What’s learning? Point Estimation

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University
September 10th, 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

- 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
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

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:
  - Naive 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”

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, jegonzal@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

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 (35%)**
  - First one goes out 9/12
  - Start early, Start early, Start early, Start early, Start early, Start early, Start early, Start early, Start early, Start early

- **Final project (25%)**
  - Details out around Oct. 1\textsuperscript{st}
  - 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:
    - ![Thumbtack flips](image)
  - You say: The probability is: \( \frac{3}{5} \)
- He says: Why???
- You say: Because…
Thumbtack – Binomial Distribution

- P(Heads) = $\theta$, P(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 $\hat{\theta}$ that maximizes the probability of observed data:
  $$\hat{\theta} = \arg \max_{\theta} \ P(D \mid \theta)$$
  $$\hat{\theta} = \arg \max_{\theta} \ \ln P(D \mid \theta)$$
Your first learning algorithm

$$\hat{\theta}_{MLE} = \arg \max_\theta \ln P(D | \theta)$$

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

Set derivative to zero:

$$\frac{d}{d\theta} \ln P(D | \theta) = 0$$

- 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 \( \varepsilon > 0 \):
  \[
P( | \hat{\theta}_{MLE} - \theta^* | \geq \varepsilon ) \leq 2e^{-2N\varepsilon^2}
  \]

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?
  \[
P( | \hat{\theta} - \theta^* | \geq \varepsilon ) \leq 2e^{-2N\varepsilon^2}
  \]

\[
\ln \left( \frac{2e}{\varepsilon} \right) \geq -2N\varepsilon^2 \\
\Rightarrow \quad N \geq \frac{\ln \left( \frac{2e}{\varepsilon} \right)}{-2\varepsilon^2}
\]
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 $\theta$, 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

Likelihood function is simply Binomial:

\[ P(\mathcal{D} \mid \theta) = \theta^H (1 - \theta)^T \]

**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)**

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

- Likelihood function:
- Posterior:

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

\[ \theta^{\alpha_H + (1 - \theta)^{\alpha_T + \beta_T - 1}} \sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T) \]

- Mean: \[ E(\theta) = \frac{\beta_H}{\alpha_H + \beta_H} \]
- Mode: \[ Mode(\theta) = \frac{\beta_H - 1}{\alpha_H + \beta_H - 2} \]
Posterior distribution

- Prior: $\text{Beta}(\beta_H, \beta_T)$
- Data: $\alpha_H$ heads and $\alpha_T$ tails
- Posterior distribution:
  \[ P(\theta \mid D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

Using Bayesian posterior

- Posterior distribution:
  \[ P(\theta \mid D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]
- Bayesian inference:
  - No longer single parameter:
    \[ E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid D) d\theta \]
  - 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:

\[ \hat{\theta}_{\text{MAP}} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]

MAP for Beta distribution

\[ E[\theta] = \alpha_H = \frac{\beta_H}{\beta_H + \beta_T} \]

\[ P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1}(1 - \theta)^{\beta_T + \alpha_T - 1}}{H(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- MAP: use most likely parameter:

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

- Beta prior equivalent to extra thumbtack flips
- As \( N \rightarrow \infty \), 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