http://www.cs.washington.edu/education/courses/cse546/16au/

What's learning? Point Estimation

Machine Learning – CSE546
Sham Kakade
University of Washington

September 28, 2016

What is Machine Learning ?

Machine Learning

Study of algorithms that

- improve their <u>performance</u>
- at some <u>task</u>
- with <u>experience</u>

Classification

from data to discrete classes

Spam filtering

data prediction

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

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: ang show details 9:52 PM (1 hour ago) Reply

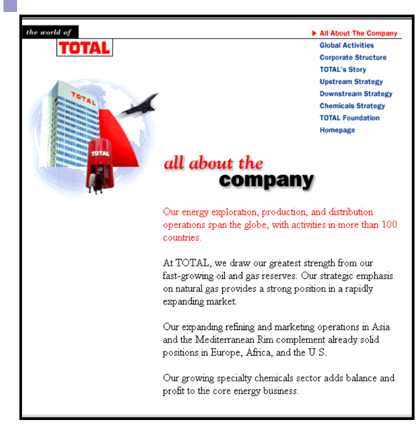
=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* BetterSexLife
* A Natural Colon Cleanse

Text classification



→ Company home page

vs

Personal home page

vs

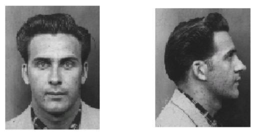
University home page

vs

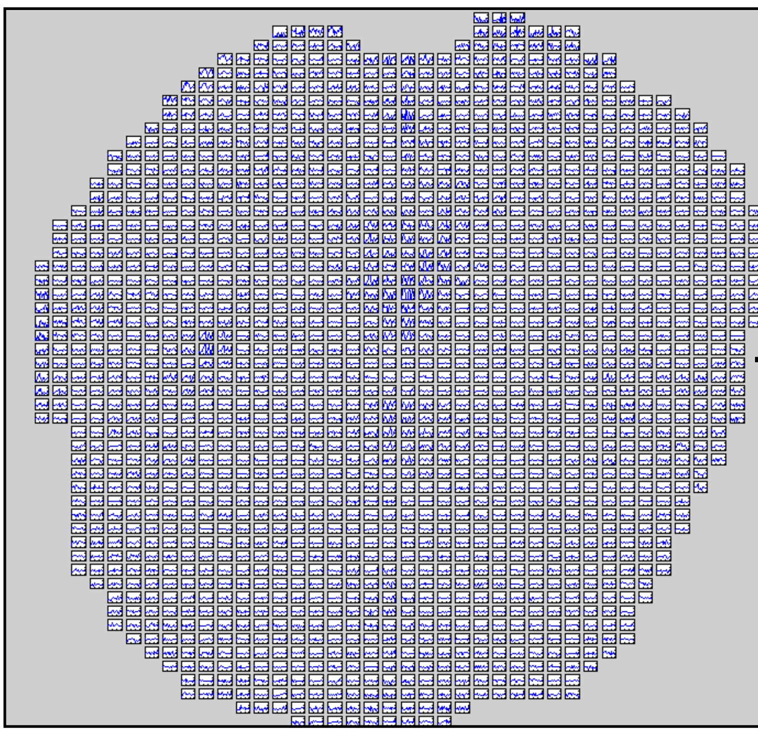
…

Object detection

(Prof. H. Schneiderman)



Example training images
for each orientation



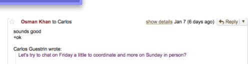
Reading
a noun
(vs verb)

[Rustandi et al.,
2005]

Weather prediction

The classification pipeline

Training

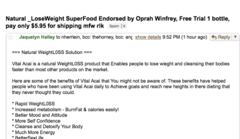


Osman Khan to Carlos — show details Jan 7 (6 days ago) Reply

sounds good
+ok

Carlos Guestrin wrote:
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos



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

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: anç — show details 9:52 PM (1 hour ago) Reply

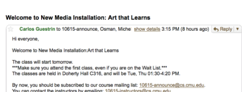
=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* BetterSexLife
* A Natural Colon Cleanse

Testing



Welcome to New Media Installation: Art that Learns

Carlos Guestrin to 10615-announce, Osman, Michel — show details 3:15 PM (8 hours ago) Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu

Regression

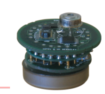
predicting a numeric value

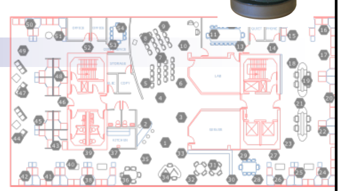
Stock market

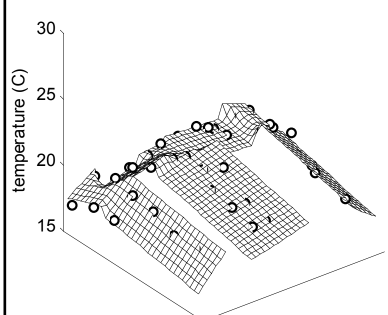
Weather prediction revisited

Temperature

Modeling sensor data





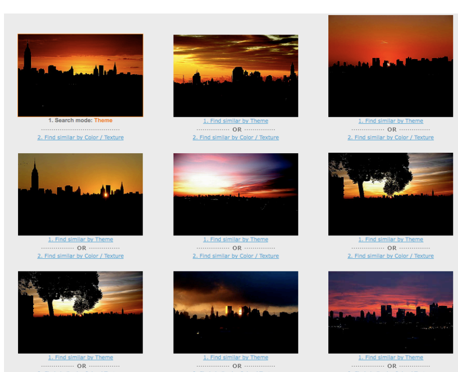


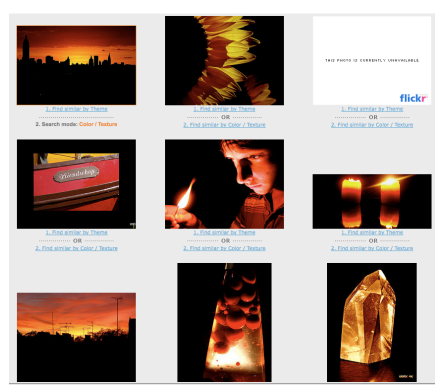
- Measure temperatures at some locations
- Predict temperatures throughout the environment

Similarity

finding data

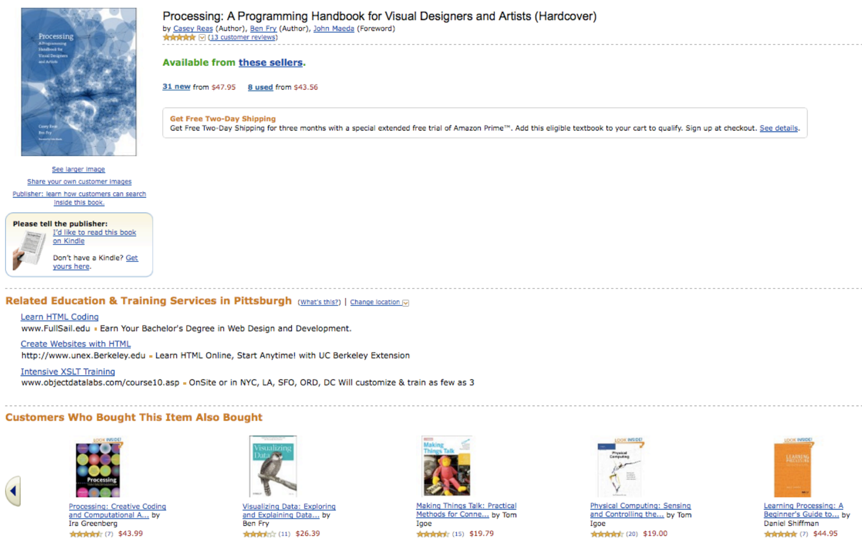
Given image, find similar images





http://www.tiltomo.com/

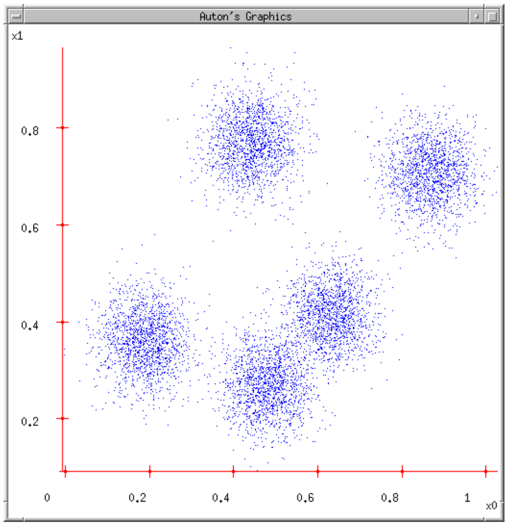
Similar products



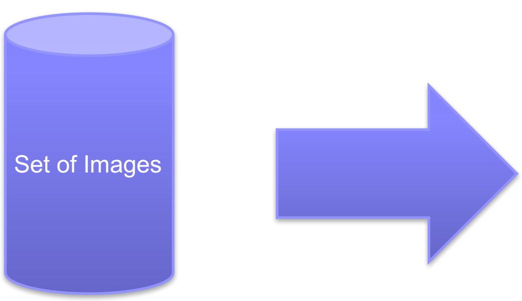
Clustering

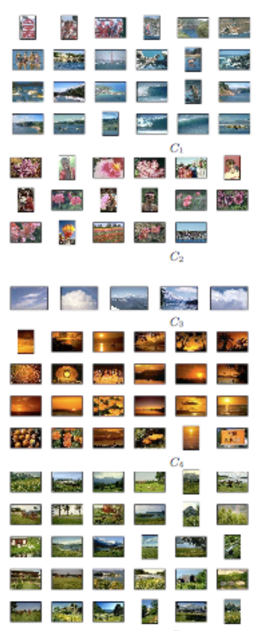
discovering structure in data

Clustering Data: Group similar things



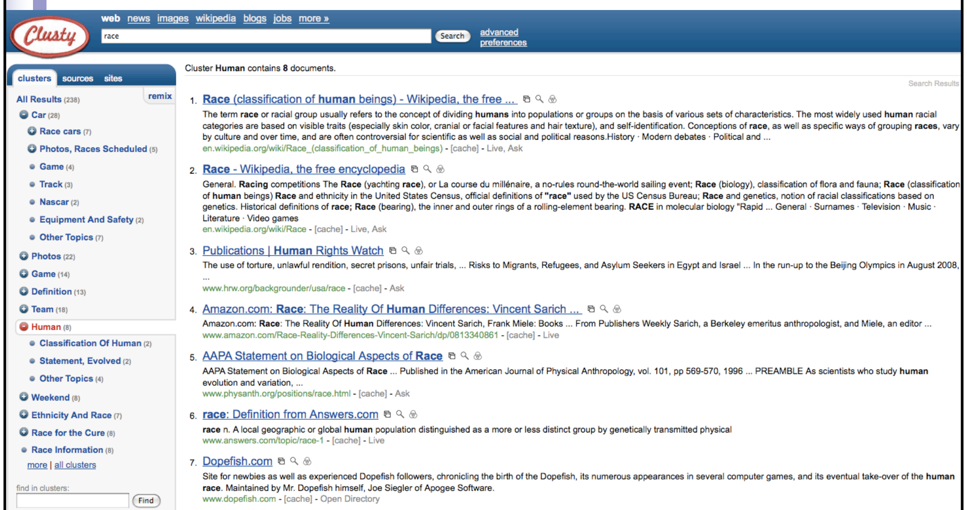
Clustering images





[Goldberger et al.]

Clustering web search results



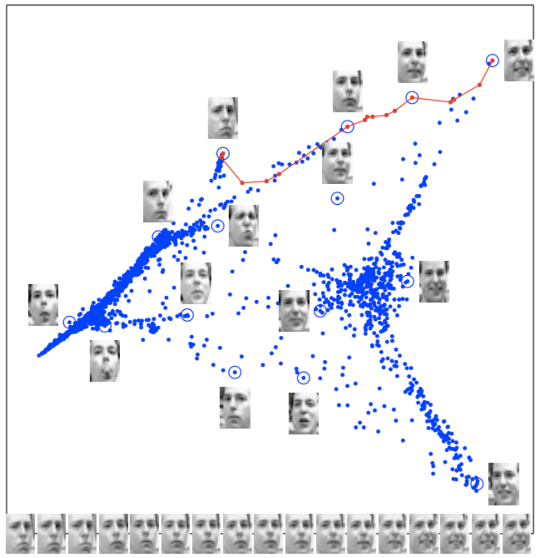
Embedding

visualizing data

Embedding images

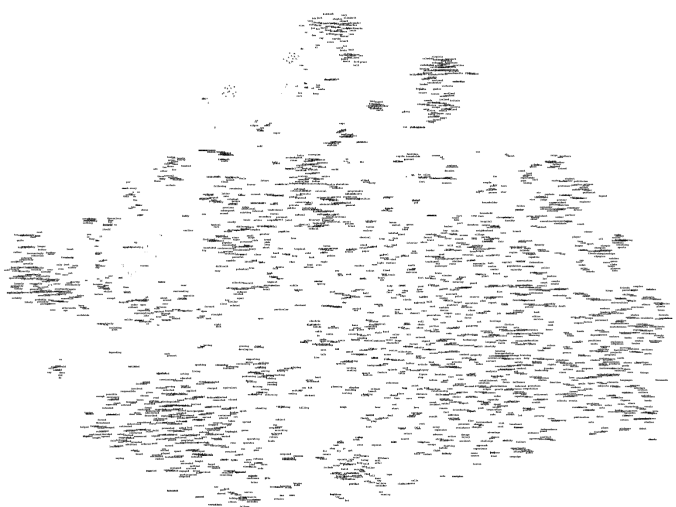
Images have thousands or millions of pixels.

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



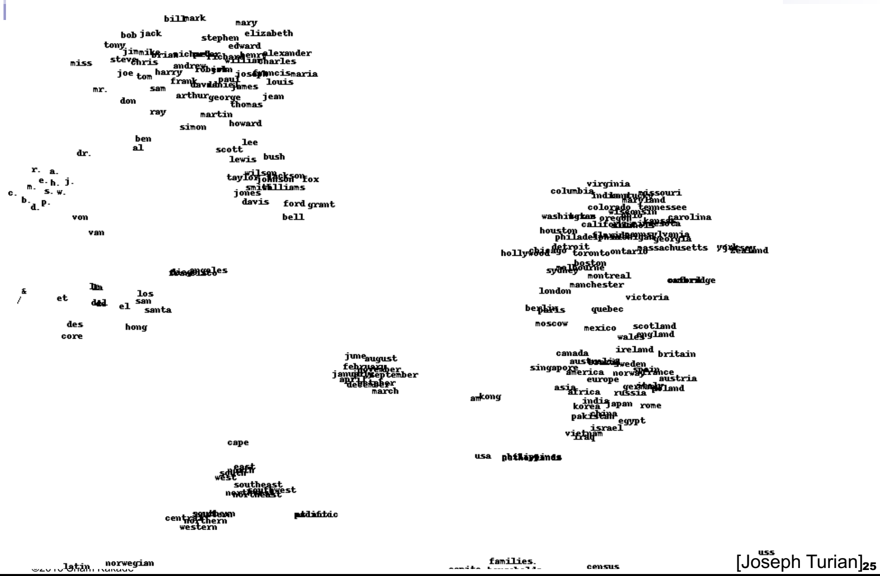
[Saul & Roweis '03]

Embedding words



[Joseph Turian]

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]

Impact

What are the biggest successes?

Successes

- Speech Recognition
  - SIRI, Alexa, etc.
- Computer vision
  - ImageNet
- Alpha-Go
  - Game playing
  - Go was 'solved' with ML/AI
- And more:
  - Natural language processing
  - Robotics (self-driving cars?)
  - Medical analysis
  - Computational biology

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

Logistics

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.**

Prerequisites

- Linear algebra:
  - SVDs, eigenvectors, matrix multiplication
- Probabilities
  - Distributions, densities, marginalization…
- Basic statistics
  - Moments, typical distributions, regression…
- Algorithms
  - Dynamic programming, basic data structures, complexity…
- Programming
  - Python will be very useful
- 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:
  - Time/Location

- We are recommending Python for homeworks!
  - There are many resources to get started with Python online
  - We’ll run an ***optional*** tutorial:
    - First recitation: next week

Staff

- Three Great TAs: Great resource for learning, interact with them!
  - **Dae Hyun Lee**
    Office hours: TBD
  - **Angli Liu**
    Office hours: TBD
  - **Alon Milchgrub**
    Office hours: TBD

Communication Channels

- Announcements on Canvas.
- Use the Discussion board!
  - 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 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:
  - Understanding Machine Learning: From Theory to Algorithms; Shai Shalev-Shwartz and Shai Ben-David.
  - 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 (65%)
  - First posted today
    - Start early!
  - HW 1,2,4 (15%)
    - Collaboration allowed
    - You must write (and submit) your own code, which we may run.
    - You must write (and understand) your own answers.
  - HW 3 midterm (20%)
    - No collaboration allowed.
- Final project (35%)
  - Full details: see website
  - Projects done individually, or groups of two students

HW Policy (SEE WEBSITE)

- Homeworks are hard/long, start early
  - Heavy programming component.
  - They will build on themselves (you will re-use your code).
- 33% subtracted per late day.
- You have 2 LATE DAYS to use for homeworks throughout the quarter
  - Please plan accordingly.
  - No exceptions (aside from university policies).
- All homeworks **must be handed in**, even for zero credit.
- Use Canvas to submit homeworks.
- No collaboration allowed on HW 3
- Collaboration: HW 1,2,4
  - Each student writes (and understands) their own answers.
  - You may **discuss** the questions.
  - 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 (35%)

- SEE WEBSITE
- An opportunity/intro for research.
  - encouraged to be related to your research, but must be something new you did this quarter
  - It's Not a project you worked on during the summer, last year, etc.
- Grading:
  - We seek some novel exploration.
  - If you write your own code, great. We take 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
- Must involve real data
  - Must be data that you have available to you by the time of the project proposals
- Must involve machine learning

(tentative) project dates (35%)

- Full details in a couple of weeks

- Mon., October 24, 5p: **Project Proposals**
- Mon., November 14, 5p: **Project Milestone**
- Thu., December 8, 9-11:30am: **Poster Session**
- Thu., December 15, 10am: **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
- Have fun..

A Data Science Job

- Someone asks you a stat/data science question:
  - She 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:
  - **She says: Why???**
  - You say: Because…

Thumbtack – Binomial Distribution

- P(Heads) = θ, P(Tails) = 1-θ
- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence *D* of α_H Heads and α_T Tails

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

Maximum Likelihood Estimation

- **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?

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

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

Your first learning algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

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

How many flips do I need?

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

- She says: I flipped 3 heads and 2 tails.
- You say: θ = 3/5, I can prove it!
- She says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it!
- **She says: What's better?**
- You say: Humm… The more the merrier???
- She 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 $\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$
- Let θ^* be the true parameter, for any $\varepsilon>0$:

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

PAC Learning

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the thumbtack parameter θ, within ε = 0.1, with probability at least 1-δ = 0.95. How many flips?

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

What about continuous variables?

- She says: If I am measuring a continuous variable, what can you do for me?
- **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 ~ $N(\mu,\sigma^2)$
  - Y = aX + b ➔ Y ~ $N(a\mu+b,a^2\sigma^2)$

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

Learning a Gaussian

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

- Learn parameters
  - Mean
  - Variance

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

MLE for Gaussian

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

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

- Log-likelihood of data:

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

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

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

MLE for variance

- Again, set derivative to zero:

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

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

Learning Gaussian parameters

- MLE:

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

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

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

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

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
- Like everything in life, there is a lot more to learn…
  - Many more facets… Many more nuances…
  - More later…