

# Learning Theory Continued...

Machine Learning – CSE446

Carlos Guestrin

University of Washington

May 13, 2013

©Carlos Guestrin 2005-2013

1

## A simple setting...

- Classification
  - N data points *iid*
  - Finite number of possible hypothesis (e.g., dec. trees of depth d)
- A learner finds a hypothesis  $h$  that is consistent with training data
  - Gets zero error in training –  $\text{error}_{\text{train}}(h) = 0$
- What is the probability that  $h$  has more than  $\varepsilon$  true error?
  - $\text{error}_{\text{true}}(h) \geq \varepsilon$  *← large for  $\varepsilon > 0$  given*

©Carlos Guestrin 2005-2013

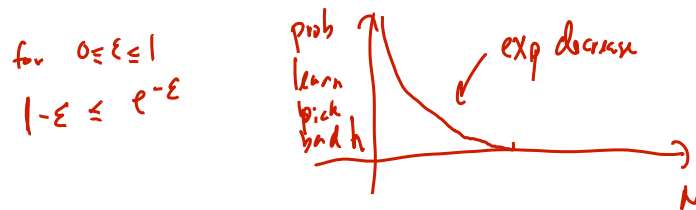
2

## Generalization error in finite hypothesis spaces [Haussler '88]

- **Theorem:** Hypothesis space  $H$  finite, dataset  $D$  with  $N$  i.i.d. samples,  $0 < \epsilon < 1$  : for any learned hypothesis  $h$  that is consistent on the training data:

$$P(\text{error}_{\text{true}}(h) > \epsilon) \leq |H|e^{-N\epsilon}$$

$$\leq |H| (1-\epsilon)^N \leq |H| (e^{-\epsilon})^N = |H| e^{-N\epsilon}$$



©Carlos Guestrin 2005-2013

3

## Limitations of Haussler '88 bound

$$P(\text{error}_{\text{true}}(h) > \epsilon) \leq |H|e^{-N\epsilon}$$

- Consistent classifier

↑  
 $\text{error}_{\text{train}}(h) = 0$   
 highly unrealistic, and bad w.r.t overfitting

- Size of hypothesis space

$\ln|H|$ , bad is  $H$  is continuous (infinite)  
 or  $H$  is very very large

©Carlos Guestrin 2005-2013

4

## What if our classifier does not have zero error on the training data?

- A learner with **zero** training errors may make mistakes in test set
- What about a learner with  $\text{error}_{\text{train}}(h)$  in training set?

is  $\text{error}_{\text{train}}(h) > 0$

what about  $\text{error}_{\text{true}}(h)$ ?

in Logistic Regression,  
there are infinitely  
many  $h$ ,  
parameterized by  
 $w$

©Carlos Guestrin 2005-2013

5

## Generalization bound for $|H|$ hypothesis

- **Theorem:** Hypothesis space  $H$  finite, dataset  $D$  with  $N$  i.i.d. samples,  $0 < \epsilon < 1$  : for any learned hypothesis  $h$ :

$$P(\text{error}_{\text{true}}(h_i) - \text{error}_{\text{train}}(h_i) > \epsilon) \leq e^{-2N\epsilon^2} \leq \delta$$

hold  $\forall h_i$

$$P(\text{error}_{\text{true}}(h) - \text{error}_{\text{train}}(h) > \epsilon) \leq |H| e^{-2N\epsilon^2}$$

with prob. at least  $1 - \delta$

$$\epsilon \geq \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2N}} \quad \left\{ \begin{array}{l} \text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2N}} \end{array} \right.$$

©Carlos Guestrin 2005-2013

6

## PAC bound and Bias-Variance tradeoff

PAC Bound  $\rightarrow$   $P(\text{error}_{\text{true}}(h) - \text{error}_{\text{train}}(h) > \epsilon) \leq e^{-2N\epsilon^2}$

or, after moving some terms around,  
with probability at least  $1-\delta$ :

Want to be small  $\rightarrow$

$$\text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2N}}$$

ITS of depth 1,600,000	'complex' hypothesis space	Small	$\ln  H $ large $\Rightarrow$ large
ITS of depth 4	'simple' hypothesis space	larger "bias"	Smaller, because $ H $ smaller "variance"

- Important: PAC bound holds for all  $h$ , but doesn't guarantee that algorithm finds best  $h$ !!!

©Carlos Guestrin 2005-2013

7

## What about the size of the hypothesis space?

$$N \geq \frac{\ln |H| + \ln \frac{1}{\delta}}{2\epsilon^2}$$

- How large is the hypothesis space?

How big is  $|H|$ ?

really big

Versus

really really big

©Carlos Guestrin 2005-2013

8

## Boolean formulas with $m$ binary features

$$x_1 \wedge \neg x_2 \vee x_2 \wedge x_3 \wedge \neg x_4 \dots$$

$$N \geq \frac{\ln |H| + \ln \frac{1}{\delta}}{2\epsilon^2}$$

H: any boolean formula

H: all conjunctions of literals

$x_1, x_2, \dots, x_m$   
 $0 \ 0 \ 0 \dots 0$   
 $0 \ 0 \ 0 \ 0 \ 1$   
 $\vdots$   
 $1 \ 1 \dots 1$

$Y$   
 $0 \text{ or } 1$   
 $0 \text{ or } 1$   
 $\vdots$   
 $0 \text{ or } 1$

$|H| = 2^m$   
 $\uparrow$  (2 choices per row)  
 $\log |H| = 2^m$   
 $\Rightarrow$  need to see all rows to learn

$x_1 \wedge \neg x_3 \wedge x_2 \wedge x_4$   
 $|H| = 3^m \leftarrow$  really big  
 $\ln |H| = m \ln 3$   
 $\leftarrow$  much more hopeful

3 choices per literal  
 - absent  
 - positive  
 - negated

©Carlos Guestrin 2005-2013

9

## Number of decision trees of depth $k$

$$N \geq \frac{\ln |H| + \ln \frac{1}{\delta}}{2\epsilon^2}$$

Recursive solution

Given  $m$  attributes

$H_k$  = Number of decision trees of depth  $k$

$H_0 = 2$

$H_{k+1} = (\text{\#choices of root attribute}) * (\text{\#possible left subtrees}) * (\text{\#possible right subtrees})$

$$= m * H_k * H_k$$

Write  $L_k = \log_2 H_k$

$L_0 = 1$

$L_{k+1} = \log_2 m + 2L_k$

So  $L_k = (2^k - 1)(1 + \log_2 m) + 1$

simplifying

$$\ln |H| \leq 2^k \log m$$

really big in depth

ok in # features

©Carlos Guestrin 2005-2013

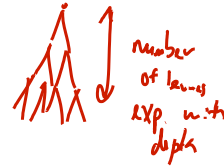
10

## PAC bound for decision trees of depth $k$

$$N \geq \frac{2^k \log m + \ln \frac{1}{\delta}}{\epsilon^2}$$

### ■ Bad!!!

- Number of points is exponential in depth!



### ■ But, for $N$ data points, decision tree can't get too big...

*no reason to expand to more than  $N$  leaves*

**Number of leaves never more than number data points**

©Carlos Guestrin 2005-2013

11

## Number of Decision Trees with $k$ Leaves

### ■ Number of decision trees of depth $k$ is really really big:

- $\ln |H|$  is about  $2^k \log m$

*← extra flexibility is unnecessary in the theory*

### ■ Decision trees with up to $k$ leaves:

- $|H|$  is about  $m^k k^{2k}$
- A very loose bound

$$\ln |H| \leq k \ln m + 2k \ln k$$

*← only really big*  
*much better!!*

©Carlos Guestrin 2005-2013

12

## PAC bound for decision trees with k leaves – Bias-Variance revisited

$$\ln |H_{\text{DTs } k \text{ leaves}}| \leq 2k(\ln m + \ln k) \quad \text{my bound } error_{true}(h) \leq 1$$

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2N}}$$

want to be small

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{2k(\ln m + \ln k) + \ln \frac{1}{\delta}}{2N}} \quad \leftarrow \# \text{ datapoints}$$

max number of leaves DT	"bias"	"variance"
$k \approx N$	goes to zero	LARGE greater than 1
$k \ll N$	potentially larger	small

©Carlos Guestrin 2005-2013

13

## What did we learn from decision trees?

- Bias-Variance tradeoff formalized

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{2k(\ln m + \ln k) + \ln \frac{1}{\delta}}{2N}}$$

- Moral of the story:

Complexity of learning not measured in terms of size hypothesis space, but in maximum number of points that allows consistent classification

- Complexity  $N$  – no bias, lots of variance
- Lower than  $N$  – some bias, less variance

↑ flexibility of the class

©Carlos Guestrin 2005-2013

14

# What about continuous hypothesis spaces?

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2N}}$$

## Continuous hypothesis space:

- ☐  $|H| = \infty$
- ☐ Infinite variance???

infinite possible  
 $|H| = \infty$   
 but they  
 all give  
 same  
 answer

## As with decision trees, only care about the maximum number of points that can be classified exactly!

- ☐ Called VC dimension... see readings for details

©Carlos Guestrin 2005-2013

15

# What you need to know

## Finite hypothesis space

- ☐ Derive results
- ☐ Counting number of hypothesis
- ☐ Mistakes on Training data

## Complexity of the classifier depends on number of points that can be classified exactly on training data

- ☐ Finite case – decision trees
- ☐ Infinite case – VC dimension

## Bias-Variance tradeoff in learning theory

## Remember: will your algorithm find best classifier?

2 is OK  
 3 is OK  
 4?  
 VC dimension  
 no line separates

©Carlos Guestrin 2005-2013

16



So for supervised learning :  $h: \mathcal{X} \rightarrow \mathcal{Y} \leftarrow \text{regression}$   
 $h: \mathcal{X} \rightarrow \{0, 1, \dots, k\} \leftarrow \text{classification}$

Unsupervised learning

## Clustering K-means

Machine Learning – CSE446

Carlos Guestrin

University of Washington

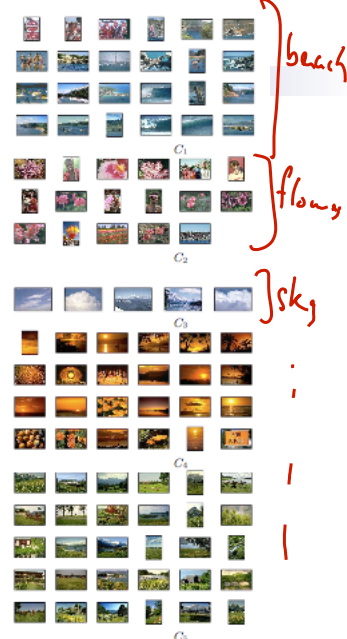
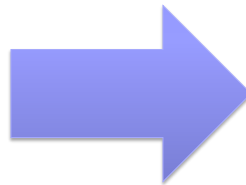
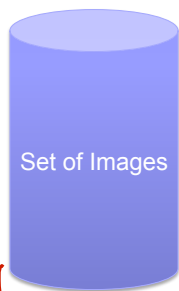
May 13, 2013

©Carlos Guestrin 2005-2013

17

## Clustering images

no  
labels  
given



©Carlos Guestrin 2005-2013

[Goldberger et al.]<sub>18</sub>

# Clustering web search results

The screenshot shows the Clusty search interface. The search term 'race' is entered in the top bar. On the left, a sidebar lists various clusters, with 'Human' selected. The main area displays a list of 8 documents related to the 'Human' cluster. A red arrow points to the 'Human' cluster in the sidebar, and another red arrow points to the first document in the list, 'Race (classification of human beings) - Wikipedia, the free encyclopedia'.

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 ...**
- 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
- 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, ...
- Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich, Frank Miele: Books ...**

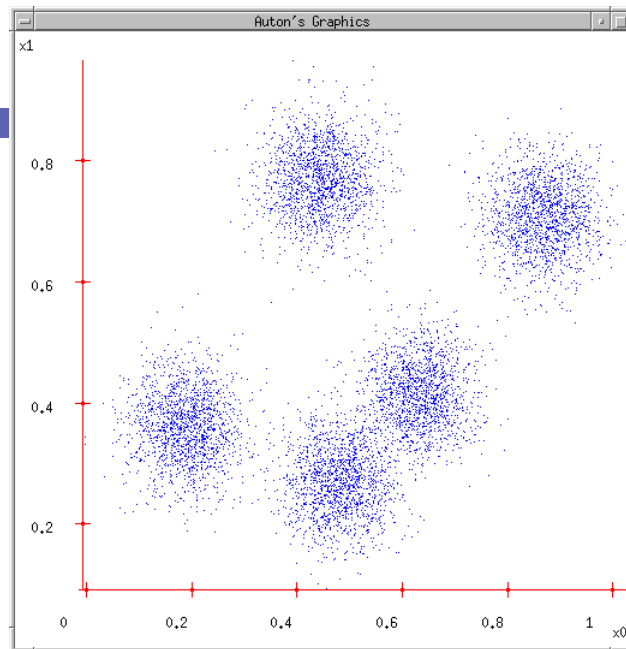
From Publishers Weekly Sarich, a Berkeley emeritus anthropologist, and Miele, an editor ...
- AAPA Statement on Biological Aspects of Race**

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, ...
- 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 ...
- Dopefish.com**

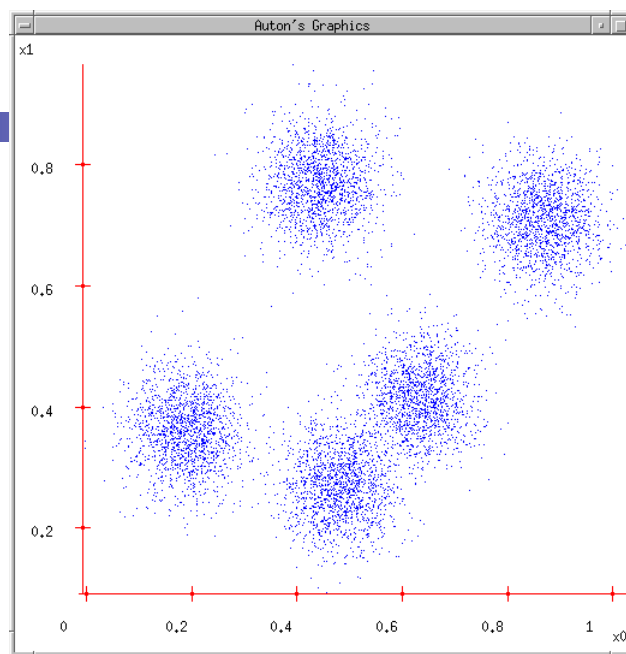
Site for newbies as well as experienced Dopefish followers, chronicling the birth of the Dopefish, its numerous appearances in several computer games, and its eventual take-over of the human race. Maintained by Mr. Dopefish himself, Joe Siegler of Apogee Software.

## Some Data



# K-means

1. Ask user how many clusters they'd like.  
(e.g.  $k=5$ )

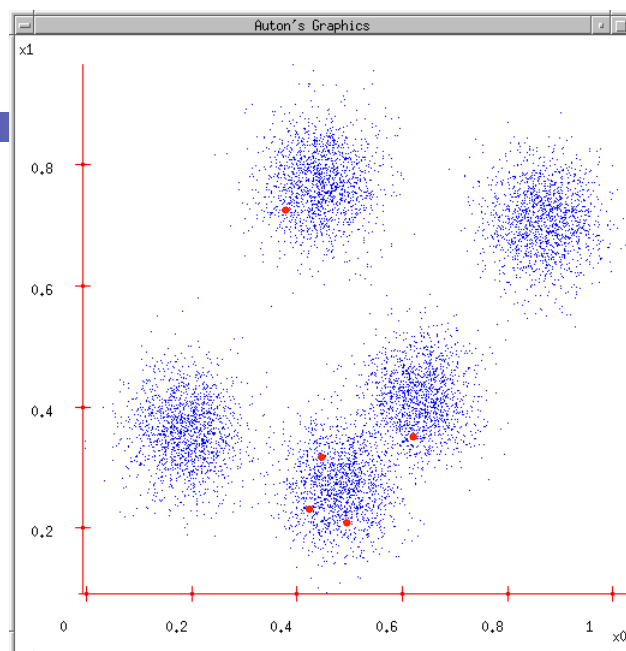


©Carlos Guestrin 2005-2013

21

# K-means

1. Ask user how many clusters they'd like.  
(e.g.  $k=5$ )
2. Randomly guess  $k$  cluster Center locations

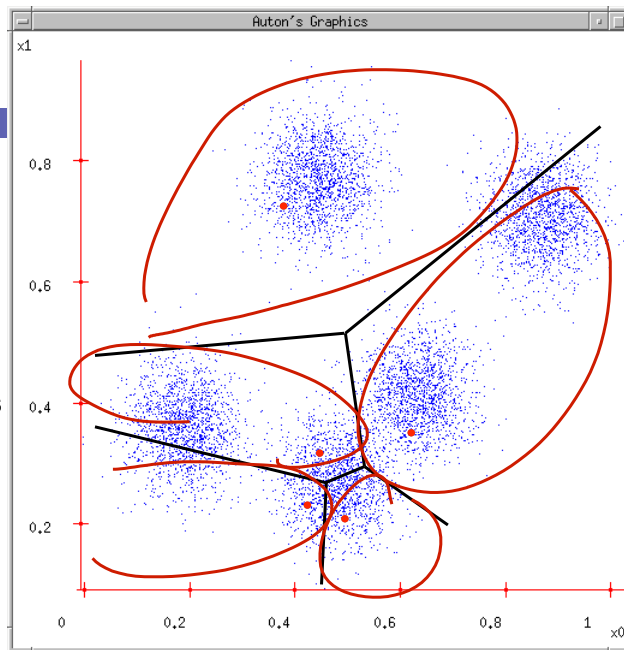


©Carlos Guestrin 2005-2013

22

# K-means

1. Ask user how many clusters they'd like.  
(e.g.  $k=5$ )
2. Randomly guess  $k$  cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)

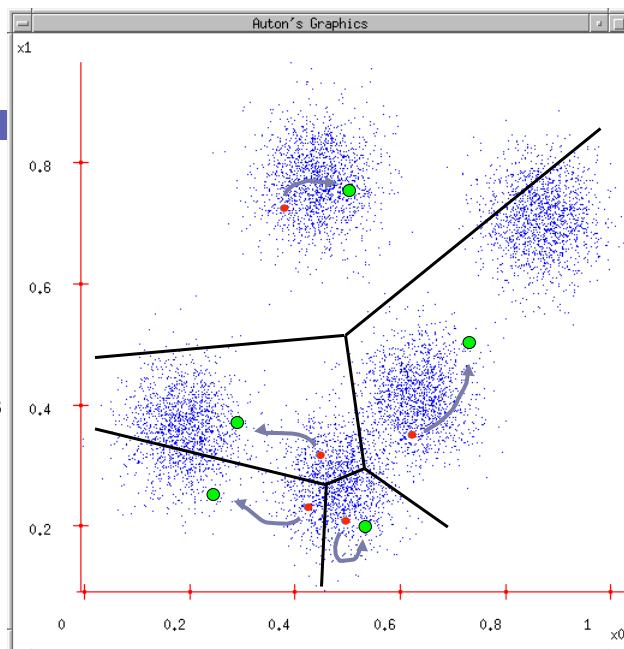


©Carlos Guestrin 2005-2013

23

# K-means

1. Ask user how many clusters they'd like.  
(e.g.  $k=5$ )
2. Randomly guess  $k$  cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns

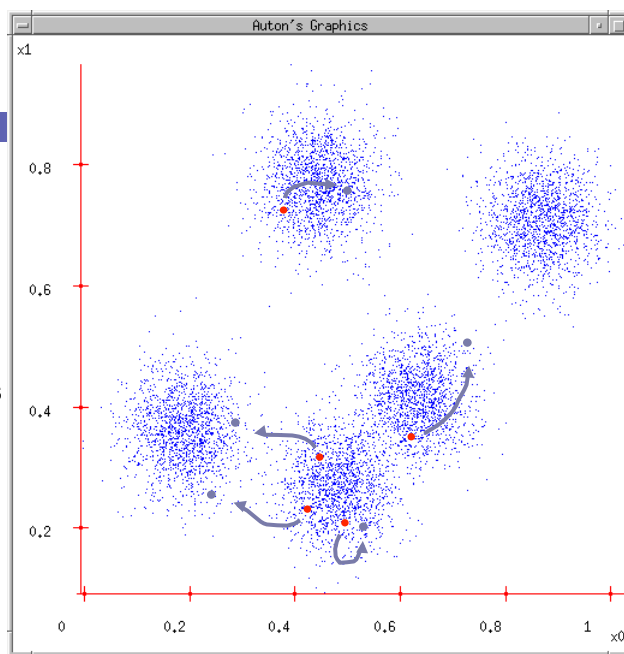


©Carlos Guestrin 2005-2013

24

# K-means

1. Ask user how many clusters they'd like.  
(e.g.  $k=5$ )
2. Randomly guess  $k$  cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns...
5. ...and jumps there
6. ...Repeat until terminated!



©Carlos Guestrin 2005-2013

25

# K-means

- Randomly initialize  $k$  centers *or smartly*
  - $\mu^{(0)} = \mu_1^{(0)}, \dots, \mu_k^{(0)}$
  - converge when nothing moves (no point changes its cluster)*
- **Classify:** Assign each point  $j \in \{1, \dots, n\}$  to nearest center:

$$\square C^{(t)}(j) \leftarrow \arg \min_i \|\mu_i - x_j\|^2$$

*Handwritten notes:  $i$  is cluster center,  $x_j$  is data point*

- **Recenter:**  $\mu_i$  becomes centroid of its point:

$$\square \mu_i^{(t+1)} \leftarrow \arg \min_{\mu} \sum_{j: C(j)=i} \|\mu - x_j\|^2$$

- Equivalent to  $\mu_i \leftarrow$  average of its points!

$$\mu_i = \frac{\sum_{j: C(j)=i} x_j}{\text{num of points assigned to cluster } i}$$

©Carlos Guestrin 2005-2013

26

## What is K-means optimizing?

- Potential function  $F(\mu, C)$  of centers  $\mu$  and point allocations  $C$ :

- $F(\mu, C) = \sum_{j=1}^N \|\mu_{C(j)} - x_j\|^2$

- Optimal K-means:

- $\min_{\mu} \min_C F(\mu, C)$

©Carlos Guestrin 2005-2013

27

## Does K-means converge??? Part 1

- Optimize potential function:

- $$\min_{\mu} \min_C F(\mu, C) = \min_{\mu} \min_C \sum_{i=1}^k \sum_{j: C(j)=i} \|\mu_i - x_j\|^2$$

- Fix  $\mu$ , optimize  $C$

©Carlos Guestrin 2005-2013

28

## Does K-means converge??? Part 2

- Optimize potential function:

$$\min_{\mu} \min_C F(\mu, C) = \min_{\mu} \min_C \sum_{i=1}^k \sum_{j: C(j)=i} \|\mu_i - x_j\|^2$$

- Fix C, optimize  $\mu$

©Carlos Guestrin 2005-2013

29

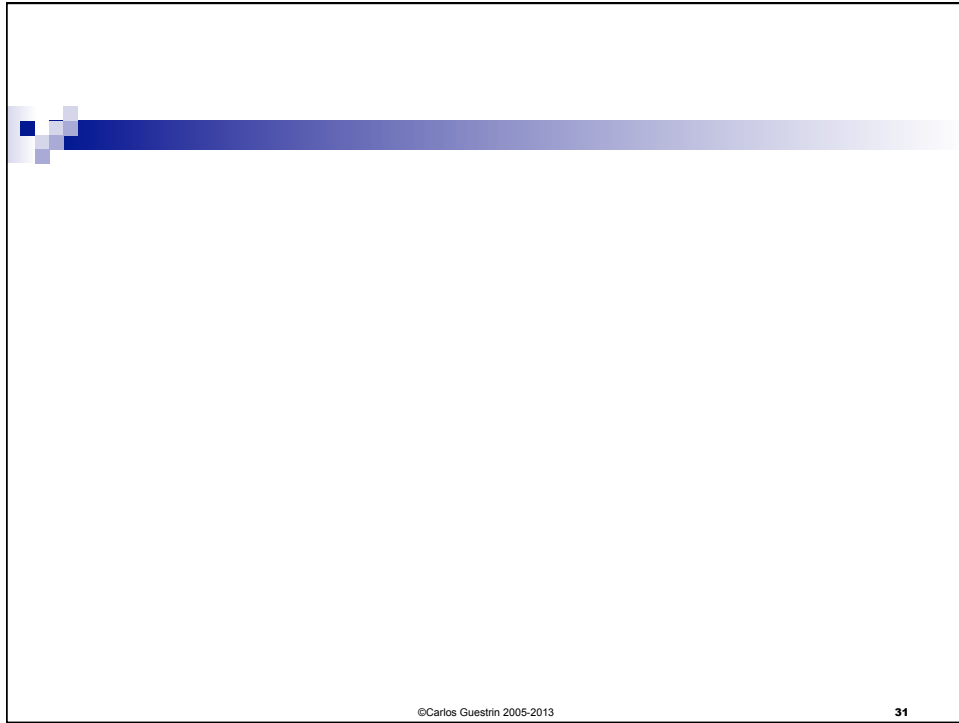
## Coordinate descent algorithms

$$\min_{\mu} \min_C F(\mu, C) = \min_{\mu} \min_C \sum_{i=1}^k \sum_{j: C(j)=i} \|\mu_i - x_j\|^2$$

- Want:  $\min_a \min_b F(a, b)$
- Coordinate descent:
  - fix a, minimize b
  - fix b, minimize a
  - repeat
- Converges!!!
  - if F is bounded
  - to a (often good) local optimum
    - as we saw in applet (play with it!)
      - (For LASSO it converged to the optimum)
- K-means is a coordinate descent algorithm!

©Carlos Guestrin 2005-2013

30







How many points can a linear boundary classify exactly? (1-D)

©Carlos Guestrin 2005-2013 34

How many points can a linear  
boundary classify exactly? (2-D)

©Carlos Guestrin 2005-2013

35

How many points can a linear  
boundary classify exactly? (d-D)

©Carlos Guestrin 2005-2013

36

## PAC bound using VC dimension

- Number of training points that can be classified exactly is VC dimension!!!
  - Measures relevant size of hypothesis space, as with decision trees with  $k$  leaves

$$\text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{VC(H) \left( \ln \frac{2m}{VC(H)} + 1 \right) + \ln \frac{4}{\delta}}{m}}$$

## Shattering a set of points

*Definition:* a **dichotomy** of a set  $S$  is a partition of  $S$  into two disjoint subsets.

*Definition:* a set of instances  $S$  is **shattered** by hypothesis space  $H$  if and only if for every dichotomy of  $S$  there exists some hypothesis in  $H$  consistent with this dichotomy.

# VC dimension

*Definition:* The **Vapnik-Chervonenkis dimension**,  $VC(H)$ , of hypothesis space  $H$  defined over instance space  $X$  is the size of the largest finite subset of  $X$  shattered by  $H$ . If arbitrarily large finite sets of  $X$  can be shattered by  $H$ , then  $VC(H) \equiv \infty$ .

## PAC bound using VC dimension

- **Number of training points that can be classified exactly is VC dimension!!!**
  - Measures relevant size of hypothesis space, as with decision trees with  $k$  leaves
  - Bound for infinite dimension hypothesis spaces:

$$\text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{VC(H) \left( \ln \frac{2m}{VC(H)} + 1 \right) + \ln \frac{4}{\delta}}{m}}$$

## Examples of VC dimension


$$\text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{VC(H) \left( \ln \frac{2m}{VC(H)} + 1 \right) + \ln \frac{4}{\delta}}{m}}$$

- Linear classifiers:

- ☐  $VC(H) = d+1$ , for  $d$  features plus constant term  $b$

- Neural networks

- ☐  $VC(H) = \text{\#parameters}$
- ☐ Local minima means NNs will probably not find best parameters

- 1-Nearest neighbor?

©Carlos Guestrin 2005-2013

41

## Another VC dim. example - What can we shatter?



- What's the VC dim. of decision stumps in 2d?

©Carlos Guestrin 2005-2013

42

## Another VC dim. example - What can't we shatter?

- What's the VC dim. of decision stumps in 2d?

## What you need to know

- Finite hypothesis space
  - Derive results
  - Counting number of hypothesis
  - Mistakes on Training data
- Complexity of the classifier depends on number of points that can be classified exactly
  - Finite case – decision trees
  - Infinite case – VC dimension
- Bias-Variance tradeoff in learning theory
- Remember: will your algorithm find best classifier?