What’s learning?
Point Estimation

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University
September 9th, 2009

What is Machine Learning?
Machine Learning

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

Data → Machine Learning → Understanding

Classification

from data to discrete classes
Object detection

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
Weather prediction revisited

Modeling sensor data

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

[Guestrin et al. ’04]
Similarity

finding data

Given image, find similar images

http://www.tiltomo.com/
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?

Embedding words

©2009 Carlos Guestrin

[Saul & Roweis '03] 23

[Joseph Turian] 24
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]

Bringing it all together…
HURLEY: Uh ... the Chinese people have water. (Sayid and Kate go to check it out.)

[SAYID]

[SAYID] holds the empty bottle in his hand and questions Sun.)

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

[SUN]

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

[SAYID]

[SUN]

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

[SUN]

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

[SUN]

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

[SUN]

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

[SUN]

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

[SUN]

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

[SUN]

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

[SUN]

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

[SUN]
Unsupervised learning of language

No supervision, only raw natural language sentences. Why?
- Certain languages do not have much annotated data
- “Learning without supervision” corresponds to the natural phenomenon of language acquisition

Machine learning can help:
- Uncover linguistic structure in observed sentences
  INPUT: Sequence of parts of speech.
  OUTPUT: Directed trees describing syntactic relations
- Model language acquisition in children
  INPUT: Speech utterance in one chunk
  OUTPUT: Segmented utterance into words

Output:

```
匈牙利总统等外国政府首脑将访华
```

Input:

```
NR NNP ETC NN NN NN AD VV NR
```

Shay Cohen et al.
Parallel Machine Learning

Yucheng Low

Need to take advantage of parallelism to stay ahead of the curve!
- Efficient Parallel / Distributed Belief Propagation
- Programming Abstractions for Machine Learning

Datasets are getting Larger.
- 13 Million Wikipedia Pages
- 3.6 Billion photos on Flickr

Processes are not getting faster.

Multi-modal activity recognition

Kate Spriggs et al.

Inputs:
First person vision (video)
Inertial measurement units

Output:
Beat eggs
Open box
Stir mix
Put pan in oven

Research challenges:
feature extraction and selection, temporal classification and segmentation, robustness to outliers
### Computational Cancer Genetics

(i,j): expression of gene i in tumor j

Goal: Infer k components of genes and the level of expression of each component in a tumor

Goal: Infer from “DNA gain and losses in chromosomal arms” data, an oncogenetic tree

*Manifold Learning, Dimensionality Reduction, Clustering, Graphical Models, Theory*

---

### 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

©2005-2009 Carlos Guestrin
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”
Recitations

- Very useful!
  - Review material
  - Present background
  - Answer questions
- Thursdays, 5:00-6:20pm in Gates Hillman 6115
- Special recitation 1:
  - tomorrow, Gates 6115, 5:00-6:20
  - Review of probabilities
- Special recitation 2 on Matlab
  - Monday, Sept. 14th 5:00-6:20pm GHC 6115

Staff

- Four Great TAs: Great resource for learning, interact with them!
  - Shay Cohen, GHC 5719, scohen@cs, Office hours: Tuesdays 2-4pm
  - Yucheng Low, GHC 8219, ylow@cs, Office hours: Wednesdays 4-6pm
  - Ekaterina Spriggs, GHC 8023, espriggs@cs, Office hours: Tuesdays 4-6pm
  - Babis Tsourakakis, GHC 8223, ctsourak@cs, Office hours: Fridays 11am-1pm
- Administrative Assistant
  - Michelle Martin, GHC 8001, x8-5537, michelle324@cs
Text Books

- Required Textbook:
  - Pattern Recognition and Machine Learning; Chris Bishop

- Secondary Textbook:
  - The Elements of Statistical Learning: Data Mining, Inference, and Prediction; Trevor Hastie, Robert Tibshirani, Jerome Friedman

- Optional Books:
  - Machine Learning; Tom Mitchell
  - Information Theory, Inference, and Learning Algorithms; David MacKay

Grading

- 5 homeworks (35%)
  - First one goes out 9/14
    - Start early, Start early, 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 Sept. 30th
  - Projects done individually, or groups of two students

- Midterm (15%)
  - Wed., Oct 21 in class

- 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 Michelle Martin, GHC 8001

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

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
Communication Channels

- Main channel for announcements, questions, etc. – Google Group:
  - [http://groups.google.com/group/10701-F09?hl=en](http://groups.google.com/group/10701-F09?hl=en)
  - Subscribe!

- For e-mailing instructors, always use:
  - [10701-instructors@cs.cmu.edu](mailto:10701-instructors@cs.cmu.edu)

- For announcements, subscribe to:
  - [10701-announce@cs](mailto:10701-announce@cs)
  - [https://mailman.srv.cs.cmu.edu/mailman/listinfo/10701-announce](https://mailman.srv.cs.cmu.edu/mailman/listinfo/10701-announce)

Sitting in & Auditing the Class

- Due to 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 $\alpha_H$ 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:

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

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

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

- 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 $\epsilon > 0$:

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

PAC Learning

PAC: Probably Approximately 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?

\[
P(|\hat{\theta} - \theta^*| \geq c) \leq 2e^{-2N\epsilon^2}
\]
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 $\theta$, we obtain a distribution over possible values of $\theta$

Bayesian Learning

- Use Bayes rule:
  \[ P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})} \]

- Or equivalently:
  \[ P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta) \]
Bayesian Learning for Thumbtack

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

- Likelihood function is simply Binomial:
  \[ P(\mathcal{D} \mid \theta) = \theta^H (1 - \theta)^T \]

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

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

- Bayesian inference:

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

- Integral is often hard to compute
MAP: Maximum a posteriori approximation

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

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

- 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