http://www.cs.washington.edu/education/courses/cse546/13au/

What's learning? Point Estimation

Machine Learning – CSE546
Carlos Guestrin
University of Washington

September 25, 2013

What is Machine Learning ?

Machine Learning

loss function

Study of algorithms that
- improve their performance
- at some task ← classification, regression, clustering...
- with experience ← data

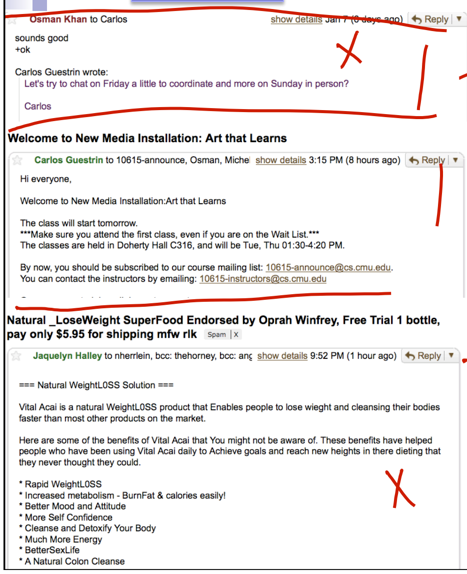
Classification

X Y = {spam, not spam}

from data to discrete classes

Spam filtering

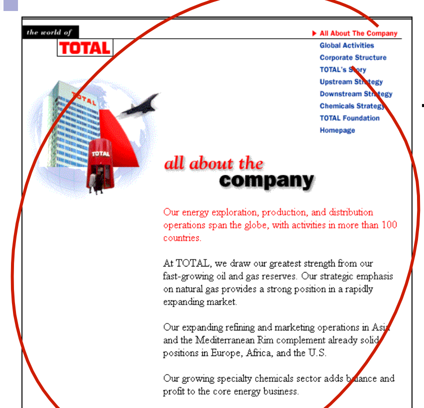
data   prediction



x f $f(x) \rightarrow$ not spam

x $f(x) \rightarrow$ Spam

Text classification



→ Company home page

vs

Personal home page

vs

Univeristy home page

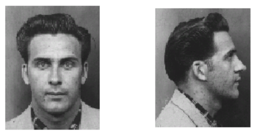
vs

…

Object detection

(Prof. H. Schneiderman)

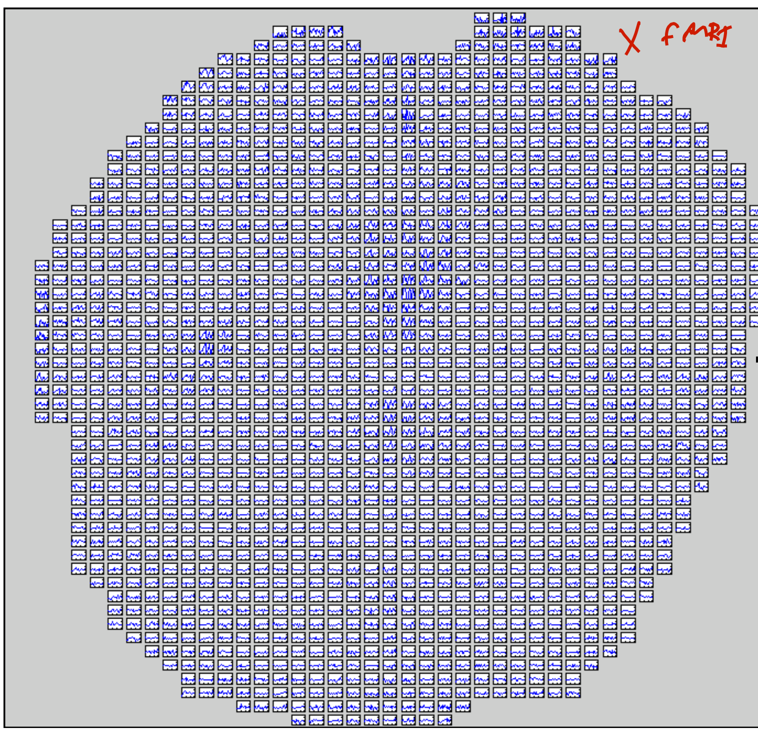
X → face, not face



Example training images
for each orientation

©2005-2013 Carlos Guestrin

7



X fMRI

Y

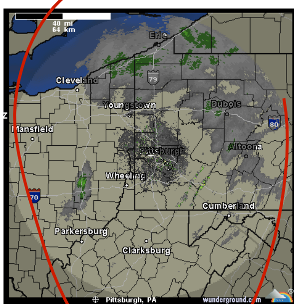
reading your mind mind

Reading
a noun
(vs verb)

[Rustandi et al.,
2005]

8

Weather prediction

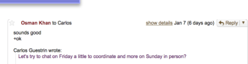


Rain, Rain, ...

The classification pipeline

email label

Training



$\{X, Y\}_i \; \ldots$

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk

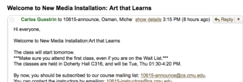
$\boxed{\text{learn } f}$

$f(x) \rightarrow y$

Testing

using new X_{new} ... predict $Y \leftarrow f(X_{new})$

Welcome to New Media Installation: Art that Learns



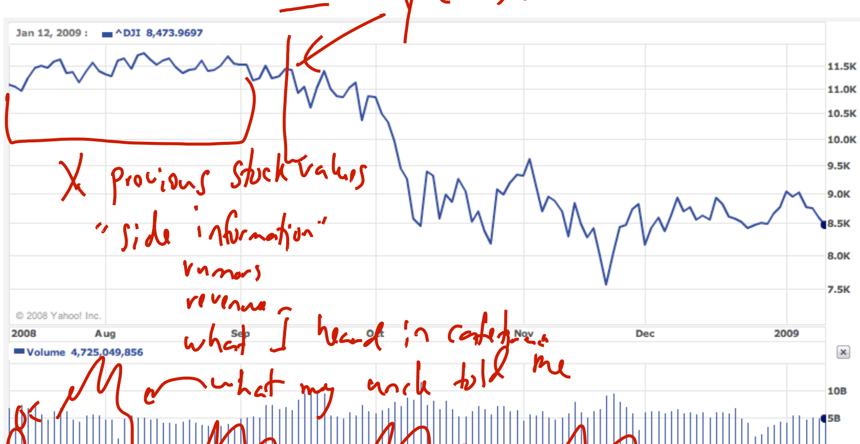
Regression

$X \rightarrow Y \leftarrow \mathbb{R}$

predicting a numeric value

Stock market

$Y \leftarrow$ stock value tomorrow



X previous stock values

"side information"

rumors

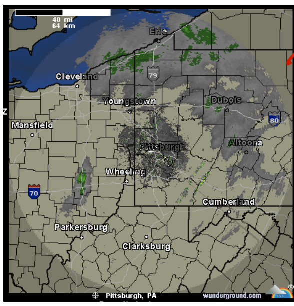
revenue

what I heard in cafeteria

or what my uncle told me

Weather prediction revisted

Rain

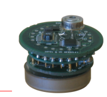


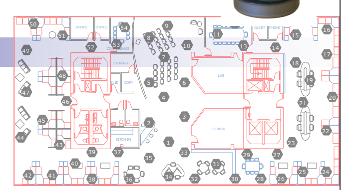
Temperature

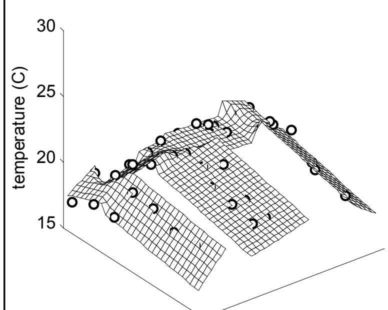
59°F

17°C ?

Modeling sensor data







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

[Guestrin et al. '04]

Unsupervised learning

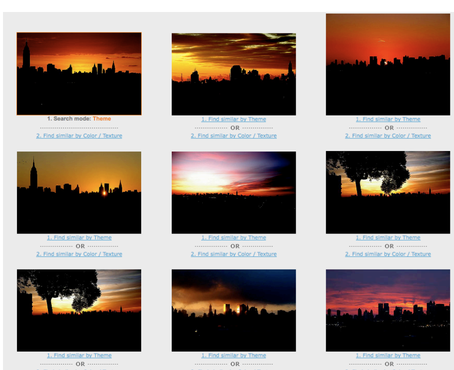
Similarity

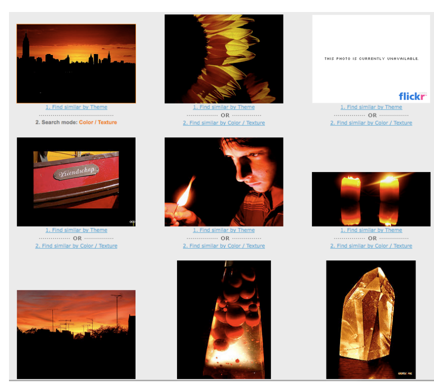
finding data

©2005-2013 Carlos Guestrin

15

Given image, find similar images



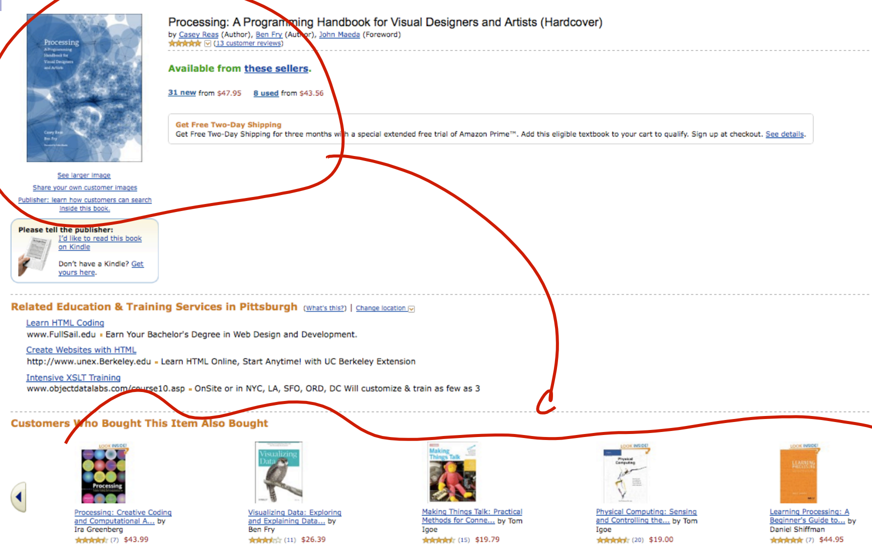


©2005-2013 Carlos Guestrin

http://www.tiltomo.com/

16

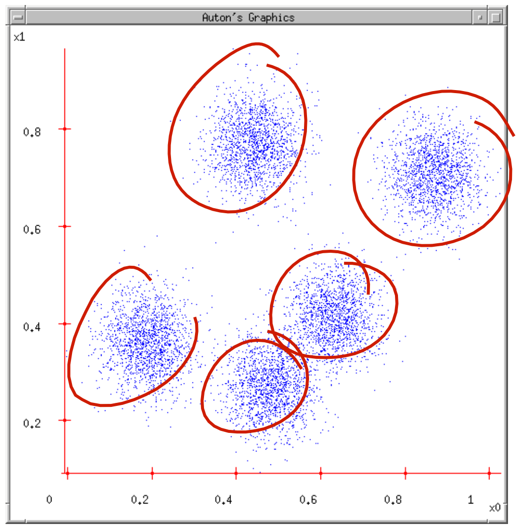
Similar products



Clustering

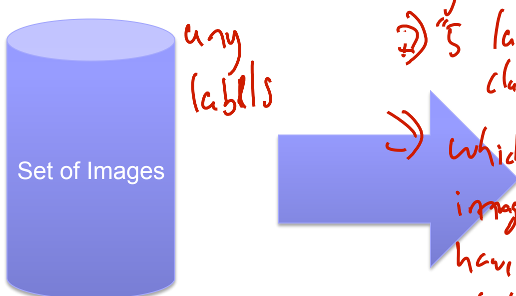
discovering structure in data

Clustering Data: Group similar things



Clustering images

with out any labels



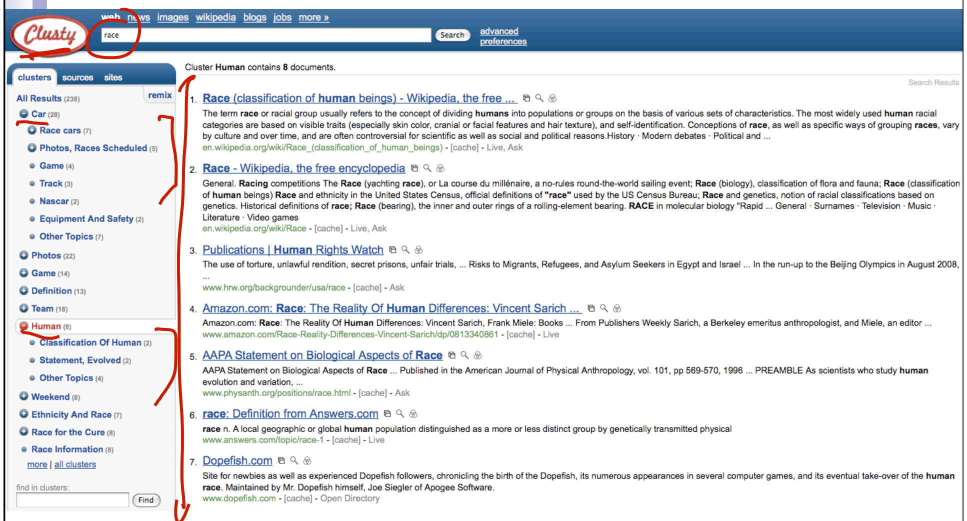
learn 2 things

1) "5 labels" classes

2) which images have which labels

[Goldberger et al.]

Clustering web search results



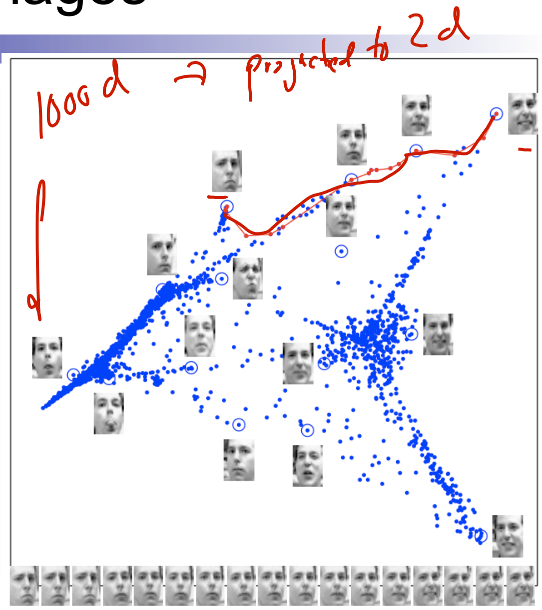
Embedding

"visualizing data"

Embedding images

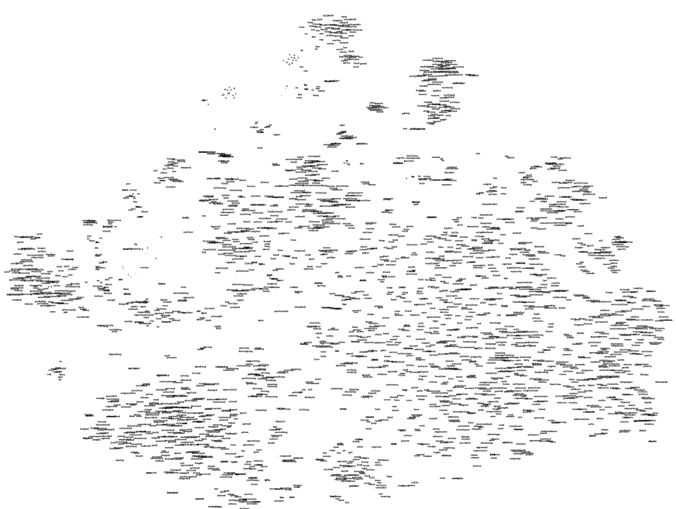
Images have thousands or millions of pixels.

Can we give each image a coordinate,
such that similar images are near each other?



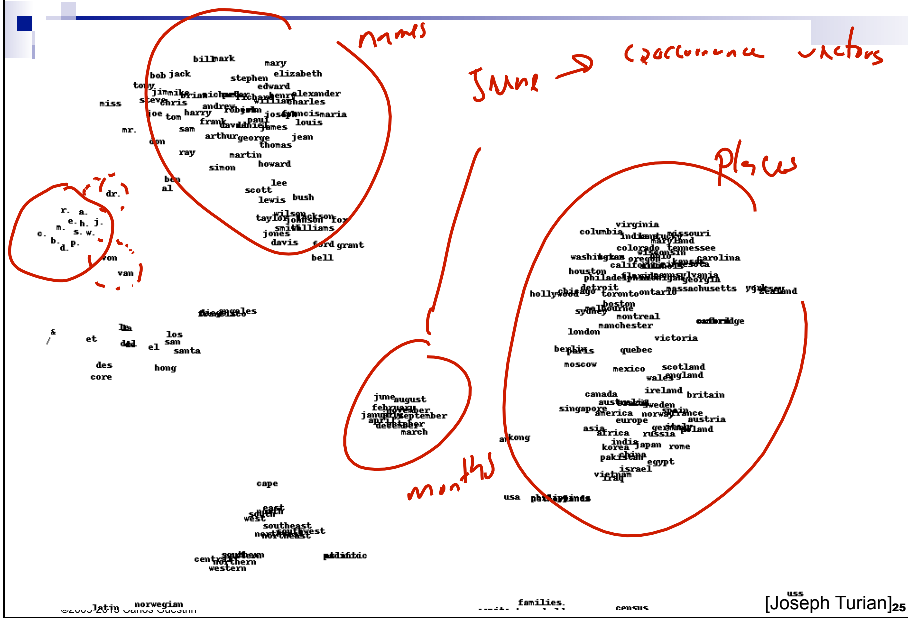
[Saul & Roweis '03]

Embedding words



[Joseph Turian]

Embedding words (zoom in)



[Joseph Turian]

MANY

supervised learning
eg classification ← → unsupervised learning
eg clustering

Reinforcement Learning

training by feedback

learning to act from sporadic feedback

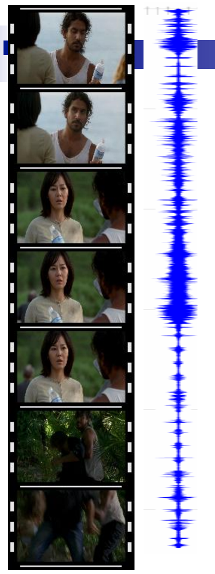
Learning to act

- 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…

Combining video, text and audio



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

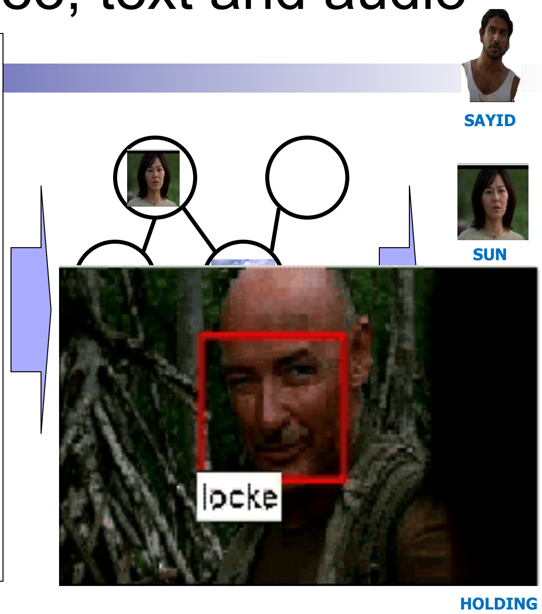
[EXT. BEACH - CRASH SITE]

(Sayid holds the empty bottle in his hand and questions Sun.)

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

[EXT. JUNGLE]

(Sawyer is walking through the jungle. He reaches a spot. He kneels down and looks back to check that no one's followed him.)

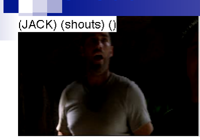


Taskar et al.

Automatically Discovered and Labeled Actions

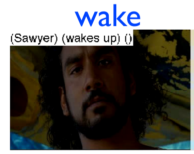
smile

shout



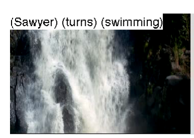
sit down

wake

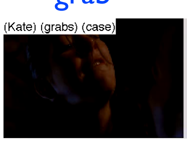


follow

swim

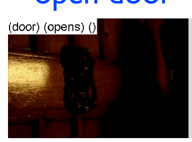


grab



kiss

open door



point

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, naïve Bayes, logistic regression, nearest-neighbor, decision trees, boosting, perceptron, 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

- Formally:
  - STAT 341, STAT 391, or equivalent
- Probabilities
  - Distributions, densities, marginalization…
- Basic statistics
  - Moments, typical distributions, regression…
- Algorithms
  - Dynamic programming, basic data structures, complexity…
- Programming
  - ~~R~~ Python will be very useful, but we'll help you get started
- We provide some background, but the class will be fast paced
- Ability to deal with "abstract mathematical concepts"

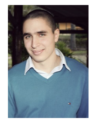
Recitations & Python

- We'll run an ***optional*** recitations:
  - Tuesdays @5:30pm
  - Location ~~TBD~~ LOW 101
- We are recommending Python for homeworks!
  - There are many resources to get started with Python online
  - We'll run an ***optional*** tutorial:
    - First recitation: Tuesday 10/1 @5:30pm

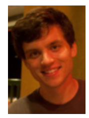
Staff

- Three Great TAs: Great resource for learning, interact with them!
  - **Eric Lei**
    Office hours: Fridays 1:30-3:30pm

  - **Marco Ribeiro**
    Office hours: Tuesdays 1:30-3:20pm



  - **Tyler Johnson**
    Office hours: Mondays 3-5pm



  - Prof: **Carlos Guestrin**
    Office hours: Wednesdays 10:30-11:30am

Communication Channels

- Only channel for announcements, questions, etc. – Catalyst Group:
  - https://catalyst.uw.edu/gopost/board/tbjohns/34218/
  - Subscribe!
  - All non-personal questions should go here
  - Answering your question will help others
  - Feel free to chime in
- For e-mailing instructors about personal issues, use:
  - cse546-instructors@cs.washington.edu

Text Books

- Required Textbook:
  - Machine Learning: a Probabilistic Perspective; Kevin Murphy
- Optional Books:
  - Pattern Recognition and Machine Learning; Chris Bishop
  - The Elements of Statistical Learning: Data Mining, Inference, and Prediction; Trevor Hastie, Robert Tibshirani, Jerome Friedman
  - Machine Learning; Tom Mitchell
  - Information Theory, Inference, and Learning Algorithms; David MacKay

Grading

- 4 homeworks (35%)
  - First one goes out 9/30
    - Start early, Start early, Start early, Start early, Start early, Start early, Start early, Start early, 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 (30%)
  - Full details out around 10/9
  - Projects done individually, or groups of two students
- Midterm (15%)
  - Wed., 10/30 in class
- Final (20%)
  - TBD by registrar

Homeworks

- Homeworks are hard, start early ☺
- Due in the beginning of class
- 33% subtracted per late day
- You have 3 LATE DAYS to use for homeworks only throughout the quarter
  - Please plan accordingly and after that don't be about deadlines, travel,… ☺
- All homeworks **must be handed in**, even for zero credit
- Use Catalyst to submit homeworks

- 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

Projects

FUN !! ☺

- An opportunity to exercise what you learned and to learn new things
- Individually or groups of two
- Must involve real data
  - Must be data that you have available to you by the time of the project proposals
- Must involve machine learning
- 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.

- Full details in a couple of weeks

- Wed., October 23 at 9:00am: **Project Proposals**
- Mon., November 11 at 9:00am: **Project Milestone**
- Wed., December 4, 3-5pm: **Poster Session** ← atrium
- Mon., December 9 at 9:00am: **Project Report**

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…

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:

$P(H) = 3/5$

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

Thumbtack – Binomial Distribution

HHTHT
indep
indentically
dist.
IID

- P(Heads) = θ, P(Tails) = 1-θ

$P(D|\theta) = P(HHTHT) = \theta\,\theta\,(1-\theta)\,\theta\,(1-\theta)$
$= \theta^3 (1-\theta)^2$

- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence *D* of α_H Heads and α_T Tails

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

ln monotonic ⇒ $\arg\max_x f(x) = \arg\max_x \ln f(x)$ $f(x) > 0\ \forall x$

Maximum Likelihood Estimation

simplifies calculations

iid

- **Data:** Observed set *D* of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
  - What's the objective function?

$\max_\theta P(D|\theta) = \max_\theta \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

$\max_x f(x) = -\min_x -f(x)$

- MLE: Choose θ that maximizes the probability of observed data:

$$\widehat{\theta}_{MLE} = \arg\max_\theta P(\mathcal{D} \mid \theta)$$
$$= \arg\max_\theta \underbrace{\ln P(\mathcal{D} \mid \theta)}_{\text{log likelihood}}$$

$-\ln P(D|\theta)$
$\min_\theta$
loss function

Your first learning algorithm

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

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

$\ln ab = \ln a + \ln b$

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

$\frac{d}{d\theta}(f+g)$ = sum deriv.

$\frac{d}{d\theta} \ln \theta = \frac{1}{\theta}$

$\frac{d}{d\theta} \ln(1-\theta) = \frac{-1}{1-\theta}$

loss function: $-\ln P(D|\theta)$ ← if minimize

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

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

$$= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0$$

$$\theta = \frac{\alpha_H}{\alpha_H + \alpha_T} \Big\} = \frac{3}{5} \ !!$$

How many flips do I need?

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{3}{5}$$

- 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)

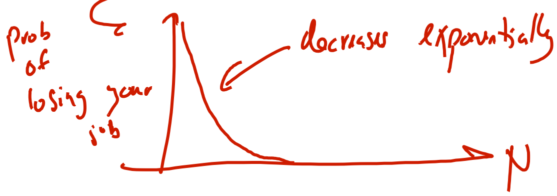
Sample complexity

- For $N = \alpha_H + \alpha_T$, and $\widehat{\theta}_{MLE} = \dfrac{\alpha_H}{\alpha_H + \alpha_T}$

$|\theta^* - \hat{\theta}_{MLE}|$?

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

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



PAC Learning

$1-\delta = 0.95$ $\quad \varepsilon = 0.1$

- 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? ← Sample complexity

$$P(|\,\widehat{\theta} - \theta^*\,| \geq \epsilon) \;\leq\; 2e^{-2N\epsilon^2} \leq \delta \equiv \text{prob lose my job}$$

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

$$N \geq \frac{\ln \frac{2}{\delta}}{2\varepsilon^2}$$

if $\delta = 0.05$, $\varepsilon = 0.1$

$N \geq 184.4$ flips

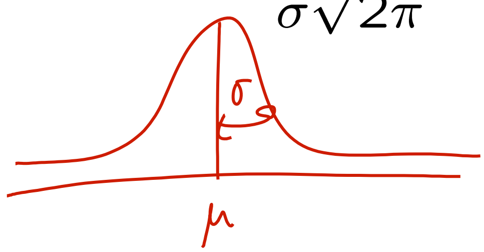
loose bound

much more sensitive to ε than δ

What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me? Salary of employees
- **You say: Let me tell you about Gaussians…**

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



Some properties of Gaussians

- affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \;\rightarrow\; Y \sim N(a\mu+b, a^2\sigma^2)$
- Sum of Gaussians
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \;\rightarrow\; Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

Learning a Gaussian

Num Score

$x_1 : 85$
$x_2 = 92$
$x_3 = 99$
$\vdots$

- Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., exam scores

why??

MLE

- Learn parameters
  - Mean : μ
  - Variance : σ^2

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

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

MLE for Gaussian

standard Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_N\}$:

$$P(\mathcal{D} \mid \mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$\mu_{MLE}, \sigma^2_{MLE} = \arg\max_{\mu,\sigma} P(D \mid \mu,\sigma^2) = \arg\max_{\mu,\sigma} \ln P(D \mid \mu,\sigma)$$

- Log-likelihood of data:

$$\max_{\mu,\sigma} \quad \ln P(\mathcal{D} \mid \mu,\sigma) = \ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right]$$

$$= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$\ln \prod = \sum \ln$

$\ln e^{\text{something}} = \text{something}$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

$$\frac{d}{d\mu}\ln P(\mathcal{D}\mid\mu,\sigma) = \frac{d}{d\mu}\left[-N\ln\cancel{\sigma\sqrt{2\pi}} - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right] = 0$$

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

$$= \sum_{i=1}^{N} \frac{x_i-\mu}{\cancel{\sigma^2}} = 0 \Rightarrow \sum_{i=1}^{N} x_i-\mu = 0$$

$$\Rightarrow N\mu = \sum_{i=1}^{N} x_i \Rightarrow \mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

!!

MLE for variance

$\max_{a,b} f(a,b)$

$\frac{\partial}{\partial a} f(a,b)=0$ & $\frac{\partial}{\partial b} f(a,b)=0$

simultaneous equ.

- Again, set derivative to zero:

from previous slide $\hat{\mu}_{MLE}$ didn't depend on σ

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

$$= \frac{d}{d\sigma}\left[-N\ln\sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N}\frac{d}{d\sigma}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

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

Learning Gaussian parameters

- MLE:

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

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

now you can justify these

- BTW. MLE for the variance of a Gaussian is **biased**
  - Expected result of estimation is **not** true parameter!
  - Unbiased variance estimator:

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

What you need to know…

- Learning is…
  - Collect some data
    - E.g., thumbtack 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
  - Collect the big bucks
- Like everything in life, there is a lot more to learn…
  - Many more facets… Many more nuances…
  - The fun will continue…