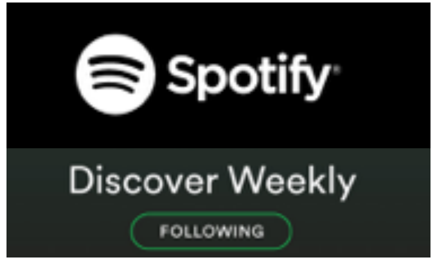
http://www.cs.washington.edu/education/courses/cse546/17au/

Machine Learning CSE546

Kevin Jamieson
University of Washington

September 28, 2017

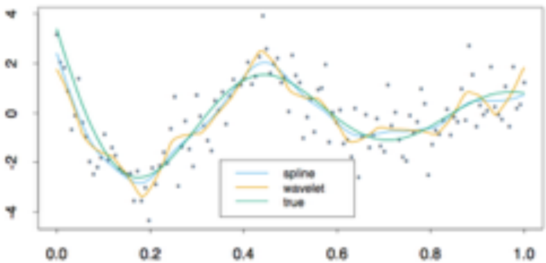


amazon prime

You may also like…

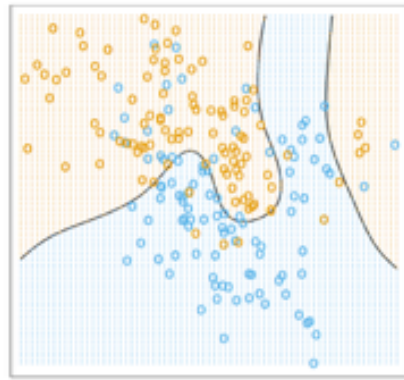
ML uses past data to make personalized predictions

Flavors of ML



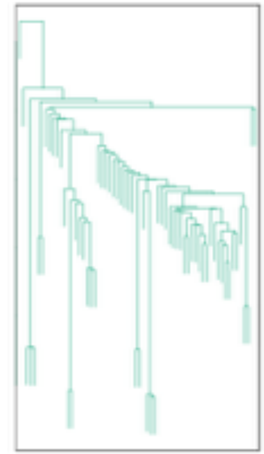
Regression

Predict continuous value:
ex: stock market, credit score, temperature, Netflix rating



Classification

Predict categorical value:
loan or not? spam or not? what disease is this?



Unsupervised Learning

Predict structure:
tree of life from DNA, find similar images, community detection

Mix of statistics (theory) and algorithms (programming)

Machine Learning Ingredients

- **Data**: past observations
- **Hypotheses/Models**: devised to capture the patterns in data
  - Does not have to be correct, just close enough to be useful
- **Prediction**: apply model to forecast future observations

Why is Machine Learning so popular, now?

- **“Big” Data**: the proliferation of the internet and smart phones has created consumer opportunities that *scale ($$$$$)*
- **Computing**: powerful, reliable, commoditized resources
- **Capitalism**: gives companies an edge (e.g., hedge funds)

Growth of Machine Learning

One of the most sought for specialties in industry today.

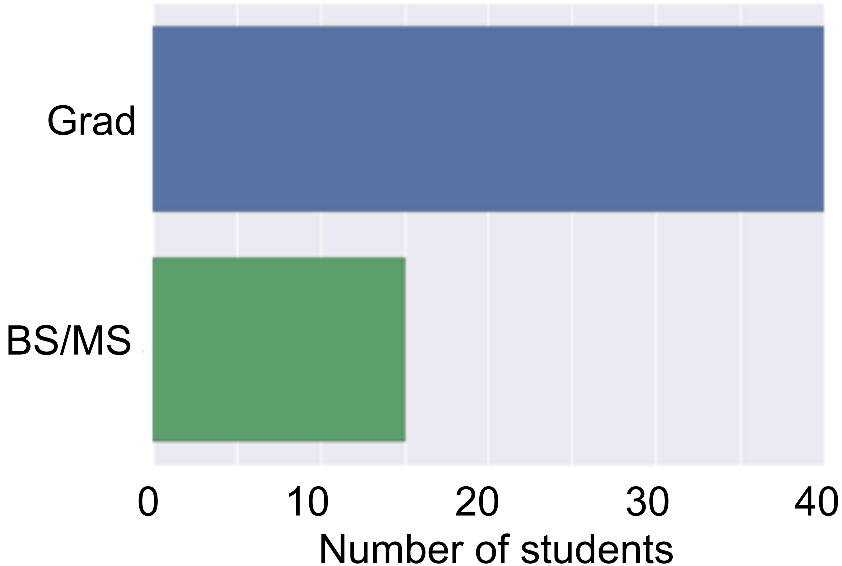
- Machine learning is preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology
  - Sensor networks
  - …

- This trend is accelerating, especially with **Big Data**
  - 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

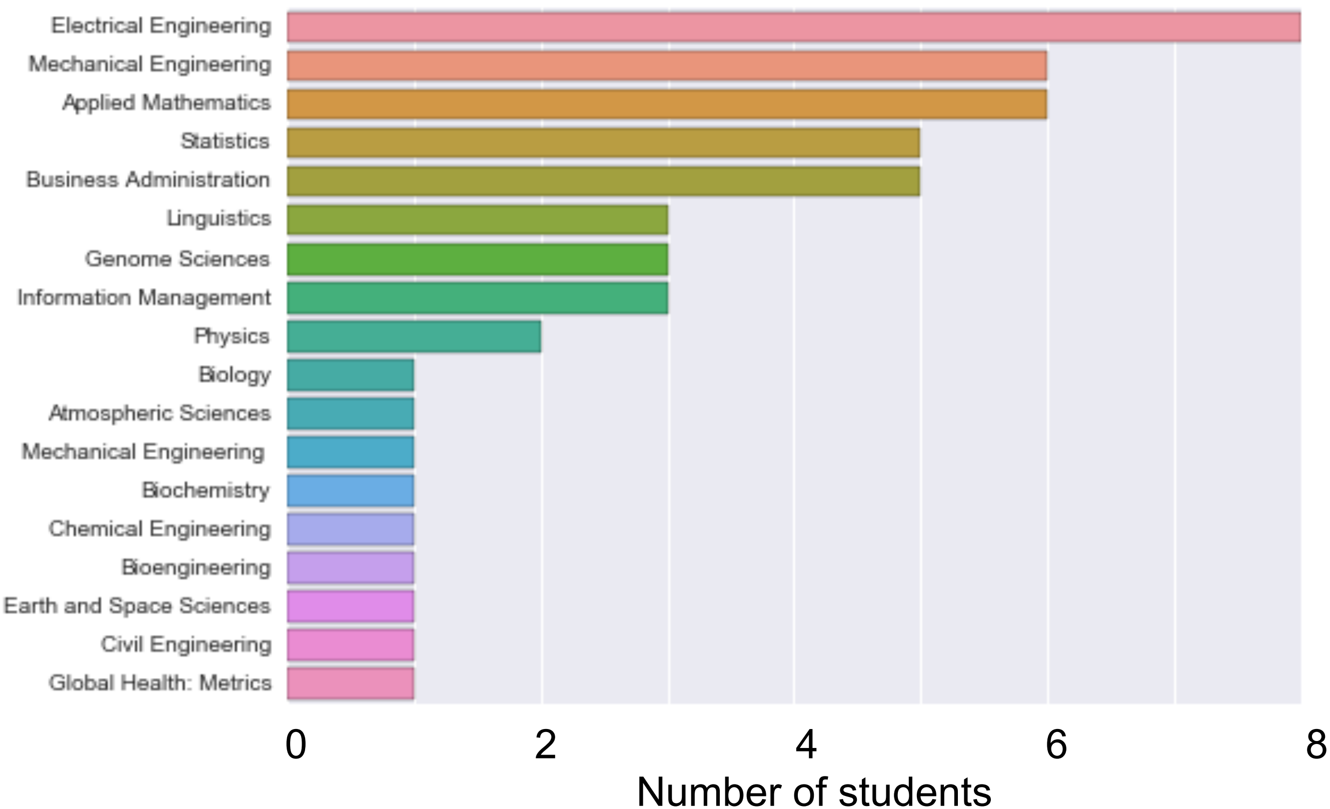
- Covers a wide range of Machine Learning techniques – from basic to state-of-the-art
- You will learn about the methods you heard about:
  - Point estimation, regression, logistic regression, optimization, nearest-neighbor, decision trees, boosting, perceptron, overfitting, regularization, dimensionality reduction, PCA, error bounds, SVMs, kernels, margin bounds, K-means, EM, mixture models, HMMs, graphical models, deep learning, reinforcement learning…
- Covers algorithms, theory and applications
- **It's going to be fun and hard work.**

Student makeup: CSE 55%



About 55 CSE students (total expected class size)

Student makeup: Non-CSE 45%



Welcome. You may also consider CSE 416 offered in the Spring.

Prerequisites

- Formally:
  - STAT 341, STAT 391, or equivalent
- Probability + statistics
  - Distributions, densities, marginalization, moments
- Math
  - Linear algebra, multivariate calculus
- Algorithms
  - Basic data structures, complexity
- Programming
  - Python
  - LaTeX
- Quick poll…

- **See website for review materials!**

Staff

- Four Great TAs: They are great resources in addition to the discussion board
  - **Nancy Wang:** Monday 4:00-5:00 PM, CSE 220
  - **Yao Lu:** Tuesday 2:30-3:30 PM, CSE 220
  - **Aravind Rajeswaran:** Wednesday 3:00-4:00 PM, CSE 220
  - **Dae Hyun Lee:** Thursday 1:30-2:30 PM, CSE 007

- Check Canvas Discussion board for exceptions/updates

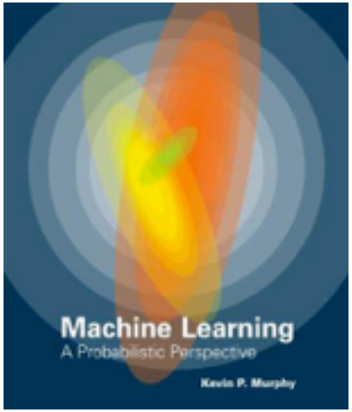
Communication Channels

- **Canvas Discussion board**
  - Announcements (e.g., office hours, due dates, etc.)
  - Questions (logistical or homework) - please participate and help others
  - All non-personal questions should go here

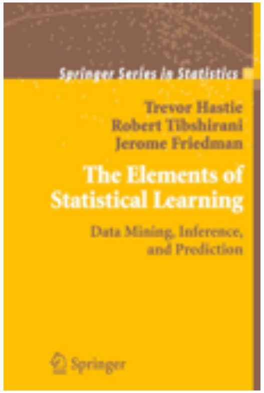
- For e-mailing instructors about personal issues and grading use:
  - cse546-instructors@cs.washington.edu
- Office hours limited to knowledge based questions. Use email for all grading questions.

Text Books

- Required Textbook:
  - ***Machine Learning: a Probabilistic Perspective*; Kevin Murphy**



- Optional Books (free PDF):
  - *The Elements of Statistical Learning: Data Mining, Inference, and Prediction;* Trevor Hastie, Robert Tibshirani, Jerome Friedman



Grading

- 5 homeworks (65%)
  - Each contains both theoretical questions and will have programming
  - Collaboration okay. You must write, submit, and understand your answers and code (which we may run)
  - Do not Google for answers.
- Final project (35%)
  - An ML project of your choice that uses real data

1. Code must be written in Python

2. Written work must be typeset using LaTeX

See website for tutorials… otherwise Google it.

Homeworks

- HW 0 is out (10 points, **Due next Thursday**)
  - Short and easy, gets you using Python and LaTeX
- HW 1,2,3,4 (25 points each)
  - They are not easy or short. Start early.
- Grade is minimum of the summed points and 100 points.
- **There is no credit for late work, receives 0 points.**
- **You must turn in all 5 assignments (even if late for 0 points) or else you will not pass.**

Projects (35%)

- An opportunity/intro for research in machine learning
- Grading:
  - We seek some novel exploration.
  - If you write your own code, great. We takes this into account for grading.
  - You may use ML toolkits (e.g. TensorFlow, etc), then we expect more ambitious project (in terms of scope, data, etc).
  - If you use simpler/smaller datasets, then we expect a more involved analysis.
- Individually or groups of two or three.
  - If in a group, the expectation are much
- Must involve real data
  - Must be data that you have available to you by the time of the project proposals
- It's encouraged to be related to your research, but must be something new you did this quarter
  - Not a project you worked on during the summer, last year, etc.
  - You also must have the data right now.

Enjoy!

- ML is becoming ubiquitous in science, engineering and beyond
- It’s one of the hottest topics in industry today
- This class should give you the basic foundation for applying ML and developing new methods
- The fun begins…

Maximum Likelihood Estimation

Machine Learning – CSE546
Kevin Jamieson
University of Washington

September 28, 2017

Your first consulting job

- *Billionaire*: I have special coin, if I flip it, what's the probability it will be heads?
- *You*: Please flip it a few times:

HHTHT

- *You*: The probability is: 3/5
- *Billionaire:* Why?

Coin – Binomial Distribution

- **Data**: sequence *D= (HHTHT…)*, **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
  - Flips are i.i.d.: $P(\underset{x_1}{H}\underset{x_2}{H}\underset{x_3}{T}\underset{x_4}{H}\underset{x_5}{T}) = P_1(H)P_2(H)P_3(T)P_4(H)P_5(T)$
    - Independent events $= \theta\,\theta(1-\theta)\theta(1-\theta)$
    - Identically distributed according to Binomial distribution $= \theta^3(1-\theta)^2$

- $P(\mathcal{D}|\theta) = \theta^k(1-\theta)^{n-k}$

Maximum Likelihood Estimation

- **Data**: sequence *D= (HHTHT…)*, **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ

$$P(\mathcal{D}|\theta) = \theta^k (1-\theta)^{n-k}$$

- Maximum likelihood estimation (MLE): Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\widehat{\theta}_{MLE} &= \arg\max_\theta \; P(\mathcal{D}|\theta) \\ &= \arg\max_\theta \; \log P(\mathcal{D}|\theta)\end{aligned}$$

Your first learning algorithm

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \ \log P(\mathcal{D}|\theta)$$

$$= \arg\max_{\theta} \ \log \theta^k (1-\theta)^{n-k}$$

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

$$\frac{\partial}{\partial\theta}\left[k \log(\theta) + (n-k)\log(1-\theta)\right]$$

$$= \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0$$

$$k(1-\theta) - (n-k)\theta = 0$$

$$k - \theta n = 0 \qquad \hat{\theta}_{MLE} = k/n$$

How many flips do I need?

$$\widehat{\theta}_{MLE} = \frac{k}{n}$$

- *You*: flip the coin 5 times. *Billionaire*: I got 3 heads.

$$\widehat{\theta}_{MLE} = 3/5$$

- *You*: flip the coin 50 times. *Billionaire*: I got 20 heads.

$$\widehat{\theta}_{MLE} = 20/50 = 2/5$$

- *Billionaire:* Which one is right? Why?

Simple bound (based on Hoeffding's inequality)

If R.V. X has density $f(x)$ then $\mathbb{E}[g(X)] = \int g(x) f(x) dx$

- For **n flips** and **k heads** the MLE is **unbiased** for true θ^*:

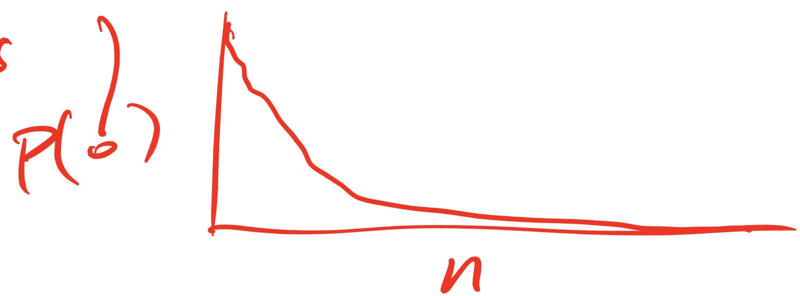
$$\widehat{\theta}_{MLE} = \frac{k}{n} \qquad\qquad \mathbb{E}[\widehat{\theta}_{MLE}] = \theta^*$$

- Hoeffding's inequality says that for any $\varepsilon > 0$:

$$P(|\widehat{\theta}_{MLE} - \theta^*| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

$n=5$ $\frac{2}{5}$ $n=50$ $\frac{3}{5}$

$\varepsilon = .05$



PAC Learning

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

$$P(|\widehat{\theta}_{MLE} - \theta^*| \geq \epsilon) \leq \boxed{2e^{-2n\epsilon^2} = \delta}$$

Solve for epsilon

$$\varepsilon = \sqrt{\frac{\log(2/\delta)}{2n}}$$

w.p. $\geq 1-\delta$

$$|\hat{\theta}_{MLE} - \theta^*| \leq \sqrt{\frac{\log(2/\delta)}{2n}} = 0.1$$

What about continuous variables?

- *Billionaire*: What if I am measuring a **continuous variable**?
- ***You*: Let me tell you about Gaussians…**

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

$$X \sim \mathcal{N}(\mu, \sigma)$$

Some properties of Gaussians

- affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \quad \rightarrow Y \sim N(a\mu+b,a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \quad \rightarrow Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\dots,x_N\}$ (e.g., exam scores):

$$P(\mathcal{D}|\mu,\sigma) = P(x_1,\dots,x_n|\mu,\sigma) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- Log-likelihood of data:

$$\log P(\mathcal{D}|\mu,\sigma) = -n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

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

$$= -\sum_{i=1}^{n}\frac{d}{d\mu}\frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= +\sum_{i=1}^{n}\frac{(x_i-\mu)}{\sigma^2} \longrightarrow 0$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu) = 0$$

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

MLE for variance

$\log(ab) = \log(a) + \log(b)$

- Again, set derivative to zero:

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

$$= -\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^3} = 0$$

mult by $\frac{\sigma^3}{n}$ both sides

$$-n\sigma^2 + \sum_i (x_i-\mu)^2 = 0$$

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

Learning Gaussian parameters

- MLE:

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\mathbb{E}[\widehat{\mu}_{MLE}] = \mu$

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

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

$$\mathbb{E}[\widehat{\sigma^2}_{MLE}] \neq \sigma^2$$

- Unbiased variance estimator:

$$\widehat{\sigma^2}_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{MLE})^2$$

Recap

- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood
  - Choose an optimization procedure
    - E.g., set derivative to zero to obtain MLE
  - Justifying the accuracy of the estimate
    - E.g., Hoeffding's inequality