

# VC-dimension for characterizing classifiers

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: <http://www.cs.cmu.edu/~awm/tutorials>. Comments and corrections gratefully received.

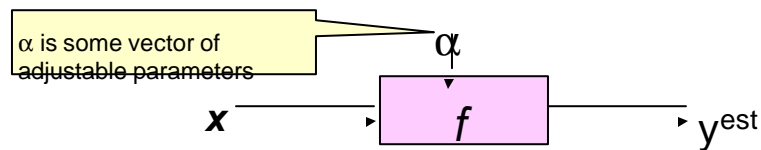
**Andrew W. Moore**  
**Associate Professor**  
**School of Computer Science**  
**Carnegie Mellon University**  
www.cs.cmu.edu/~awm  
awm@cs.cmu.edu  
412-268-7599

Copyright © 2001, Andrew W. Moore

Nov 20th, 2001

## A learning machine

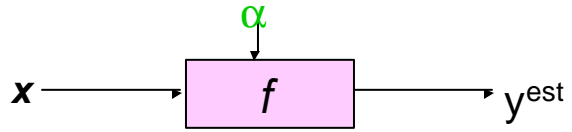
- A learning machine  $f$  takes an input  $x$  and transforms it, somehow using weights  $\alpha$ , into a predicted output  $y^{est} = +/- 1$



Copyright © 2001, Andrew W. Moore

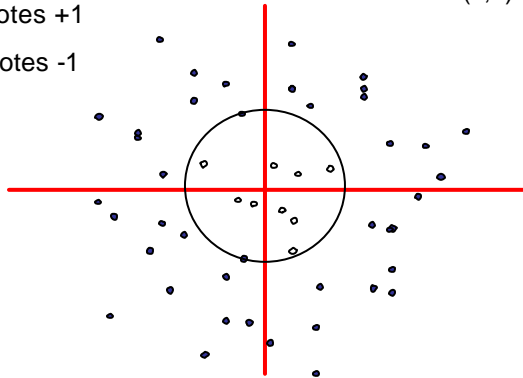
VC-dimension: Slide 2

# Examples



$$f(x,b) = \text{sign}(x \cdot x - b)$$

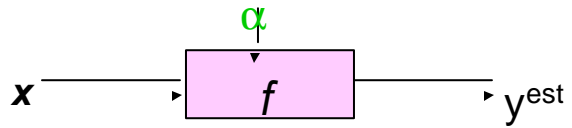
- denotes +1
- denotes -1



Copyright © 2001, Andrew W. Moore

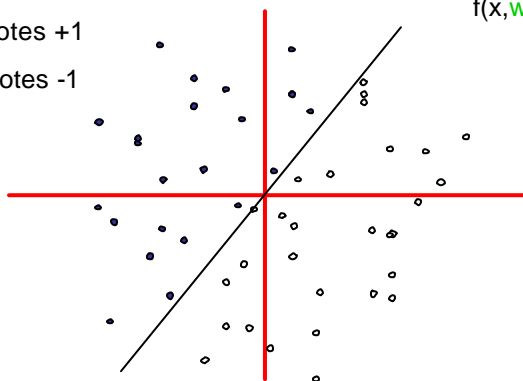
VC-dimension: Slide 3

# Examples



$$f(x,w) = \text{sign}(x \cdot w)$$

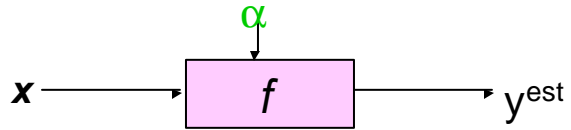
- denotes +1
- denotes -1



Copyright © 2001, Andrew W. Moore

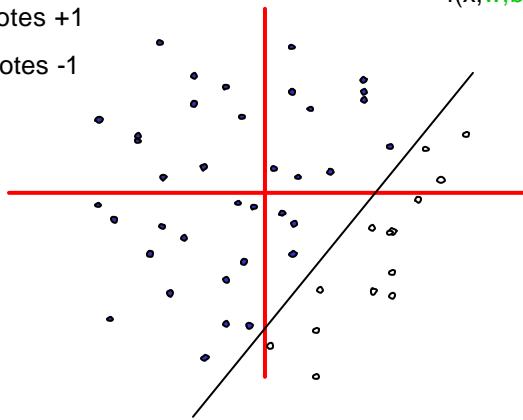
VC-dimension: Slide 4

## Examples



$$f(x, w, b) = \text{sign}(x \cdot w + b)$$

- denotes +1
- denotes -1



Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 5

## How do we characterize “power”?

- Different machines have different amounts of “power”.
- Tradeoff between:
  - More power: Can model more complex classifiers but might overfit.
  - Less power: Not going to overfit, but restricted in what it can model.
- How do we characterize the amount of power?

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 6

## Some definitions

- Given some machine  $f$
- And under the assumption that all training points  $(x_k, y_k)$  were drawn i.i.d from some distribution.
- And under the assumption that future test points will be drawn from the same distribution
- Define

$$R(\mathbf{a}) = \text{TESTERR}(\mathbf{a}) = E \left[ \frac{1}{2} |y - f(x, \mathbf{a})| \right] = \text{Probabilit y of Misclassification}$$

Official terminology

Terminology we'll use

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 7

## Some definitions

- Given some machine  $f$
- And under the assumption that all training points  $(x_k, y_k)$  were drawn i.i.d from some distribution.
- And under the assumption that future test points will be drawn from the same distribution
- Define

$$R(\mathbf{a}) = \text{TESTERR}(\mathbf{a}) = E \left[ \frac{1}{2} |y - f(x, \mathbf{a})| \right] = \text{Probabilit y of Misclassification}$$

Official terminology

Terminology we'll use

$$R^{emp}(\mathbf{a}) = \text{TRAINERR}(\mathbf{a}) = \frac{1}{R} \sum_{k=1}^R \frac{1}{2} |y_k - f(x_k, \mathbf{a})| = \text{Fraction Training Set misclassified}$$

R = #training set data points

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 8

## Vapnik-Chervonenkis dimension

$$\text{TESTERR}(\mathbf{a}) = E\left[\frac{1}{2}|y - f(x, \mathbf{a})|\right] \quad \text{TRAINERR}(\mathbf{a}) = \frac{1}{R} \sum_{k=1}^R \frac{1}{2}|y_k - f(x_k, \mathbf{a})|$$

- Given some machine  $f$ , let  $h$  be its VC dimension.
- $h$  is a measure of  $f$ 's power ( $h$  does not depend on the choice of training set)
- Vapnik showed that with probability  $1-\eta$

$$\text{TESTERR}(\mathbf{a}) \leq \text{TRAINERR}(\mathbf{a}) + \sqrt{\frac{h(\log(2R/h) + 1) - \log(h/4)}{R}}$$

This gives us a way to estimate the error on future data based only on the training error and the VC-dimension of  $f$

## What VC-dimension is used for

$$\text{TESTERR}(\mathbf{a}) = E\left[\frac{1}{2}|y - f(x, \mathbf{a})|\right] \quad \text{TRAINERR}(\mathbf{a}) = \frac{1}{R} \sum_{k=1}^R \frac{1}{2}|y_k - f(x_k, \mathbf{a})|$$

- Given some machine  $f$ , let  $h$  be its VC dimension.
- $h$  is a measure of  $f$ 's power
- Vapnik showed that

$$\text{TESTERR}(\mathbf{a}) \leq \text{TRAINERR}(\mathbf{a}) + \sqrt{\frac{h(\log(2R/h) + 1) - \log(h/4)}{R}}$$

This gives us a way to estimate the error on future data based only on the training error and the VC-dimension of  $f$

# Shattering

- Machine  $f$  can *shatter* a set of points  $x_1, x_2 \dots x_r$  if and only if...  
For every possible training set of the form  $(x_1, y_1), (x_2, y_2), \dots (x_r, y_r)$   
...There exists some value of  $\alpha$  that gets zero training error.

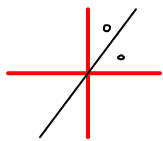
There are  $2^r$  such training sets to consider, each with a different combination of +1's and -1's for the y's

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 11

# Shattering

- Machine  $f$  can *shatter* a set of points  $x_1, x_2 \dots x_r$  if and only if...  
For every possible training set of the form  $(x_1, y_1), (x_2, y_2), \dots (x_r, y_r)$   
...There exists some value of  $\alpha$  that gets zero training error.
- Question: Can the following  $f$  shatter the following points?



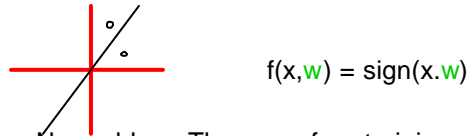
$$f(x, w) = \text{sign}(x \cdot w)$$

Copyright © 2001, Andrew W. Moore

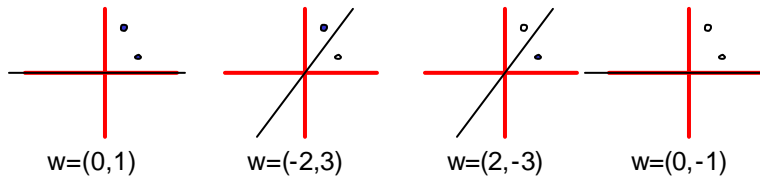
VC-dimension: Slide 12

# Shattering

- Machine  $f$  can *shatter* a set of points  $x_1, x_2 \dots x_r$  if and only if...  
For every possible training set of the form  $(x_1, y_1), (x_2, y_2), \dots (x_r, y_r)$   
...There exists some value of  $\alpha$  that gets zero training error.
- Question: Can the following  $f$  shatter the following points?



- Answer: No problem. There are four training sets to consider

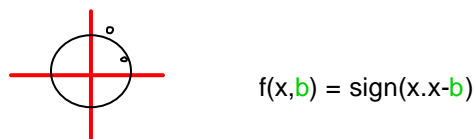


Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 13

# Shattering

- Machine  $f$  can *shatter* a set of points  $x_1, x_2 \dots x_r$  if and only if...  
For every possible training set of the form  $(x_1, y_1), (x_2, y_2), \dots (x_r, y_r)$   
...There exists some value of  $\alpha$  that gets zero training error.
- Question: Can the following  $f$  shatter the following points?

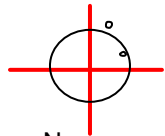


Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 14

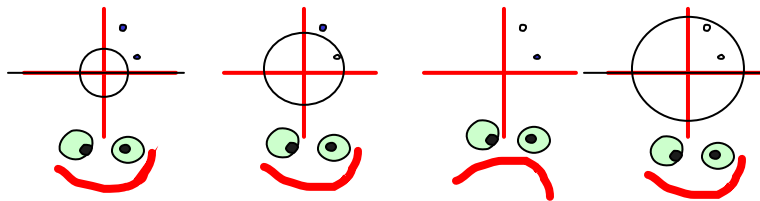
# Shattering

- Machine  $f$  can *shatter* a set of points  $x_1, x_2 \dots x_r$  if and only if...  
For every possible training set of the form  $(x_1, y_1), (x_2, y_2), \dots (x_r, y_r)$   
...There exists some value of  $\alpha$  that gets zero training error.
- Question: Can the following  $f$  shatter the following points?



$$f(x,b) = \text{sign}(x \cdot x - b)$$

- Answer: No way my friend.



Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 15

# Definition of VC dimension

Given machine  $f$ , the VC-dimension  $h$  is  
The maximum number of points that can be  
arranged so that  $f$  shatter them.

**Example:** What's VC dimension of  $f(x,b) = \text{sign}(x \cdot x - b)$

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 16

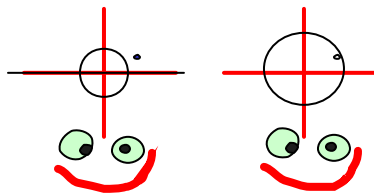
## VC dim of trivial circle

Given machine  $f$ , the VC-dimension  $h$  is

The maximum number of points that can be arranged so that  $f$  shatter them.

**Example:** What's VC dimension of  $f(x,b) = \text{sign}(x \cdot x - b)$

**Answer = 1:** we can't even shatter two points! (but it's clear we can shatter 1)



Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 17

## Reformulated circle

Given machine  $f$ , the VC-dimension  $h$  is

The maximum number of points that can be arranged so that  $f$  shatter them.

**Example:** For 2-d inputs, what's VC dimension of  $f(x,q,b) = \text{sign}(qx \cdot x - b)$

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 18

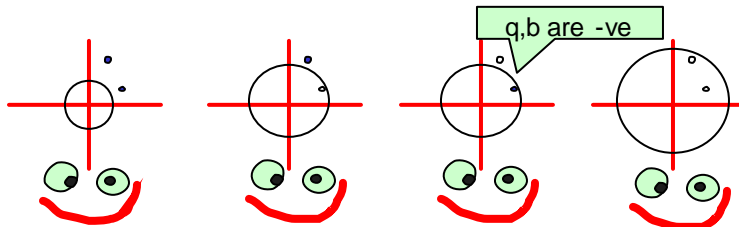
## Reformulated circle

Given machine  $f$ , the VC-dimension  $h$  is

The maximum number of points that can be arranged so that  $f$  shatter them.

**Example:** What's VC dimension of  $f(x, q, b) = \text{sign}(qx \cdot x - b)$

- Answer = 2



Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 19

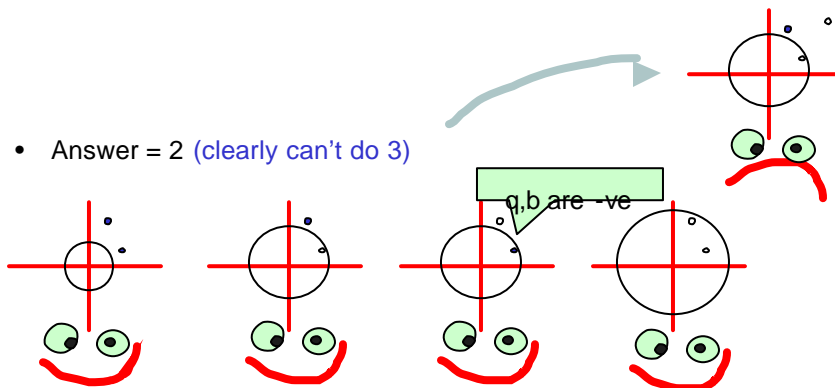
## Reformulated circle

Given machine  $f$ , the VC-dimension  $h$  is

The maximum number of points that can be arranged so that  $f$  shatter them.

**Example:** What's VC dimension of  $f(x, q, b) = \text{sign}(qx \cdot x - b)$

- Answer = 2 (clearly can't do 3)



Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 20

## VC dim of separating line

Given machine  $f$ , the VC-dimension  $h$  is

The maximum number of points that can be arranged so that  $f$  shatter them.

**Example:** For 2-d inputs, what's VC-dim of  $f(x, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot x + b)$ ?  
Well, can  $f$  shatter these three points?



Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 21

## VC dim of line machine

Given machine  $f$ , the VC-dimension  $h$  is

The maximum number of points that can be arranged so that  $f$  shatter them.

**Example:** For 2-d inputs, what's VC-dim of  $f(x, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot x + b)$ ?  
Well, can  $f$  shatter these three points?



Yes, of course.

All -ve or all +ve is trivial

One +ve can be picked off by a line

One -ve can be picked off too.

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 22

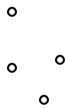
## VC dim of line machine

Given machine  $f$ , the VC-dimension  $h$  is

The maximum number of points that can be arranged so that  $f$  shatter them.

**Example:** For 2-d inputs, what's VC-dim of  $f(x, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot x + b)$ ?

Well, can we find four points that  $f$  can shatter?



Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 23

## VC dim of line machine

Given machine  $f$ , the VC-dimension  $h$  is

The maximum number of points that can be arranged so that  $f$  shatter them.

**Example:** For 2-d inputs, what's VC-dim of  $f(x, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot x + b)$ ?

Well, can we find four points that  $f$  can shatter?



Can always draw six lines between pairs of four points.

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 24

## VC dim of line machine

Given machine  $f$ , the VC-dimension  $h$  is

The maximum number of points that can be arranged so that  $f$  shatter them.

**Example:** For 2-d inputs, what's VC-dim of  $f(x, \mathbf{w}, \mathbf{b}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$ ?

Well, can we find four points that  $f$  can shatter?



Can always draw six lines between pairs of four points.

Two of those lines will cross.

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 25

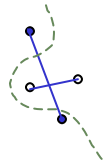
## VC dim of line machine

Given machine  $f$ , the VC-dimension  $h$  is

The maximum number of points that can be arranged so that  $f$  shatter them.

**Example:** For 2-d inputs, what's VC-dim of  $f(x, \mathbf{w}, \mathbf{b}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$ ?

Well, can we find four points that  $f$  can shatter?



Can always draw six lines between pairs of four points.

Two of those lines will cross.

If we put points linked by the crossing lines in the same class they can't be linearly separated

So a line can shatter 3 points but not 4

So VC-dim of Line Machine is 3

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 26

## VC dim of linear classifiers in m-dimensions

If input space is m-dimensional and if  $f$  is  $\text{sign}(\mathbf{w} \cdot \mathbf{x} - b)$ , what is the VC-dimension?

Proof that  $h \geq m$ : Show that  $m$  points can be shattered

*Can you guess how?*

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 27

## VC dim of linear classifiers in m-dimensions

If input space is m-dimensional and if  $f$  is  $\text{sign}(\mathbf{w} \cdot \mathbf{x} - b)$ , what is the VC-dimension?

Proof that  $h \geq m$ : Show that  $m$  points can be shattered

Define  $m$  input points thus:

$$\mathbf{x}_1 = (1, 0, 0, \dots, 0)$$

$$\mathbf{x}_2 = (0, 1, 0, \dots, 0)$$

:

$$\mathbf{x}_m = (0, 0, 0, \dots, 1) \quad \text{So } x_k[j] = 1 \text{ if } k=j \text{ and } 0 \text{ otherwise}$$

Let  $y_1, y_2, \dots, y_m$ , be any one of the  $2^m$  combinations of class labels.

Guess how we can define  $w_1, w_2, \dots, w_m$  and  $b$  to ensure

$\text{sign}(\mathbf{w} \cdot \mathbf{x}_k + b) = y_k$  for all  $k$ ? Note:

$$\text{sign}(\mathbf{w} \cdot \mathbf{x}_k + b) = \text{sign}\left(b + \sum_{j=1}^m w_j \cdot x_k[j]\right)$$

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 28

## VC dim of linear classifiers in m-dimensions

If input space is m-dimensional and if  $f$  is  $\text{sign}(\mathbf{w} \cdot \mathbf{x} - b)$ , what is the VC-dimension?

Proof that  $h \geq m$ : Show that  $m$  points can be shattered

Define  $m$  input points thus:

$$\mathbf{x}_1 = (1, 0, 0, \dots, 0)$$

$$\mathbf{x}_2 = (0, 1, 0, \dots, 0)$$

:

$$\mathbf{x}_m = (0, 0, 0, \dots, 1) \quad \text{So } x_k[j] = 1 \text{ if } k=j \text{ and } 0 \text{ otherwise}$$

Let  $y_1, y_2, \dots, y_m$ , be any one of the  $2^m$  combinations of class labels.

Guess how we can define  $w_1, w_2, \dots, w_m$  and  $b$  to ensure  $\text{sign}(\mathbf{w} \cdot \mathbf{x}_k + b) = y_k$  for all  $k$ ? Note:

Answer:  $b=0$  and  $w_k = y_k$  for all  $k$ .

$$\text{sign}(\mathbf{w} \cdot \mathbf{x}_k + b) = \text{sign}\left(b + \sum_{j=1}^m w_j \cdot x_k[j]\right)$$

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 29

## VC dim of linear classifiers in m-dimensions

If input space is m-dimensional and if  $f$  is  $\text{sign}(\mathbf{w} \cdot \mathbf{x} - b)$ , what is the VC-dimension?

- Now we know that  $h \geq m$
- In fact,  $h=m+1$
- Proof that  $h \geq m+1$  is easy
- Proof that  $h < m+2$  is moderate

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 30

## What does VC-dim measure?

- Is it the number of parameters?  
Related but not really the same.
- I can create a machine with one numeric parameter that really encodes 7 parameters (How?)
- And I can create a machine with 7 parameters which has a VC-dim of 1 (How?)
- *Andrew's private opinion: it often is the number of parameters that counts.*

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 31

## Structural Risk Minimization

- Let  $f(f)$  = the set of functions representable by  $f$ .
- Suppose  $f(f_1) \subseteq f(f_2) \subseteq \dots \subseteq f(f_n)$
- Then  $h(f_1) \leq h(f_2) \leq \dots \leq h(f_n)$  (Hey, can you formally prove this?)
- We're trying to decide which machine to use.
- We train each machine and make a table...

$$\text{TESTERR}(\mathbf{a}) \leq \text{TRAINERR}(\mathbf{a}) + \sqrt{\frac{h(\log(2R/h) + 1) - \log(h/4)}{R}}$$

$i$	$f_i$	TRAINER R	VC Conf	Probable upper bound on TESTERR	Choice
1	$f_1$				
2	$f_2$				
3	$f_3$				
4	$f_4$				
5	$f_5$				
6	$f_6$				

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 32

# Using VC-dimensionality

That's what VC-dimensionality is about

People have worked hard to find VC-dimension for..

- Decision Trees
- Perceptrons
- Neural Nets
- Decision Lists
- Support Vector Machines
- And many many more

All with the goals of

1. Understanding which learning machines are more or less powerful under which circumstances
2. Using Structural Risk Minimization for to choose the best learning machine

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 33

## Alternatives to VC-dim-based model selection

- What could we do instead of the scheme below?

$$\text{TESTERR}(\mathbf{a}) \leq \text{TRAINERR}(\mathbf{a}) + \sqrt{\frac{h(\log(2R/h) + 1) - \log(h/4)}{R}}$$












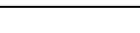
$i$	$f_i$	TRAINER R	VC Conf	Probable upper bound on TESTERR	Choice
1	$f_1$				
2	$f_2$				
3	$f_3$				
4	$f_4$				
5	$f_5$				
6	$f_6$				

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 34

## Alternatives to VC-dim-based model selection

- What could we do instead of the scheme below?
  1. Cross-validation

$i$	$f_i$	TRAINER	10-FOLD-CV-ERR	Choice
1	$f_1$			
2	$f_2$			
3	$f_3$			
4	$f_4$			
5	$f_5$			
6	$f_6$			

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 35



















## Alternatives to VC-dim-based model selection

- What could we do instead of the scheme below?
  1. Cross-validation
  2. AIC (Akaike Information Criterion)

As the amount of data goes to infinity, AIC promises\* to select the model that'll have the best likelihood for future data

\*Subject to about a million caveats

$$\text{AICSCORE} = LL(\text{Data} | \text{MLE params}) - (\# \text{ parameters})$$

$i$	$f_i$	LOGLIKE(TRAINERR)	#parameters	AIC	Choice
1	$f_1$				
2	$f_2$				
3	$f_3$				
4	$f_4$				
5	$f_5$				
6	$f_6$				

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 36

## Alternatives to VC-dim-based model selection

- What could we do instead of the scheme below?
  1. Cross-validation
  2. AIC (Akaike Information Criterion)
  3. BIC (Bayesian Information Criterion)

As the amount of data goes to infinity, BIC promises\* to select the model that the data was generated from. More conservative than AIC.

$$\text{BICSCORE} = LL(\text{Data} | \text{MLE params}) - \frac{\# \text{ params}}{2} \log R$$

\*Another million caveats

$i$	$f_i$	LOGLIKE(TRAINERR)	#parameters	BIC	Choice
1	$f_1$				
2	$f_2$				
3	$f_3$				
4	$f_4$				
5	$f_5$				
6	$f_6$				

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 37

## Which model selection method is best?

1. (CV) Cross-validation
  2. AIC (Akaike Information Criterion)
  3. BIC (Bayesian Information Criterion)
  4. (SRMVC) Structural Risk Minimize with VC-dimension
- AIC, BIC and SRMVC have the advantage that you only need the training error.
  - CV error might have more variance
  - SRMVC is wildly conservative
  - Asymptotically AIC and Leave-one-out CV should be the same
  - Asymptotically BIC and a carefully chosen k-fold should be the same
  - BIC is what you want if you want the best structure instead of the best predictor (e.g. for clustering or Bayes Net structure finding)
  - Many alternatives to the above including proper Bayesian approaches.
  - It's an emotional issue.

Copyright © 2001, Andrew W. Moore

VC-dimension: Slide 38

## Extra Comments

- Beware: that second “VC-confidence” term is usually very very conservative (at least hundreds of times larger than the empirical overfitting effect).
- An excellent tutorial on VC-dimension and Support Vector Machines (which we’ll be studying soon):  
C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):955-974, 1998.  
<http://citeseer.nj.nec.com/burges98tutorial.html>

## What you should know

- The definition of a learning machine:  $f(\mathbf{x}, \mathbf{a})$
- The definition of Shattering
- Be able to work through simple examples of shattering
- The definition of VC-dimension
- Be able to work through simple examples of VC-dimension
- Structural Risk Minimization for model selection
- Awareness of other model selection methods