (20%)
  - Details out March 1st

- Midterm (20%)
  - March 8th

- Final (30%)
  - 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 4616

Collaboration
- You may discuss the questions
- Each student writes their own answers
- Write on your homework anyone with whom you collaborate
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…
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

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

[Guestrin et al. '04]
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]
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

\[ \hat{\theta} = \arg \max_{\theta} \ln P(\mathcal{D} | \theta) \]
\[ = \arg \max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T} \]

- Set derivative to zero:
  \[ \frac{d}{d\theta} \ln P(\mathcal{D} | \theta) = 0 \]
How many flips do I need?

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

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: \( \theta = \frac{3}{5} \), 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

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

Let $\theta^*$ be the true parameter, for any $\varepsilon > 0$:

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

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

$$P(|\hat{\theta} - \theta^*| \geq \varepsilon) \leq 2e^{-2N\varepsilon^2}$$
What about prior

- Billionaire says: Wait, I know that the thumbtack is “close” to 50-50. What can you?
- **You say:** I can learn it the Bayesian way...

- Rather than estimating a single $\theta$, we obtain a distribution over possible values of $\theta$
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

\[ P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta) \]

- 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 – $P(\theta)$

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

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

- Prior: $Beta(\beta_H, \beta_T)$
- Data: $\alpha_H$ heads and $\alpha_T$ tails
- Posterior distribution:

$$P(\theta \mid D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$
Using Bayesian posterior

- Posterior distribution:
  \[ P(\theta \mid \mathcal{D}) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

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

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

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

- As more data is observed, Beta is more certain

- MAP: use most likely parameter:

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

- 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
  - Bayesian learning
  - MAP