

VC Dimension

Machine Learning – 10701/15781

Carlos Guestrin

Carnegie Mellon University

October 29th, 2007

©2005-2007 Carlos Guestrin

1

What about continuous hypothesis spaces?

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

■ Continuous hypothesis space:

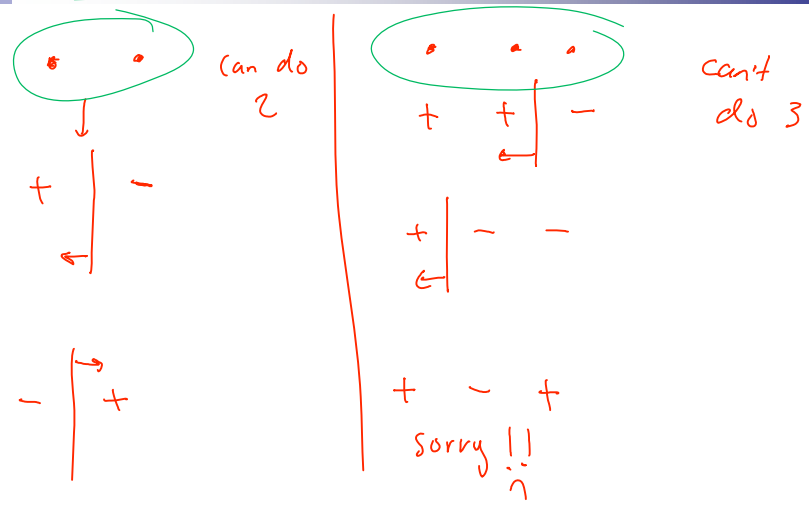
- $|H| = \infty$
- Infinite variance???

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

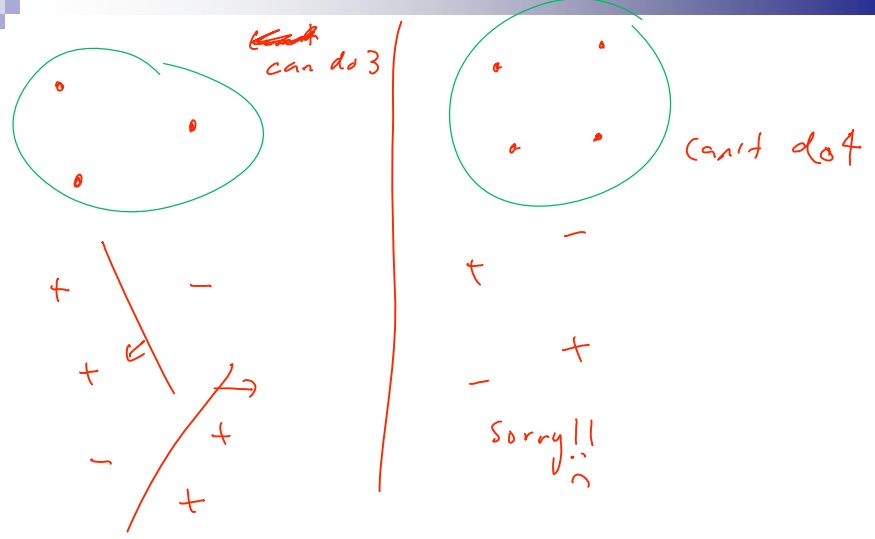
©2005-2007 Carlos Guestrin

2


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



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



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



can do $d+1$ points

how many parameters in a linear classifier in d -dimensions?

$$w_0 + \sum_{i=1}^d w_i x_i$$

$d+1$

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

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

only depend on $VC(H)$
not on $|H|$

Shattering a set of points

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

$$\rightarrow S = \{x_1, x_2, \dots, x_n\}$$

$$S^+ = \{x_1, x_7, x_{12}\}$$

$$S^- = S - S^+ = \{x_2, x_3, x_4, x_5, x_6, x_8, x_9, x_{10}, x_{11}\}$$

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.

\forall partitions of S ,
 $\exists h \in H$, classifies all S^+ as positive
 all S^- as negative

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 $VC(H) \equiv \infty$.

prove can't shatter 4 points:

\forall 4 points, adversary can always pick a label that can't be separated $VC(H) < 4$

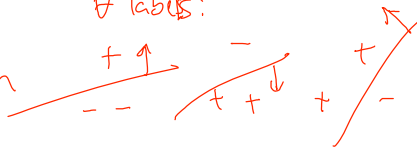
$$VC(H) = 3$$

$$VC(H) \geq 3$$

for linear classifiers in 2D:

you give me a set of points

adversary labels, \forall labels:



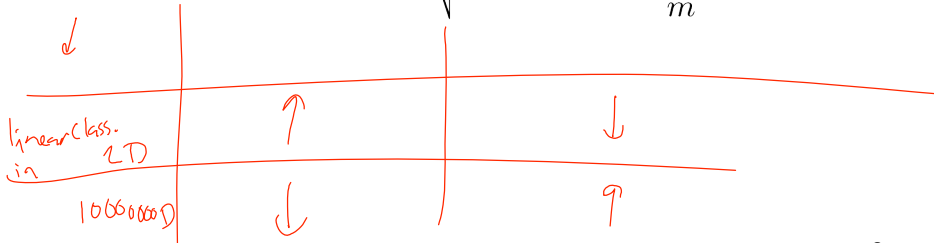
\Rightarrow can shatter 3 points !!

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?

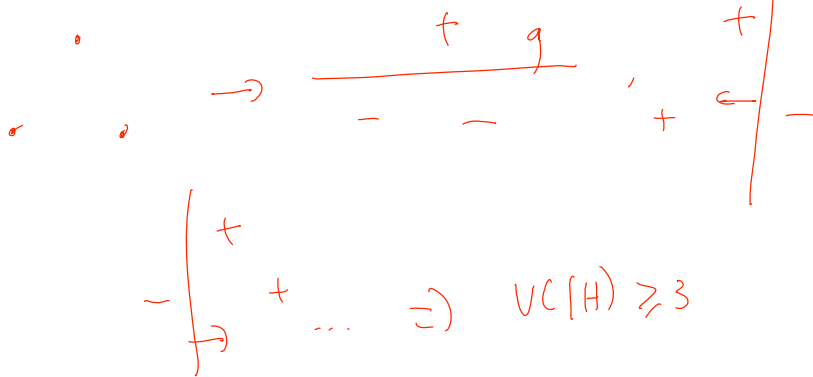


if you can find a NN, with small # param & error_train low \Rightarrow error_true "low"

\leftarrow # labels, classify exactly $\Rightarrow VC(H) = \infty$

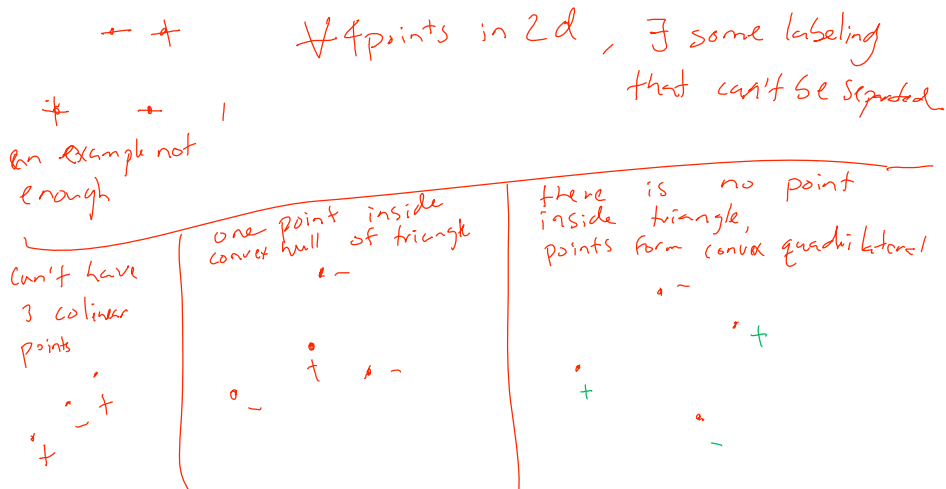
Another VC dim. example - What can we shatter?

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



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?

Questions / Suggestions

- Discussion board, *hear about it soon*
- Privacy
-
-

Bayesian Networks – Representation

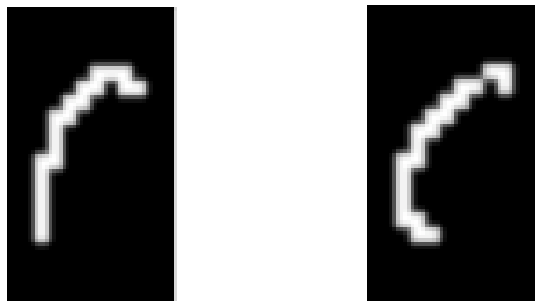
Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

October 29th, 2007

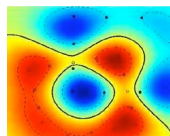
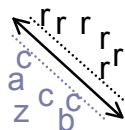
©2005-2007 Carlos Guestrin

15

Handwriting recognition



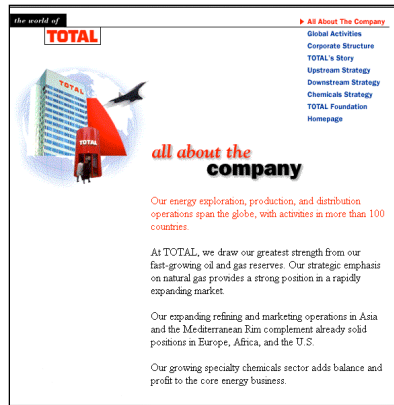
Character recognition, e.g., kernel SVMs



©2005-2007 Carlos Guestrin

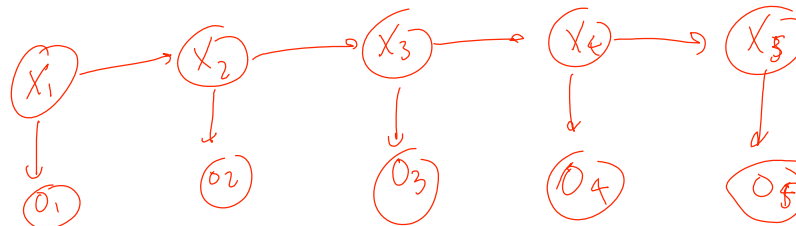
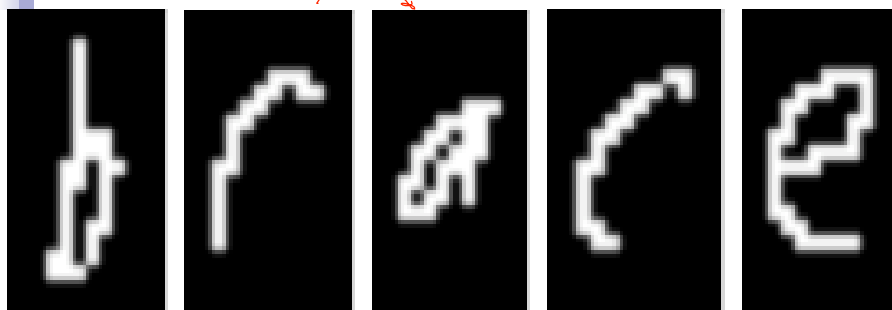
16

Webpage classification

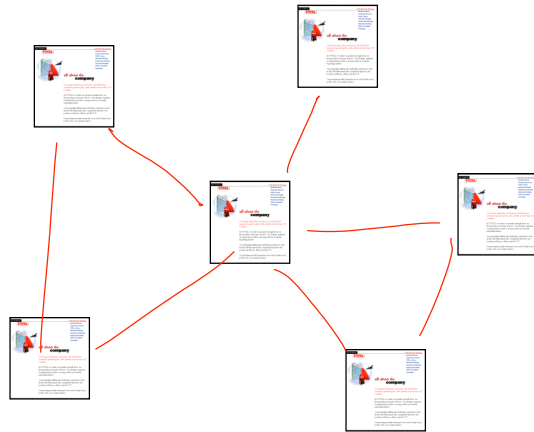


→ Company home page
VS
Personal home page
VS
University home page
VS
...

Handwriting recognition 2



Webpage classification 2

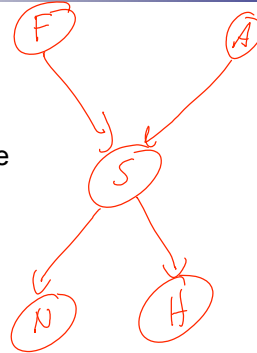


Today – Bayesian networks

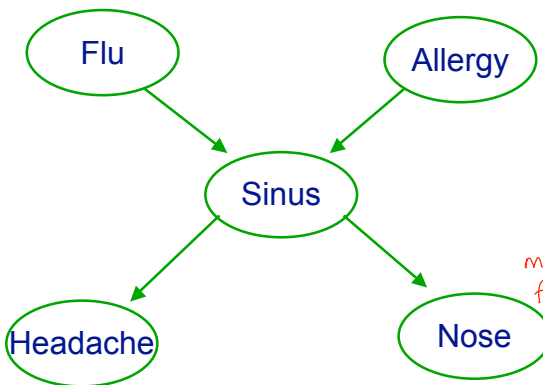
- One of the most exciting advancements in statistical AI in the last 10-15 years
- Generalizes naïve Bayes and logistic regression classifiers
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies

Causal structure

- Suppose we know the following:
 - The flu causes sinus inflammation
 - Allergies cause sinus inflammation
 - Sinus inflammation causes a runny nose
 - Sinus inflammation causes headaches
- How are these connected?



Possible queries



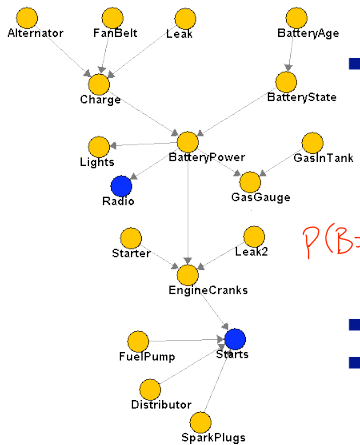
- Inference
 $P(F=t \mid H=t, N=f)$

- Most probable explanation

$\max_{f, a, s} P(f, a, s \mid H=t, N=f)$

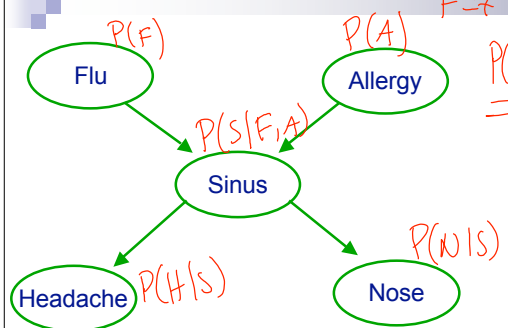
- Active data collection
what should I measure

Car starts BN



- 18 binary attributes
- Inference
 - $P(\text{BatteryAge} | \text{Starts}=f)$
 - $P(B=g | S=f) = \frac{P(B=g, S=f)}{P(S=f)}$
 - $P(B=g, S=f) = \sum_{a_1, f_1, b_1, \dots} P(B=g, S=f, A_1=a_1, F_1=f_1, \dots)$ (2¹⁶ terms)
- 2¹⁶ terms, why so fast?
- Not impressed?
 - HailFinder BN – more than 3⁵⁴ = 58149737003040059690390169 terms

Factored joint distribution - Preview



Notation $F, A \rightarrow$ I am not specify an assignment
 $f, a \rightarrow$ specific assignments
 $F=t \rightarrow \text{Flu} = \text{true}$ (a particular assignment)

$P(F, A, S, H, N)$
 $\uparrow 2^5 - 1$ (because sums to 1)
 $r = 32 - 1 = 31$

$P(F) =$

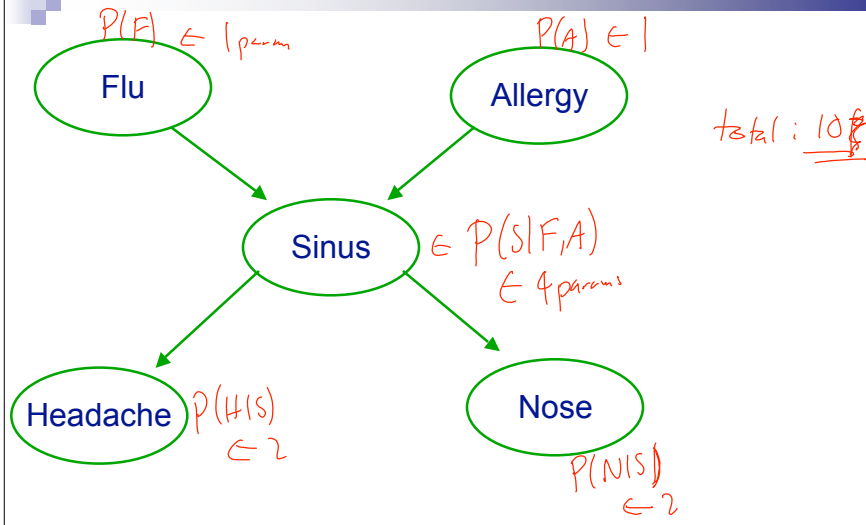
t	0.1
f	0.9

$P(H|S) :$ 2 numbers

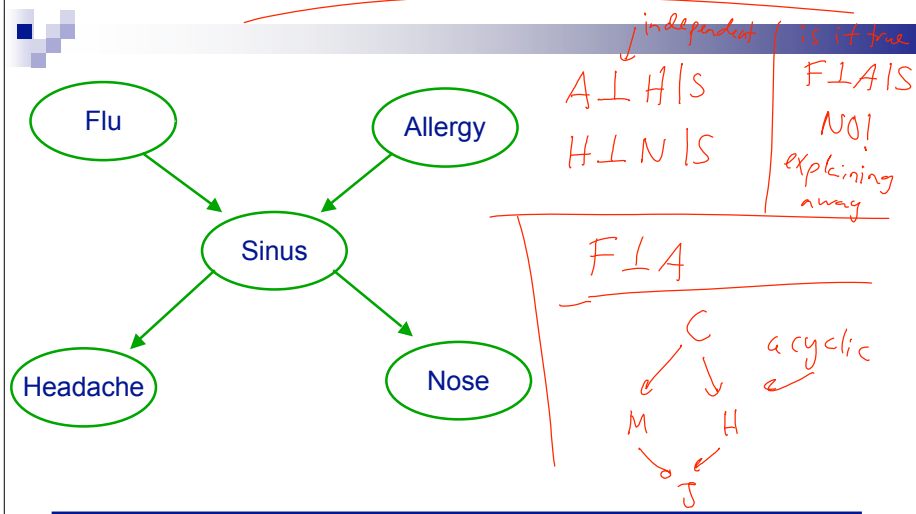
S	t	f
t	0.8	0.3
f	1-0.8 0.2	0.7

$P(F, A, S, H, N) = P(F) \cdot P(A) \cdot P(S|F, A) \cdot P(H|S) \cdot P(N|S)$

Number of parameters



Key: Independence assumptions



Knowing sinus separates the variables from each other

A BN is always acyclic / cycle: \rightarrow , not cycle: \rightarrow

(Marginal) Independence

- Flu and Allergy are (marginally) independent

$$F \perp A$$

$$P(F, A) = P(F) \cdot P(A)$$

- More Generally:

Flu = t	0.2
Flu = f	0.8

Allergy = t	0.3
Allergy = f	0.7

	Flu = t	Flu = f
Allergy = t	0.3 × 0.2	0.3 × 0.8
Allergy = f	0.2 × 0.7	0.7 × 0.8

Marginally independent random variables

- Sets of variables X, Y
- X is independent of Y if $\forall x \in \text{Val}(X), y \in \text{Val}(Y)$
 - $P(X=x \wedge Y=y) = P(X=x) \cdot P(Y=y)$
- Shorthand: $P(X=x | Y=y) = P(X=x)$
 - Marginal independence: $X \perp Y$
- Proposition: P satisfies $(X \perp Y)$ if and only if
 - $P(X, Y) = P(X) P(Y)$
 - $P(X | Y) = P(X)$