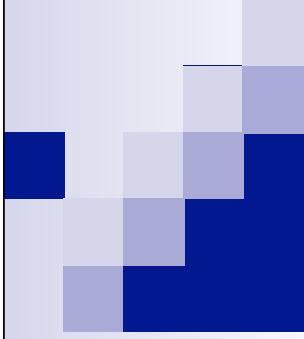


<http://courses.cs.washington.edu/courses/csep546/14wi/>



What's learning? Point Estimation

Machine Learning – CSEP546

Carlos Guestrin

University of Washington

December 6, 2014

©2005-2014 Carlos Guestrin

1



What is Machine Learning ?

©2005-2014 Carlos Guestrin

2

Machine Learning

Study of algorithms that

- improve their performance
- at some task
- with experience



©2005-2014 Carlos Guestrin

3

Classification

from data to discrete classes

©2005-2014 Carlos Guestrin

4

	data	prediction
1	<p>Osman Khan to Carlos show details Jan 7 (6 days ago) Reply</p> <p>sounds good +ok</p> <p>Carlos Guestrin wrote: Let's try to chat on Friday a little to coordinate and more on Sunday in person?</p> <p>Carlos</p>	
2	<p>Welcome to New Media Installation: Art that Learns</p> <p>Carlos Guestrin to 10615-announce, Osman, Miche show details 3:15 PM (8 hours ago) Reply</p> <p>Hi everyone,</p> <p>Welcome to New Media Installation:Art that Learns</p> <p>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.</p> <p>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</p>	
3	<p>Natural LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only \$5.95 for shipping mfw rik <small>Seem X</small></p> <p>Jacquelyn Halley to rherlein, boc; thehorney, boc; ang show details 9:52 PM (1 hour ago) Reply</p> <p>=== Natural WeightLOSS Solution ===</p> <p>Vital Acai is a natural WeightLOSS product that Enables people to lose weight and cleansing their bodies faster than most other products on the market.</p> <p>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.</p> <ul style="list-style-type: none"> * Rapid WeightLOSS * 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 	

Company home page

vs

Personal home page

vs

Univeristy home page

vs

...

Object detection

(Prof. H. Schneiderman)

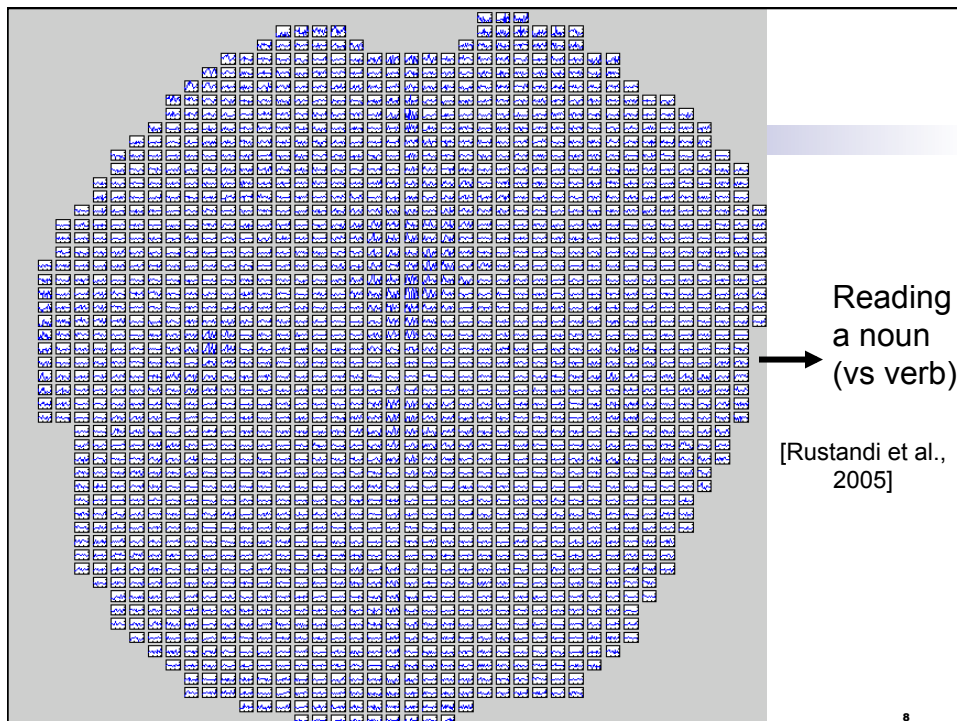


Example training images
for each orientation

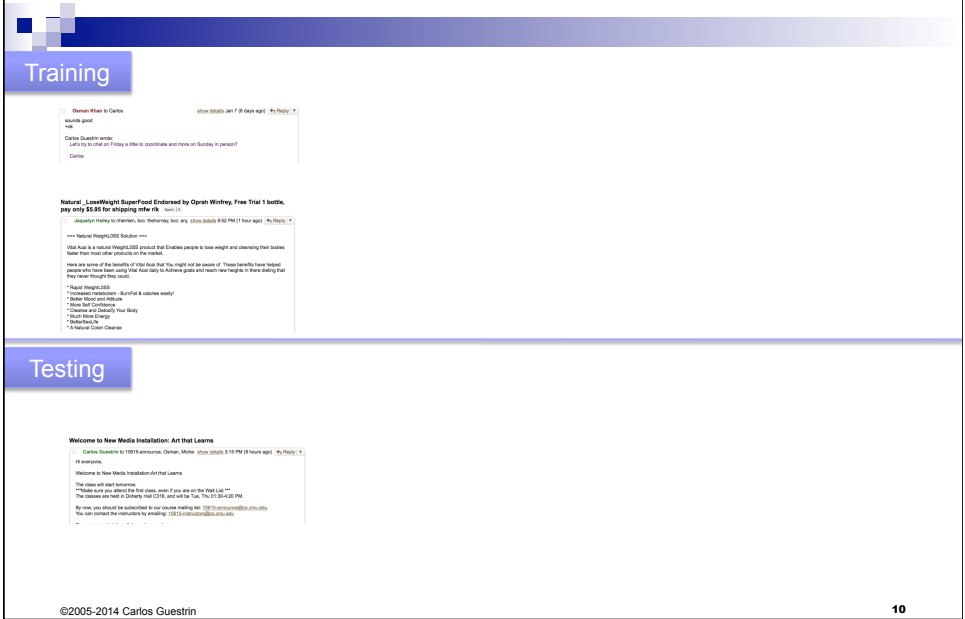
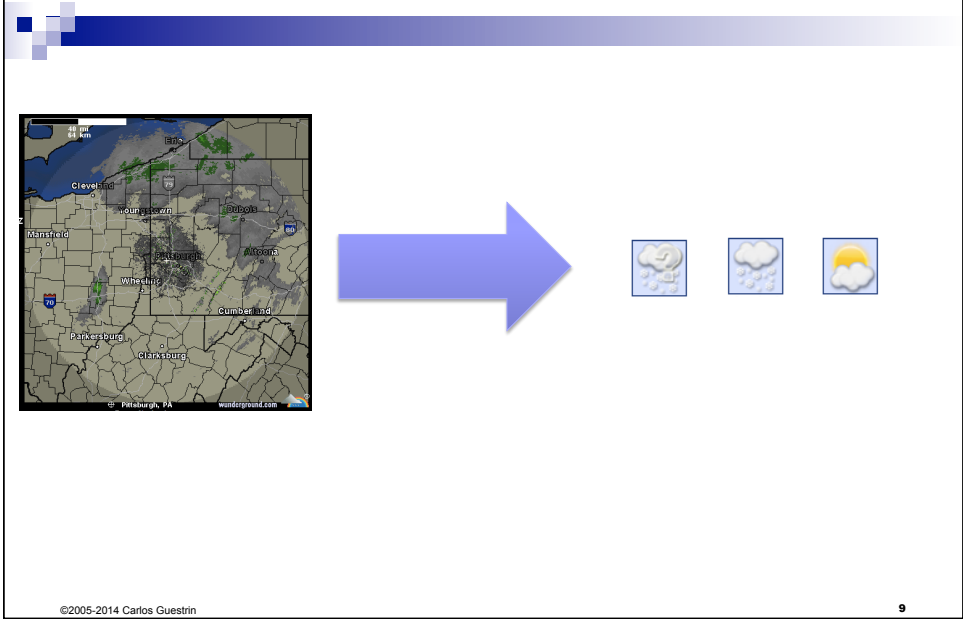


©2005-2014 Carlos Guestrin

7



8



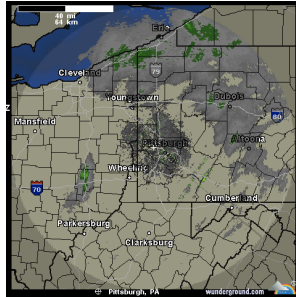
Regression

predicting a numeric value

Stock market



Weather prediction revisited

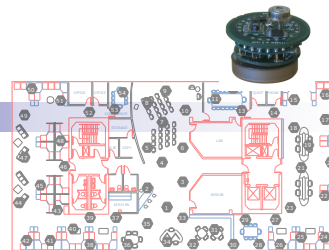
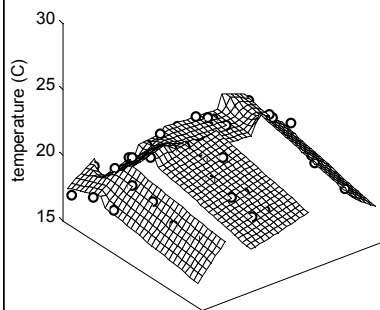


Temperature

©2005-2014 Carlos Guestrin

13

Modeling sensor data



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

[Guestrin et al. '04]

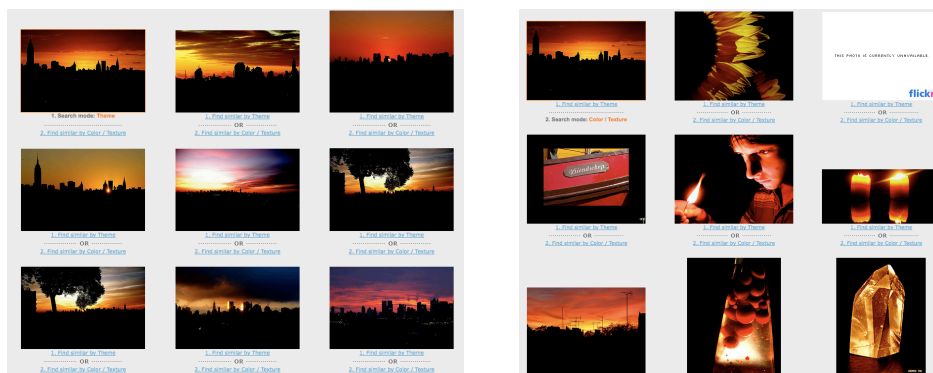
©2005-2014 Carlos Guestrin

14

Similarity

finding data

Given image, find similar images



Similar products



Processing: A Programming Handbook for Visual Designers and Artists (Hardcover)

by Casey Reas (Author), Ben Fry (Author), John Maeda (Foreword)

★★★★★ (13 customer reviews)

Available from these sellers.

31 new from \$47.95 8 used from \$43.56

Get Free Two-Day Shipping

Get Free Two-Day Shipping for three months with a special extended free trial of Amazon Prime™. Add this eligible textbook to your cart to qualify. Sign up at checkout. [See details.](#)

[See larger image](#)

[Share your own customer images](#)

[Publisher: learn how customers can search](#)

[Book this book](#)

Please tell the publisher:

I'd like to read this book on Kindle

Don't have a Kindle? [Get yours here](#)

Related Education & Training Services in Pittsburgh [\(show map\)](#) | [change location](#)

[Learn HTML Coding](#)

[www.FullSail.edu](#) • Earn Your Bachelor's Degree in Web Design and Development.

[Create Websites with HTML](#)

[http://www.unix.Berkeley.edu](#) • Learn HTML Online, Start Anytime! with UC Berkeley Extension

[Intensive XML Training](#)

[www.objectdatabs.com/course10.asp](#) • OnSite or in NYC, LA, SFO, ORD, DC Will customize & train as few as 3

Customers Who Bought This Item Also Bought



Processing: Creative Coding and Computational A... by Ira Greenberg
★★★★★ (7) \$43.99



Visualizing Data: Exploring and Explaining Data... by Ben Fry
★★★★★ (11) \$26.39



Making Things Talk: Practical Methods for Control... by Tom Igoe
★★★★★ (11) \$19.79



Physical Computing: Sensing and Controlling the... by Tom Igoe
★★★★★ (20) \$19.00



Learning Processing: A Beginner's Guide to... by Daniel Shiffman
★★★★★ (7) \$44.05

©2005-2014 Carlos Guestrin

17

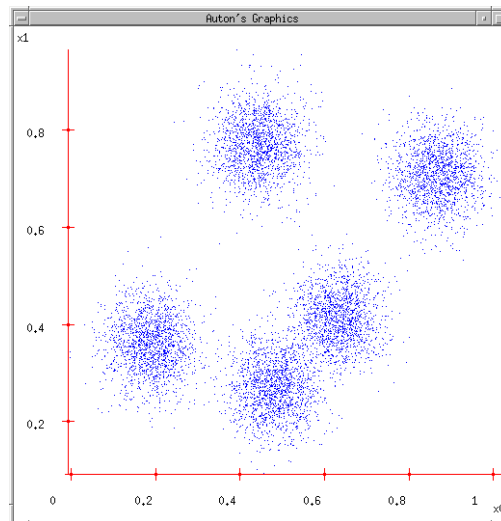
Clustering

discovering structure in data

©2005-2014 Carlos Guestrin

18

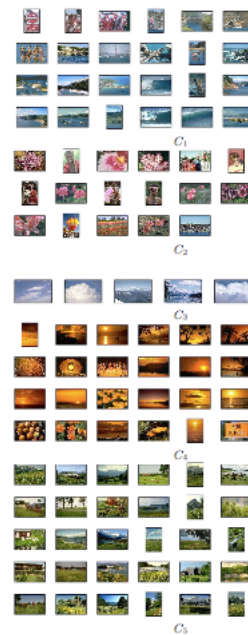
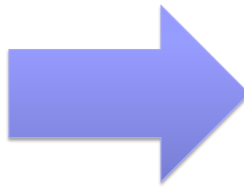
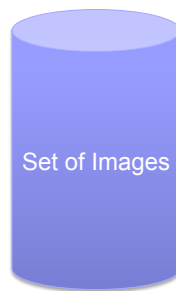
Clustering Data: Group similar things



©2005-2014 Carlos Guestrin

19

Clustering images



©2005-2014 Carlos Guestrin

[Goldberger et al.]₂₀

Clustering web search results

The screenshot shows the Clusty web search interface. At the top, there's a navigation bar with links for web, news, images, wikipedia, blogs, jobs, and more. Below this is a search bar with the query 'race' and a search button. To the left of the search results, there's a sidebar with a 'clusters' section. This section lists various clusters related to the search term, such as 'Car (28)', 'Race cars (7)', 'Photos, Races Scheduled (5)', 'Game (4)', 'Track (3)', 'Nascar (2)', 'Equipment And Safety (2)', 'Other Topics (7)', 'Photos (22)', 'Game (14)', 'Definition (13)', 'Team (18)', 'Human (8)', 'Classification Of Human (2)', 'Statement, Evolved (2)', 'Other Topics (4)', 'Weekend (8)', 'Ethnicity And Race (7)', 'Race for the Cure (8)', and 'Race Information (8)'. Below the clusters list is a 'Find in clusters' search bar. The main search results area on the right shows a list of 8 documents. The first document is 'Race (classification of human beings) - Wikipedia, the free encyclopedia'. The second document is 'Race - Wikipedia, the free encyclopedia'. The third document is 'Publications | Human Rights Watch'. The fourth document is 'Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich, Frank Miele: Books ...'. The fifth document is 'AAPA Statement on Biological Aspects of Race ...'. The sixth document is 'race: Definition from Answers.com'. The seventh document is 'Dopefish.com'. The eighth document is 'Dopefish.com'. The footer of the page shows the copyright information: '©2005-2014 Carlos Guestrin' and the page number '21'.

web news images wikipedia blogs jobs more »

Clusty

race

Search advanced preferences

clusters sources sites remix

All Results (238)

- Car (28)
- Race cars (7)
- Photos, Races Scheduled (5)
- Game (4)
- Track (3)
- Nascar (2)
- Equipment And Safety (2)
- Other Topics (7)
- Photos (22)
- Game (14)
- Definition (13)
- Team (18)
- Human (8)
 - Classification Of Human (2)
 - Statement, Evolved (2)
 - Other Topics (4)
- Weekend (8)
- Ethnicity And Race (7)
- Race for the Cure (8)
- Race Information (8)

more | all clusters

Find in clusters Find

Cluster Human contains 8 documents.

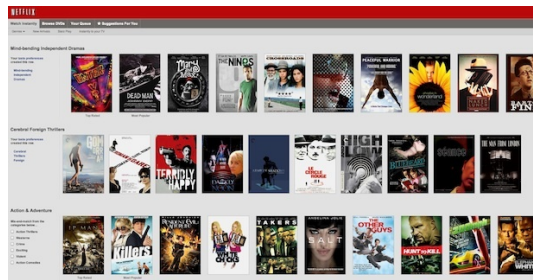
- Race (classification of human beings) - Wikipedia, the free encyclopedia**
The term **race** or racial group usually refers to the concept of dividing **humans** into populations or groups on the basis of various sets of characteristics. The most widely used **human** racial categories are based on visible traits (especially skin color, cranial or facial features and hair texture), and self-identification. Conceptions of **race**, as well as specific ways of grouping **rac**es, vary by culture and over time, and are often controversial for scientific as well as social and political reasons.**History** · **Modern debates** · **Political and ...**
en.wikipedia.org/wiki/Race_(classification_of_human_beings) · [cache] · Live, Ask
- Race - Wikipedia, the free encyclopedia**
General. **Racing** competitions The **Race** (yachting **race**), or La course du millénaire, a no-rules round-the-world sailing event; **Race** (biology); classification of flora and fauna; **Race** (classification of human beings) **Race** and ethnicity in the United States Census, official definitions of "race" used by the US Census Bureau; **Race** and genetics, notion of racial classifications based on genetics. Historical definitions of **race**; **Race** (bearing), the inner and outer rings of a rolling-element bearing. **RACE** in molecular biology "Rapid ... General · Surnames · Television · Music · Literature · Video games
en.wikipedia.org/wiki/Race · [cache] · Live, Ask
- Publications | Human Rights Watch**
The use of torture, unlawful rendition, secret prisons, unfair trials, ... Risks to Migrants, Refugees, and Asylum Seekers in Egypt and Israel ... In the run-up to the Beijing Olympics in August 2008, ...
www.hrw.org/backgrounder/usa/race · [cache] · Ask
- Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich, Frank Miele: Books ...**
www.amazon.com/Race-Reality-Differences-Vincent-Sarich/dp/0813340861 · [cache] · Live
- AAPA Statement on Biological Aspects of Race**
AAPA Statement on Biological Aspects of Race ... Published in the American Journal of Physical Anthropology, vol. 101, pp 569-570, 1996 ... PREAMBLE As scientists who study human evolution and variation, ...
www.physanth.org/positions/race.html · [cache] · Ask
- race: Definition from Answers.com**
race n. A local geographic or global human population distinguished as a more or less distinct group by genetically transmitted physical
www.answers.com/topic/race-1 · [cache] · Live
- Dopefish.com**
Site for newbies as well as experienced Dopefish followers, chronicling the birth of the Dopefish, its numerous appearances in several computer games, and its eventual take-over of the human **race**. Maintained by Mr. Dopefish himself, Joe Siegler of Apogee Software.
www.dopefish.com · [cache] · Open Directory

Recommender Systems

figuring out what your customers want

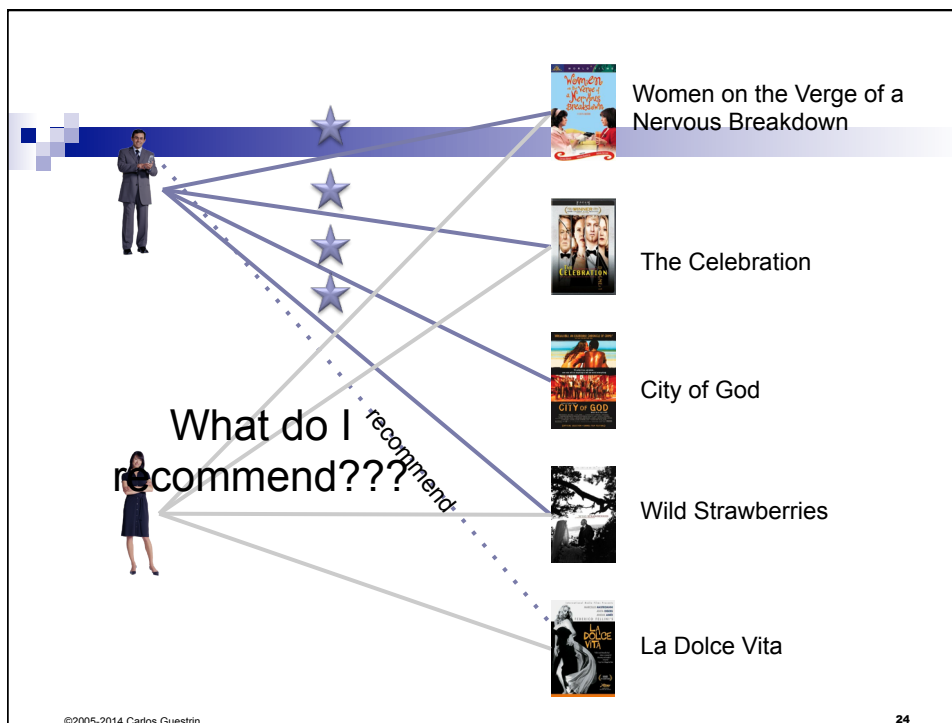
Collaborative Filtering

- **Goal:** Find movies of interest to a user based on movies watched by the user and others
- **Methods:** matrix factorization



©2005-2014 Carlos Guestrin

23



©2005-2014 Carlos Guestrin

24

Embedding

visualizing data

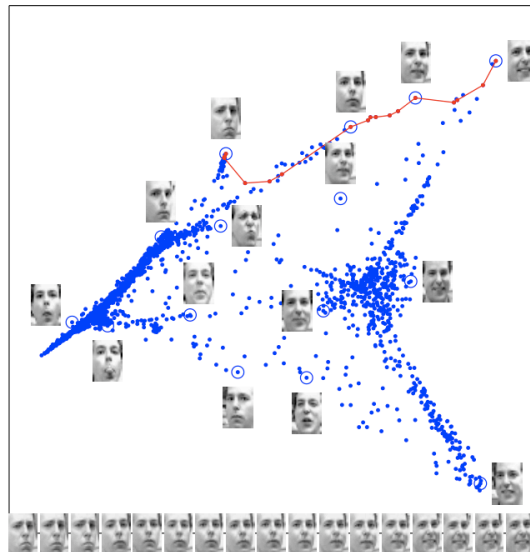
©2005-2014 Carlos Guestrin

25

Embedding images

Images have thousands or millions of pixels.

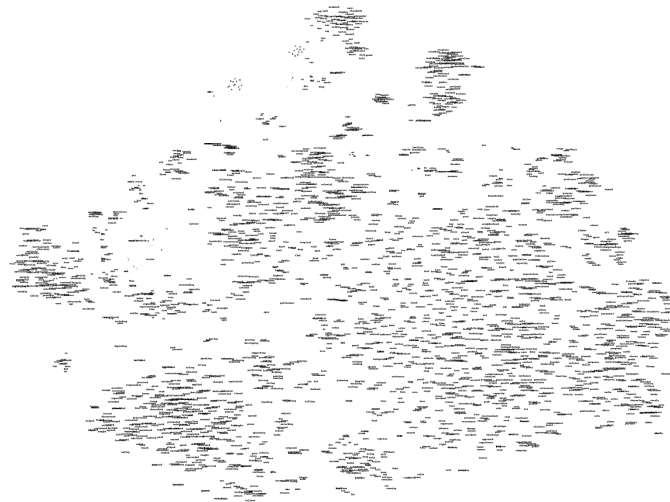
Can we give each image a coordinate, such that similar images are near each other?



©2005-2014 Carlos Guestrin

[Saul & Roweis '03] 26

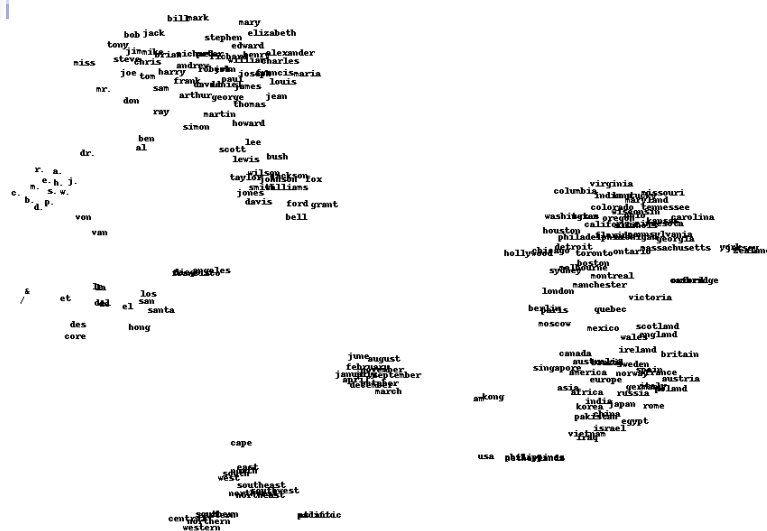
Embedding words



©2005-2014 Carlos Guestrin

[Joseph Turian] 27

Embedding words (zoom in)



©2005-2014 Carlos Guestrin

[Joseph Turian] 28

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...

©2005-2014 Carlos Guestrin

31

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.)

SAYID

SUN

locke

HOLDING

Taskar et al.

©2005-2014 Carlos Guestrin

Automatically Discovered and Labeled Actions

■ **shout**
(JACK) (shouts) ()

smile
(Kate) (smiles) ()

follow
(Kate) (follows) (Jack)

sit down
(Locke) (sits down) ()

wake
(Sawyer) (wakes up) ()

swim
(Sawyer) (turns) (swimming)

grab
(Kate) (grabs) (case)

kiss
(Shannon) (kisses) (ear)

open door
(door) (opens) ()

point
(JACK) (points) ()

©2005-2014 Carlos Guestrin

Growth of Machine Learning

One of the most sought for specialties in industry!!!!

- 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, recommender systems, matrix factorization, SVMs, kernels, margin bounds, K-means, EM, mixture models, semi-supervised learning, neural networks, reinforcement learning...
- Covers algorithms, theory and applications
- **It's going to be fun and hard work 😊**

©2005-2014 Carlos Guestrin

35

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
 - Python will be very useful
- We provide some background, but the class will be fast paced
- Ability to deal with “abstract mathematical concepts”

©2005-2014 Carlos Guestrin

36

Staff

- Two Great TAs: Great resource for learning, interact with them!

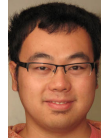
- **Akshay Srinivasan**

- Office hours: Wednesdays 4:30-6:30pm



- **TianyiZhou**

- Office hours: Tuesdays 4:30-6:30pm



- Prof: **Carlos Guestrin**

- Office hours: Mondays 5:30-6:30pm

©2005-2014 Carlos Guestrin

37

Communication Channels

- Only channel for announcements, questions, etc. – Catalyst Group:

- <https://catalyst.uw.edu/gopost/board/tianyizh/35317/>

- 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:

- csep546-instructors@cs.washington.edu

©2005-2014 Carlos Guestrin

38

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 (70%)**
 - First one goes out this week
 - 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 next week
 - Projects done individually, or groups of two students

Homeworks

- Homeworks are hard, start early ☺
- Due in the beginning of class
- 33% subtracted per late day
- 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

- 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 week or so
- Mon., January 27 by 6:30pm: **Project Proposals**
- Mon., February 24 by 6:30pm: **Project Milestone**
- Mon., March 17 by 6:30pm: **Poster Session**
- Mon., March 19 by 6:30pm: **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...

Point Estimation MLE

Machine Learning – CSEP546
Carlos Guestrin
University of Washington
January 6, 2014

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...

©2005-2014 Carlos Guestrin

45

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 α_H Heads and α_T Tails
$$P(D \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

©2005-2014 Carlos Guestrin

46

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:

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(\mathcal{D} | \theta) \\ &= \arg \max_{\theta} \ln P(\mathcal{D} | \theta)\end{aligned}$$

©2005-2014 Carlos Guestrin

47

Your first learning algorithm

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

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

©2005-2014 Carlos Guestrin

48

What about continuous variables?

- Billionaire 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)
 - $X \sim N(\mu, \sigma^2)$
 - $Y = aX + b \rightarrow Y \sim N(a\mu + b, a^2\sigma^2)$
- Sum of Gaussians
 - $X \sim N(\mu_X, \sigma_X^2)$
 - $Y \sim N(\mu_Y, \sigma_Y^2)$
 - $Z = X + Y \rightarrow Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

Learning a Gaussian

- Collect a bunch of data

- Hopefully, i.i.d. samples
- e.g., exam scores

- Learn parameters

- Mean
- Variance

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

©2005-2014 Carlos Guestrin

51

MLE for Gaussian

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

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

- Log-likelihood of data:

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

©2005-2014 Carlos Guestrin

52

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:

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

Learning Gaussian parameters

- MLE:

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

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

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

- ☐ Expected result of estimation is **not** true parameter!

- ☐ Unbiased variance estimator:

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

©2005-2014 Carlos Guestrin

55

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...

©2005-2014 Carlos Guestrin

56



Linear Regression

Machine Learning – CSEP546

Carlos Guestrin

University of Washington

January 6, 2014

©2005-2014 Carlos Guestrin

57

Prediction of continuous variables



- Billionaire sayz: Wait, that's not what I meant!
- You sayz: Chill out, dude.
- He sayz: I want to predict a continuous variable for continuous inputs: I want to predict salaries from GPA.
- You sayz: **I can regress that...**

©2005-2014 Carlos Guestrin

58

The regression problem

- **Instances:** $\langle \mathbf{x}_j, t_j \rangle$
- **Learn:** Mapping from \mathbf{x} to $t(\mathbf{x})$
- **Hypothesis space:** $H = \{h_1, \dots, h_K\}$
 - Given, basis functions
 - Find coeffs $\mathbf{w} = \{w_1, \dots, w_K\}$ $\underbrace{t(\mathbf{x})}_{\text{data}} \approx \hat{f}(\mathbf{x}) = \sum_i w_i h_i(\mathbf{x})$
 - Why is this called linear regression???
 - model is linear in the parameters
- Precisely, minimize the **residual squared error**:

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

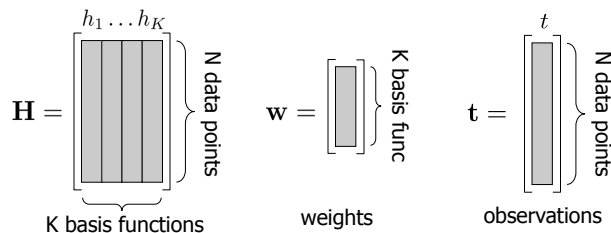
©2005-2014 Carlos Guestrin

59

The regression problem in matrix notation

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \underbrace{(\mathbf{H}\mathbf{w} - \mathbf{t})^T (\mathbf{H}\mathbf{w} - \mathbf{t})}_{\text{residual error}}$$



©2005-2014 Carlos Guestrin

60

Minimizing the Residual

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \underbrace{(\mathbf{H}\mathbf{w} - \mathbf{t})^T (\mathbf{H}\mathbf{w} - \mathbf{t})}_{\text{residual error}}$$

©2005-2014 Carlos Guestrin

61

Regression solution = simple matrix operations

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \underbrace{(\mathbf{H}\mathbf{w} - \mathbf{t})^T (\mathbf{H}\mathbf{w} - \mathbf{t})}_{\text{residual error}}$$

$$\text{solution: } \mathbf{w}^* = \underbrace{(\mathbf{H}^T \mathbf{H})^{-1}}_{\mathbf{A}^{-1}} \underbrace{\mathbf{H}^T \mathbf{t}}_{\mathbf{b}} = \mathbf{A}^{-1} \mathbf{b}$$

$$\text{where } \mathbf{A} = \mathbf{H}^T \mathbf{H} = \begin{bmatrix} \square & \square & \square \\ \square & \square & \square \\ \square & \square & \square \end{bmatrix} \quad \mathbf{b} = \mathbf{H}^T \mathbf{t} = \begin{bmatrix} \square \\ \square \\ \square \end{bmatrix}$$

$k \times k$ matrix
for k basis functions
 $k \times 1$ vector

©2005-2014 Carlos Guestrin

62

But, why?

- Billionaire (again) says: Why sum squared error???
- You say: Gaussians, Dr. Gateson, Gaussians...
- Model: prediction is linear function plus Gaussian noise
 - $t(\mathbf{x}) = \sum_i w_i h_i(\mathbf{x}) + \varepsilon_{\mathbf{x}}$

- Learn \mathbf{w} using MLE

$$P(t \mid \mathbf{x}, \mathbf{w}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{[t - \sum_i w_i h_i(\mathbf{x})]^2}{2\sigma^2}}$$

©2005-2014 Carlos Guestrin

63

Maximizing log-likelihood

Maximize:

$$\ln P(\mathcal{D} \mid \mathbf{w}, \sigma) = \ln \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^N \prod_{j=1}^N e^{-\frac{[t_j - \sum_i w_i h_i(\mathbf{x}_j)]^2}{2\sigma^2}}$$

Least-squares Linear Regression is MLE for Gaussians!!!

©2005-2014 Carlos Guestrin

64



Bias-Variance Tradeoff

Machine Learning – CSEP546

Carlos Guestrin

University of Washington

January 6, 2014

©2005-2014 Carlos Guestrin

65

Bias-Variance tradeoff – Intuition



- Model too “simple” → does not fit the data well
 - A biased solution
- Model too complex → small changes to the data, solution changes a lot
 - A high-variance solution

©2005-2014 Carlos Guestrin

66

(Squared) Bias of learner

- Given dataset D with N samples, learn function $h_D(x)$
- If you sample a different dataset D' with N samples, you will learn different $h_{D'}(x)$
- **Expected hypothesis:** $E_D[h_D(x)]$
- **Bias:** difference between what you expect to learn and truth
 - Measures how well you expect to represent true solution
 - Decreases with more complex model
 - Bias² at one point x :
 - Average Bias²:

©2005-2014 Carlos Guestrin

67

Variance of learner

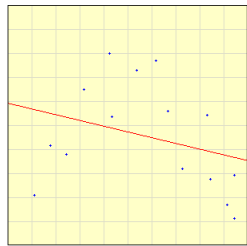
- Given dataset D with N samples, learn function $h_D(x)$
- If you sample a different dataset D' with N samples, you will learn different $h_{D'}(x)$
- **Variance:** difference between what you expect to learn and what you learn from a particular dataset
 - Measures how sensitive learner is to specific dataset
 - Decreases with simpler model
 - Variance at one point x :
 - Average variance:

©2005-2014 Carlos Guestrin

68

Bias-Variance Tradeoff

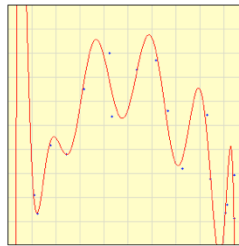
- Choice of hypothesis class introduces learning bias
 - More complex class → less bias
 - More complex class → more variance



Select points by clicking on the graph or press [Example](#)

Degree of polynomial: ☒ Fit Y to X
☐ Fit X to Y

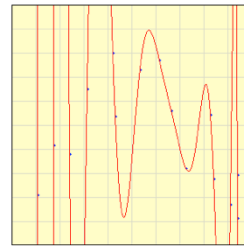
[Calculate](#) [View Polynomial](#) [Reset](#)



Select points by clicking on the graph or press [Example](#)

Degree of polynomial: ☒ Fit Y to X
☐ Fit X to Y

[Calculate](#) [View Polynomial](#) [Reset](#)



Select points by clicking on the graph or press [Example](#)

Degree of polynomial: ☒ Fit Y to X
☐ Fit X to Y

[Calculate](#) [View Polynomial](#) [Reset](#)

©2005-2014 Carlos Guestrin

69

Bias-Variance Decomposition of Error

$$\bar{h}_N(x) = E_D[h_D(x)]$$

- Expected mean squared error: $\text{MSE} = E_D \left[E_x \left[(t(x) - h_D(x))^2 \right] \right]$
- To simplify derivation, drop x :
- Expanding the square:

©2005-2014 Carlos Guestrin

70

Moral of the Story: Bias-Variance Tradeoff Key in ML

- Error can be decomposed:

$$\begin{aligned}\text{MSE} &= E_D \left[E_x \left[(t(x) - h_D(x))^2 \right] \right] \\ &= E_x \left[(t(x) - \bar{h}_N(x))^2 \right] + E_D \left[E_x \left[(\bar{h}(x) - h_D(x))^2 \right] \right]\end{aligned}$$

- Choice of hypothesis class introduces learning bias

- More complex class → less bias
- More complex class → more variance

©2005-2014 Carlos Guestrin

71

What you need to know

- Regression

- Basis function = features
- Optimizing sum squared error
- Relationship between regression and Gaussians

- Bias-variance trade-off

- Play with Applet

©2005-2014 Carlos Guestrin

72

Overfitting

Machine Learning – CSE446

Carlos Guestrin

University of Washington

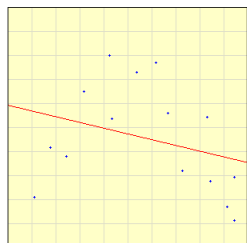
January 6, 2014

©2005-2014 Carlos Guestrin

73

Bias-Variance Tradeoff

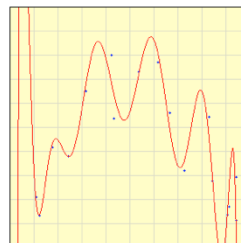
- Choice of hypothesis class introduces learning bias
 - More complex class → less bias
 - More complex class → more variance



Select points by clicking on the graph or press [Example](#)

Degree of polynomial: ☒ Fit Y to X
☐ Fit X to Y

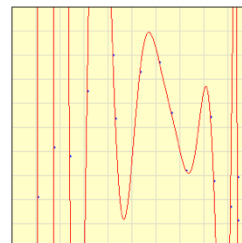
[Calculate](#) [View Polynomial](#) [Reset](#)



Select points by clicking on the graph or press [Example](#)

Degree of polynomial: ☒ Fit Y to X
☐ Fit X to Y

[Calculate](#) [View Polynomial](#) [Reset](#)



Select points by clicking on the graph or press [Example](#)

Degree of polynomial: ☒ Fit Y to X
☐ Fit X to Y

[Calculate](#) [View Polynomial](#) [Reset](#)

©2005-2014 Carlos Guestrin

74

Training set error

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

- Given a dataset (Training data)
- Choose a loss function
 - e.g., squared error (L_2) for regression
- **Training set error:** For a particular set of parameters, loss function on training data:

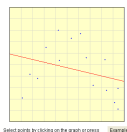
$$error_{train}(\mathbf{w}) = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

©2005-2014 Carlos Guestrin

75

Training set error as a function of model complexity

$$error_{train}(\mathbf{w}) = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$



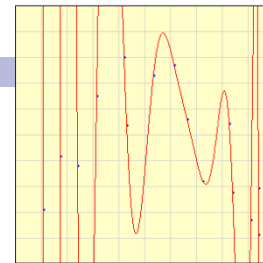
Degree of polynomial: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000

©2005-2014 Carlos Guestrin

76

Prediction error

- Training set error can be poor measure of “quality” of solution
- **Prediction error:** We really care about error over all possible input points, not just training data:



Select points by clicking on the graph or press [Example](#)

Degree of polynomial: ☒ Fit Y to X ☐ Fit X to Y

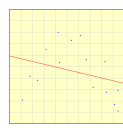
$$\begin{aligned} error_{true}(\mathbf{w}) &= E_{\mathbf{x}} \left[\left(t(\mathbf{x}) - \sum_i w_i h_i(\mathbf{x}) \right)^2 \right] \\ &= \int_{\mathbf{x}} \left(t(\mathbf{x}) - \sum_i w_i h_i(\mathbf{x}) \right)^2 p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

©2005-2014 Carlos Guestrin

77

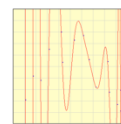
Prediction error as a function of model complexity: Bias/Variance tradeoff

$$error_{true}(\mathbf{w}) = \int_{\mathbf{x}} \left(t(\mathbf{x}) - \sum_i w_i h_i(\mathbf{x}) \right)^2 p(\mathbf{x}) d\mathbf{x}$$



Select points by clicking on the graph or press [Example](#)

Degree of polynomial: ☒ Fit Y to X ☐ Fit X to Y



Select points by clicking on the graph or press [Example](#)

Degree of polynomial: ☒ Fit Y to X ☐ Fit X to Y

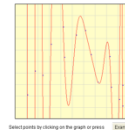
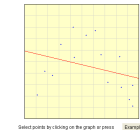
©2005-2014 Carlos Guestrin

78

Prediction error as a function of model complexity: train v. true error

$$error_{train}(\mathbf{w}) = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

$$error_{true}(\mathbf{w}) = \int_{\mathbf{x}} \left(t(\mathbf{x}) - \sum_i w_i h_i(\mathbf{x}) \right)^2 p(\mathbf{x}) d\mathbf{x}$$



©2005-2014 Carlos Guestrin

79

Computing prediction error

■ Computing prediction

- ☐ Hard integral
- ☐ May not know $t(\mathbf{x})$ for every \mathbf{x}

$$error_{true}(\mathbf{w}) = \int_{\mathbf{x}} \left(t(\mathbf{x}) - \sum_i w_i h_i(\mathbf{x}) \right)^2 p(\mathbf{x}) d\mathbf{x}$$

■ Monte Carlo integration (sampling approximation)

- ☐ Sample a set of i.i.d. points $\{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ from $p(\mathbf{x})$
- ☐ Approximate integral with sample average

$$error_{true}(\mathbf{w}) \approx \frac{1}{M} \sum_{j=1}^M \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

©2005-2014 Carlos Guestrin

80

Why training set error doesn't approximate prediction error?

- Sampling approximation of prediction error:

$$error_{true}(\mathbf{w}) \approx \frac{1}{M} \sum_{j=1}^M \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

- Training error :

$$error_{train}(\mathbf{w}) = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

- Very similar equations!!!

- ☐ Why is training set a bad measure of prediction error???

©2005-2014 Carlos Guestrin

81

Why training set error doesn't approximate prediction error?

-

Because you cheated!!!

Training error good estimate for a single \mathbf{w} ,
But you optimized \mathbf{w} with respect to the training error,
and found \mathbf{w} that is good for this set of samples

-

**Training error is a (optimistically) biased
estimate of prediction error**

- Very similar equations!!!

- ☐ Why is training set a bad measure of prediction error???

©2005-2014 Carlos Guestrin

82

Test set error

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

- Given a dataset, **randomly** split it into two parts:
 - Training data – $\{\mathbf{x}_1, \dots, \mathbf{x}_{N_{\text{train}}}\}$
 - Test data – $\{\mathbf{x}_1, \dots, \mathbf{x}_{N_{\text{test}}}\}$
- Use training data to optimize parameters \mathbf{w}
- **Test set error:** For the **final output** $\hat{\mathbf{w}}$, evaluate the error using:

$$error_{test}(\mathbf{w}) = \frac{1}{N_{test}} \sum_{j=1}^{N_{test}} \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

©2005-2014 Carlos Guestrin

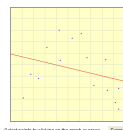
83

Test set error as a function of model complexity

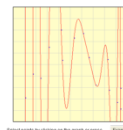
$$error_{train}(\mathbf{w}) = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

$$error_{true}(\mathbf{w}) = \int_{\mathbf{x}} \left(t(\mathbf{x}) - \sum_i w_i h_i(\mathbf{x}) \right)^2 p(\mathbf{x}) d\mathbf{x}$$

$$error_{test}(\mathbf{w}) = \frac{1}{N_{test}} \sum_{j=1}^{N_{test}} \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$



Graph of polynomial $f(x) = 0.5x^2 + 0.5x + 0.5$
[Zoom](#) [Reset](#) [Help](#)



Graph of polynomial $f(x) = 0.5x^2 + 0.5x + 0.5$
[Zoom](#) [Reset](#) [Help](#)

©2005-2014 Carlos Guestrin

84

Overfitting

- **Overfitting:** a learning algorithm overfits the training data if it outputs a solution \mathbf{w} when there exists another solution \mathbf{w}' such that:

$$[error_{train}(\mathbf{w}) < error_{train}(\mathbf{w}')] \wedge [error_{true}(\mathbf{w}') < error_{true}(\mathbf{w})]$$

How many points to I use for training/testing?

- Very hard question to answer!
 - Too few training points, learned \mathbf{w} is bad
 - Too few test points, you never know if you reached a good solution
 - Some theoretical bounds can be useful in theory... ☺
- More on this later this quarter, but still hard to answer
- Typically:
 - If you have a reasonable amount of data, pick test set “large enough” for a “reasonable” estimate of error, and use the rest for learning
 - If you have little data, then you need to pull out the big guns...
 - e.g., bootstrapping

Error estimators

$$error_{true}(\mathbf{w}) = \int_{\mathbf{x}} \left(t(\mathbf{x}) - \sum_i w_i h_i(\mathbf{x}) \right)^2 p(\mathbf{x}) d\mathbf{x}$$

$$error_{train}(\mathbf{w}) = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

$$error_{test}(\mathbf{w}) = \frac{1}{N_{test}} \sum_{j=1}^{N_{test}} \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

©2005-2014 Carlos Guestrin

87

Error as a function of number of training examples for a fixed model complexity



$$error_{train}(\mathbf{w}) = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

$$error_{true}(\mathbf{w}) = \int_{\mathbf{x}} \left(t(\mathbf{x}) - \sum_i w_i h_i(\mathbf{x}) \right)^2 p(\mathbf{x}) d\mathbf{x}$$

little data

infinite data

©2005-2014 Carlos Guestrin

88

Error estimators

Be careful!!!

Test set only unbiased if you never never ever ever
do any any any any learning on the test data

For example, if you use the test set to select
the degree of the polynomial... no longer unbiased!!!
(We will address this problem later in the quarter)

$$error_{test}(\mathbf{w}) = \frac{1}{N_{test}} \sum_{j=1}^{N_{test}} \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2$$

What you need to know

- True error, training error, test error
 - ☐ Never learn on the test data
 - ☐ Never learn on the test data
 - ☐ Never learn on the test data
 - ☐ Never learn on the test data
 - ☐ Never learn on the test data
- Overfitting