VC Dimension

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

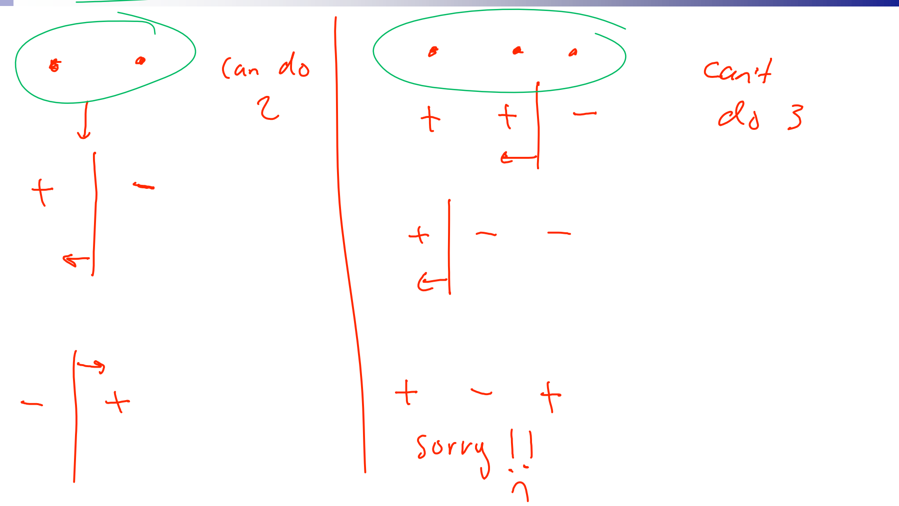
October 29th, 2007

What about continuous hypothesis spaces?

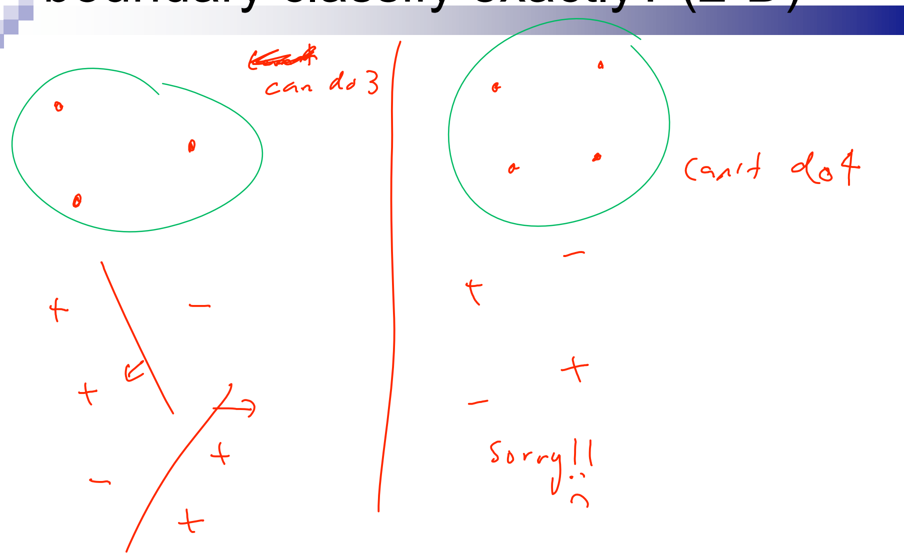
$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}}$$

- Continuous hypothesis space:
  - $|H| = \infty$
  - Infinite variance???
- **As with decision trees, only care about the maximum number of points that can be classified exactly!**

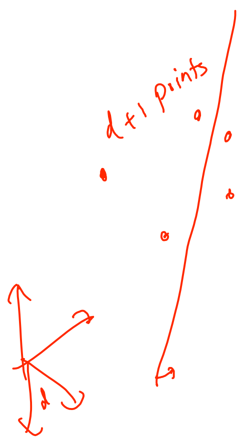
How many points can a linear boundary classify exactly? (1-D)



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**

e.g., linear classifiers

may be continuous

like ln|H|

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

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.

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.

$S = \{x_1, x_2, \ldots, x_n\}$

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

$S^- = S - S^+ = \{x_2, x_3, x_4, x_5, x_6, x_8, \ldots\}$

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

VC dimension

$VC(H) \geq 3$

$VC(H) = 3$

can't separate, $VC < 3$, because you pick points

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 $VC(H) \equiv \infty$.

for linear classifiers in 2D:

you give me a set of points

adversary labels,

$\forall$ labels:

$\exists h$

$\Rightarrow$ can shatter 3 points !!

prove can't shatter 4 points:

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

$\forall$ 4 points, adversary can find XOR

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:**

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

linear class. in 2D

1000000D

Examples of VC dimension

with prob. at least $1-\delta$:

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

- Linear classifiers:
  - VC(H) = d+1, for *d* features plus constant term *b*

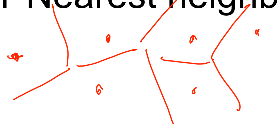
h you output need not be consistent with all train data, i.e., $\text{error}_{train}(h) \ge 0$

SVMs with RBF kernels and low bandwidth ⇒ $VC(H)=\infty$ need more powerful margin-based bounds.

- Neural networks
  - VC(H) = #parameters
  - Local minima means NNs will probably not find best parameters

if you can find a NN, with small # param., & error train low ⇒ error true "low"

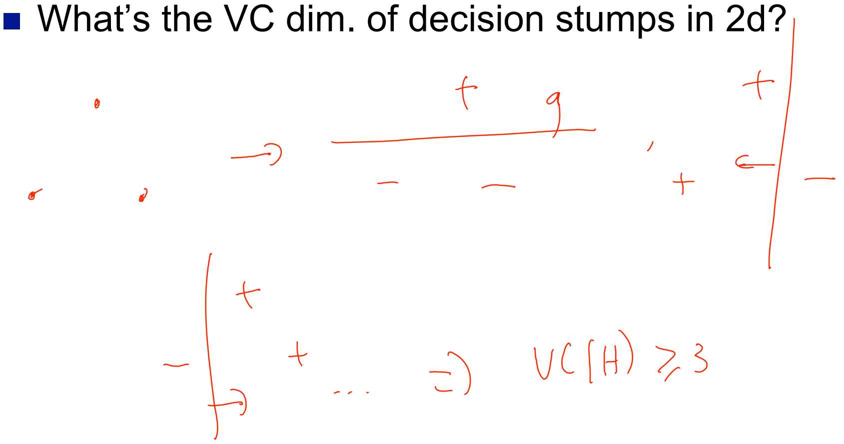
- 1-Nearest neighbor?



← ∀ labels, classify exactly ⇒ $VC(H)=\infty$

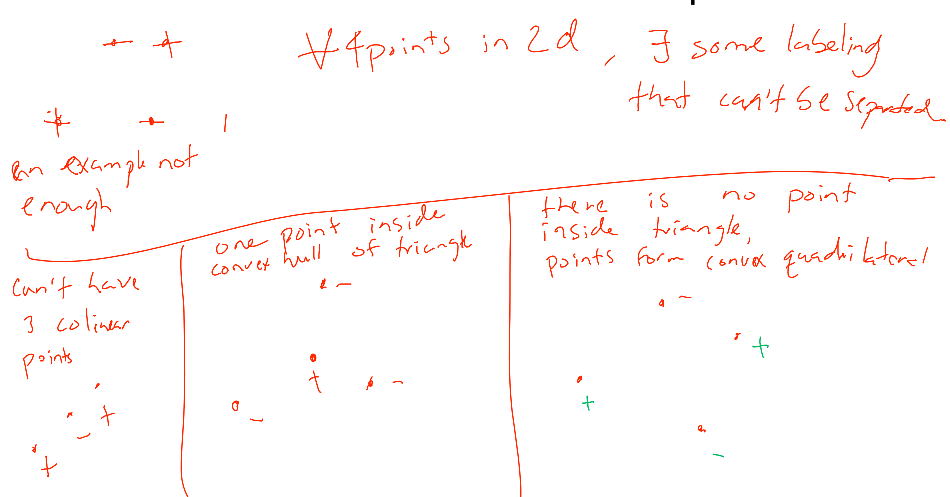
Another VC dim. example - What can we shatter?

- 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

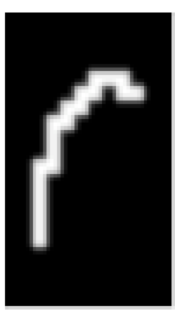
- Discussion board, hear about it soon
- Privacy
-
-

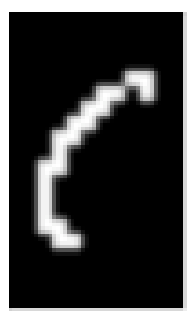
Bayesian Networks – Representation

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

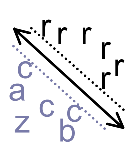
October 29th, 2007

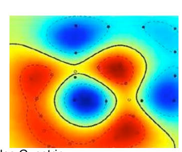
Handwriting recognition





Character recognition, e.g., kernel SVMs





Webpage classification

→ Company home page

vs

Personal home page

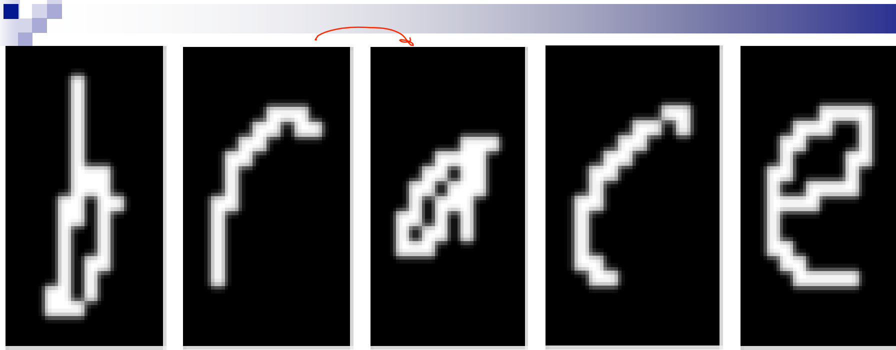
vs

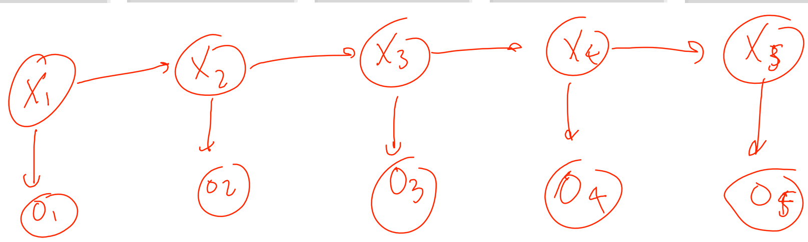
University home page

vs

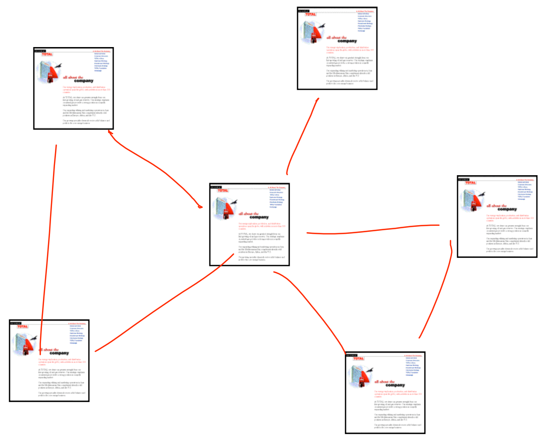
…

Handwriting recognition 2





Webpage classification 2

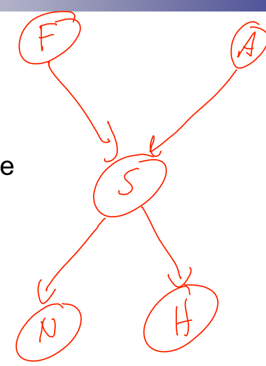


Today – Bayesian networks

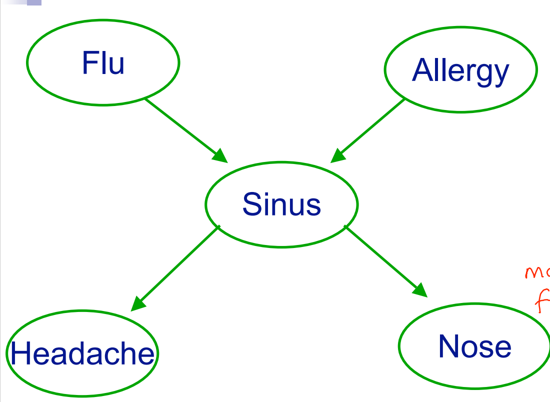
- One of the most exciting advancements in statistical AI in the last 10-15 years
- Generalizes naïve Bayes and logistic regression classifiers
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies

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

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

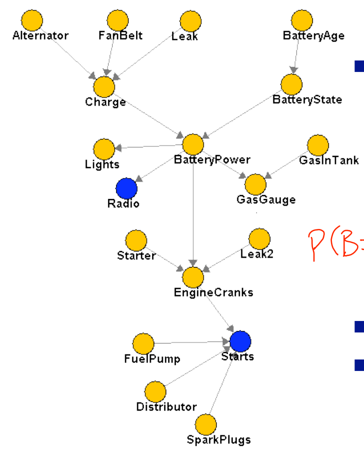
- Most probable explanation

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

- Active data collection

  what should I measure

Car starts BN



- 18 binary attributes
- Inference
  - P(BatteryAge|Starts=f)

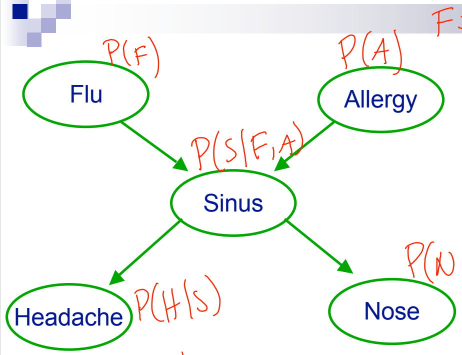
$$P(B=g \mid S=f) = \frac{P(B=g, S=f)}{P(S=f)}$$

$$P(B=g, S=f) = \sum_{a,f,bs,\ldots} P(B=g, S=f, A=a, Fb=fb, \ldots)$$

- 2^{16} terms, why so fast?
- Not impressed?
  - HailFinder BN – more than 3^{54} = 58149737003040059690390169 terms

Factored joint distribution - Preview

Notation: F, A → I am not specify an assignment
f, a → specific assignments
F=t → Flu = true (a particular assignment)



P(F,A,S,H,N)

$2^5 - 1$ (because sums to 1)

r = 32 − 1 = 31

P(F,A,S,H,N) = P(F) · P(A) · P(S|F,A) P(H|S) P(N|S)

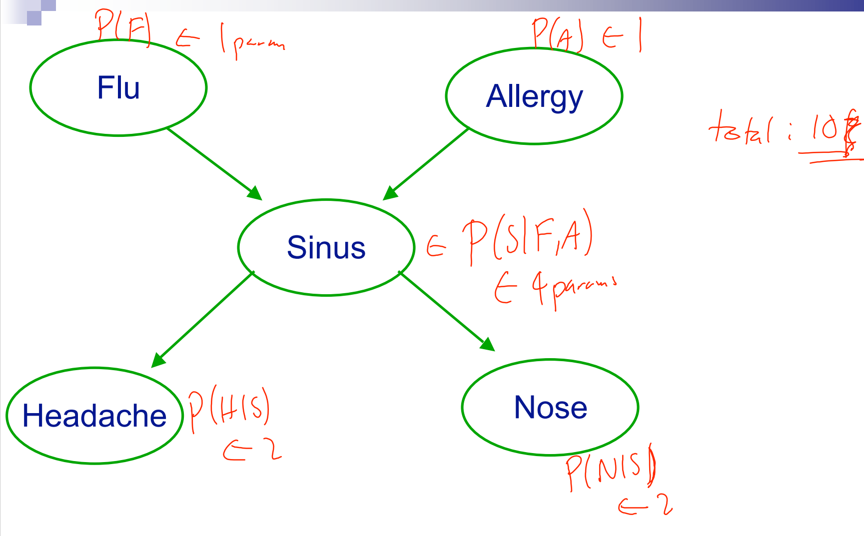
P(F) =

F	
t	0.1
f	0.9

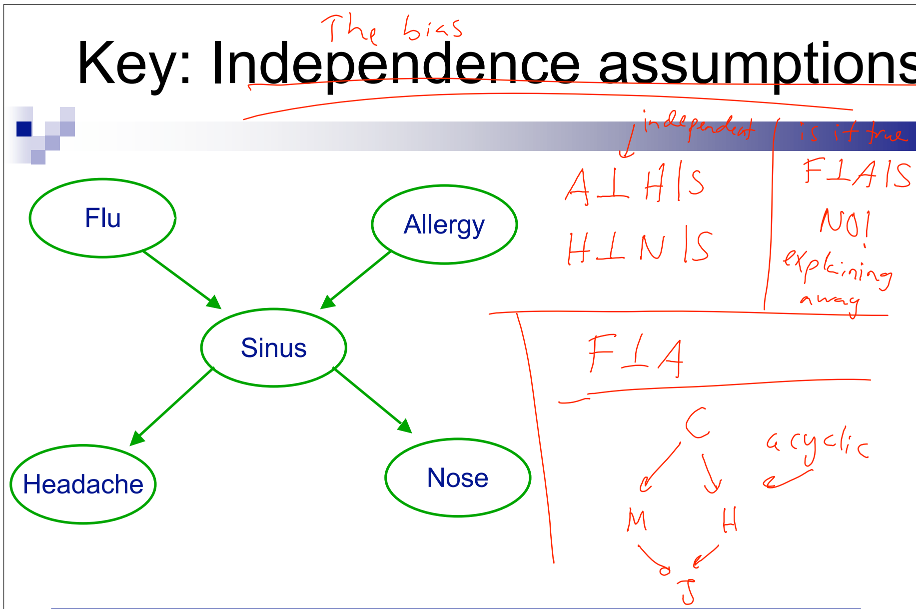
P(H|S): 2 numbers

H \ S	t	f
t	0.8	0.3
f	1−0.8 = 0.2	0.7

Number of parameters



Key: Independence assumptions



Knowing sinus separates the variables from each other

A BN is always acyclic | cycle: , not cycle:

(Marginal) Independence

- Flu and Allergy are (marginally) independent

$F \perp A$

$\Updownarrow$

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

- 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 x 0.2	0.3 x 0.8
Allergy = f	0.2 x 0.7	0.7 x 0.8

Marginally independent random variables

- **Sets** of variables **X**, **Y**
- X is independent of Y if $\forall x \in Val(X), y \in Val(Y)$
  - ~~P²~~(**X=x**⊥**Y=y**), ~~8 **x2**Val(**X**), **y2**Val(**Y**)~~

$P(X=x, Y=y) = P(X=x) \cdot P(Y=y)$

- Shorthand: $P(X=x \mid Y=y) = P(X=x)$
  - **Marginal independence:** ~~P²~~ (**X** ⊥ **Y**)
- **Proposition:** *P* satisfies (**X** ⊥ **Y**) if and only if
  - P(**X**,**Y**) = P(**X**) P(**Y**)

$P(X \mid Y) = P(X)$