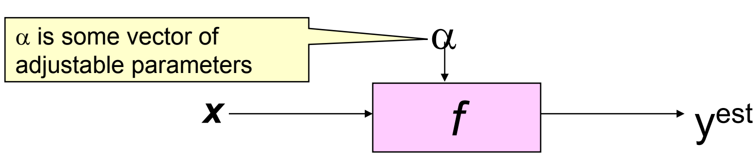
VC-dimension for Characterizing Classifiers

Ronald J. Williams
CSG220
Fall 2004

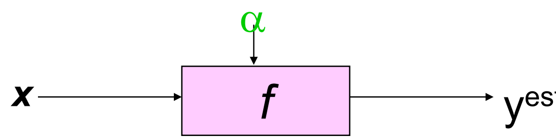
A slightly modified version of the Andrew Moore tutorial with this same title

A learning machine

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

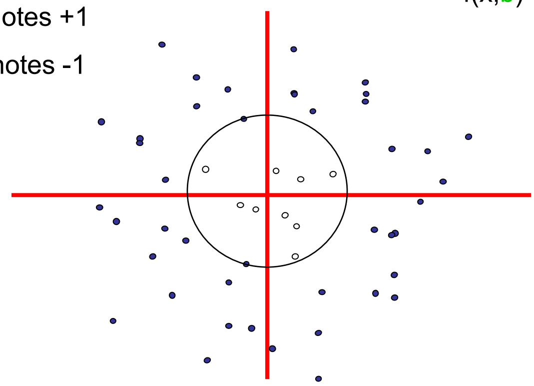


Examples

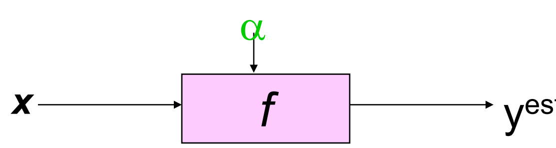


$f(x,b) = \mathrm{sgn}(x \cdot x - b)$

- denotes +1
- denotes -1

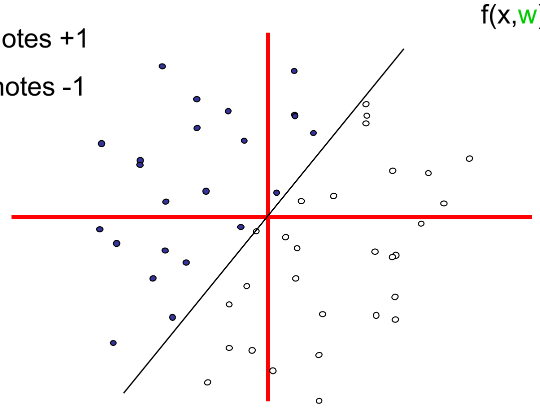


Examples

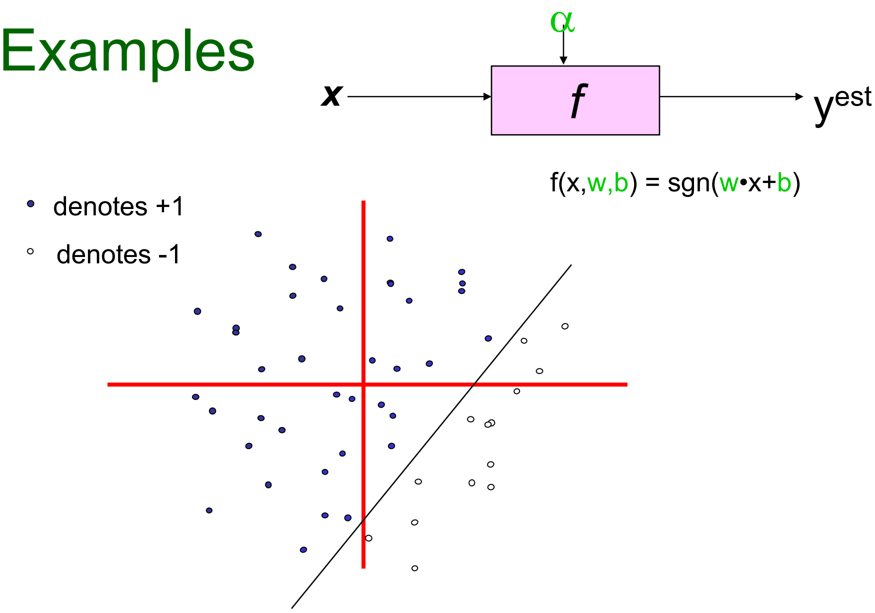


$f(x,w) = \mathrm{sgn}(w \cdot x)$

- denotes +1
- denotes -1



Examples



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?
- In the literature: “power” often called *capacity.*

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

$$R(\alpha) = \text{TESTERR}(\alpha) = E\left[\frac{1}{2}\left|y - f(x,\alpha)\right|\right] = \begin{matrix}\text{Probability of} \\ \text{Misclassification}\end{matrix}$$

Official terminology: *(Actual) Risk*

Terminology we'll use

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

$$R(\alpha) = \text{TESTERR}(\alpha) = E\left[\frac{1}{2}\left|y - f(x,\alpha)\right|\right] = \begin{matrix}\text{Probability of} \\ \text{Misclassification}\end{matrix}$$

Official terminology: *Empirical Risk*

Terminology we'll use

$$R^{emp}(\alpha) = \text{TRAINERR}(\alpha) = \frac{1}{R}\sum_{k=1}^{R}\frac{1}{2}\left|y_k - f(x_k,\alpha)\right| = \begin{matrix}\text{Fraction Training} \\ \text{Set misclassified}\end{matrix}$$

R = #training set data points

Vapnik-Chervonenkis dimension

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

- 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-δ

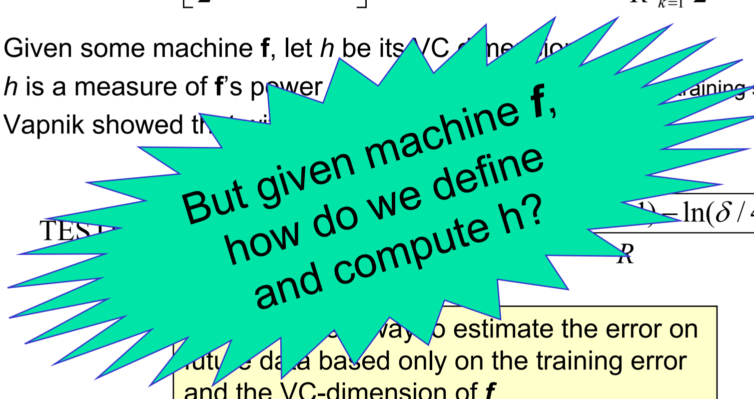
$$\text{TESTERR}(\alpha) \leq \text{TRAINERR}(\alpha) + \sqrt{\frac{h(\ln(2R/h)+1) - \ln(\delta/4)}{R}}$$

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

Vapnik-Chervonenkis dimension

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

- Given some machine **f**, let *h* be its
- *h* is a measure of **f**'s power … set)
- Vapnik showed th



$$\ldots - \ln(\delta/4) \quad R$$

… estimate the error on … based only on the training error and the VC-dimension of ***f***

Shattering

- Machine f can *shatter* a set of points $x_1, x_2 .. x_r$ if and only if…

  For every possible training set of the form (x_1,y_1) , (x_2,y_2) ,… (x_r ,y_r)

  …There exists some value of α that gets zero training error.

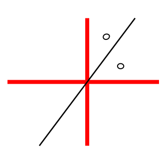
There are 2^r such training sets to consider, each with a different combination of +1's and –1's for the y's

Shattering

- Machine f can *shatter* a set of points $x_1, x_2 .. x_r$ if and only if…

  For every possible training set of the form (x_1,y_1) , (x_2,y_2) ,… (x_r ,y_r)

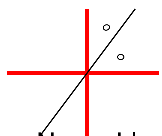
  …There exists some value of α that gets zero training error.
- Question: Can the following f shatter the following points?



$f(x,w) = sgn(w \cdot x)$

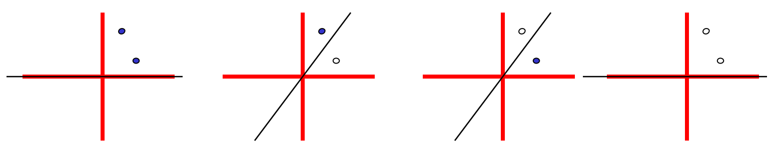
Shattering

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



f(x,w) = sgn(w•x)

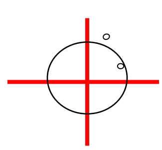
- Answer: No problem. There are four training sets to consider



w=(0,1)   w=(-2,3)   w=(2,-3)   w=(0,-1)

Shattering

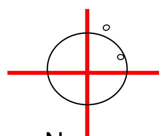
- Machine f can *shatter* a set of points $x_1, x_2 .. x_r$ if and only if…
  For every possible training set of the form (x_1,y_1) , (x_2,y_2) ,… (x_r ,y_r)
  …There exists some value of α that gets zero training error.
- Question: Can the following f shatter the following points?



f(x,b) = sgn(x•x-b)

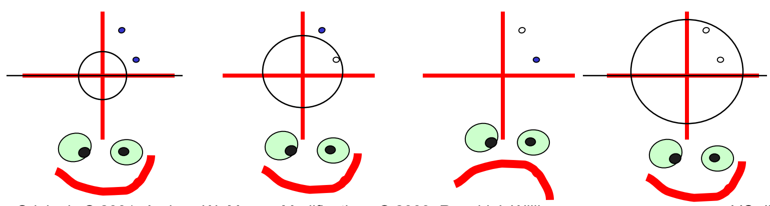
Shattering

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



$f(x,b) = sgn(x \bullet x-b)$

- Answer: No way my friend.



Definition of VC dimension

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

The maximum number of points that can be arranged so that ***f*** shatters them.

If any number of points can be shattered by ***f***, VC-dimension = +∞.

Definition of VC dimension

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

The maximum number of points that can be arranged so that ***f*** shatters them.

Example: What's VC dimension of f(x,b) = sgn(x•x-b)

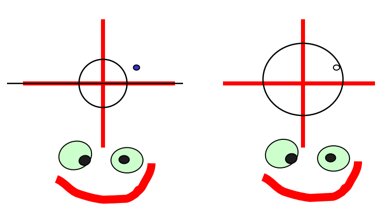
VC dim of trivial circle

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

The maximum number of points that can be arranged so that ***f*** shatters them.

Example: What's VC dimension of f(x,b) = sgn(x•x-b)

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



Reformulated circle

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

The maximum number of points that can be arranged so that ***f*** shatters them.

Example: For 2-d inputs, what's VC dimension of f(x,q,b) = sgn(qx•x-b)

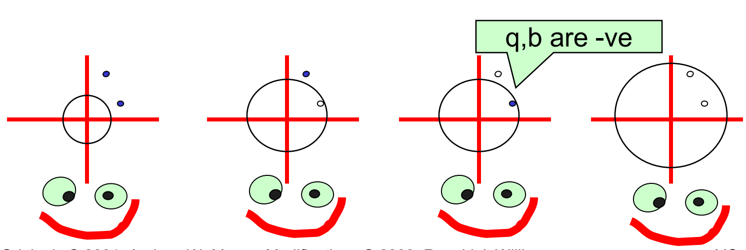
Reformulated circle

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

The maximum number of points that can be arranged so that ***f*** shatters them.

Example: What's VC dimension of f(x,q,b) = sgn(qx•x-b)

- Answer = 2

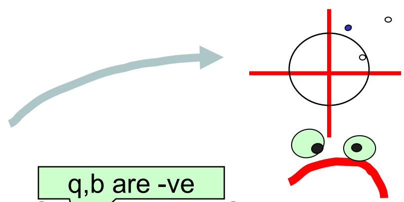


Reformulated circle

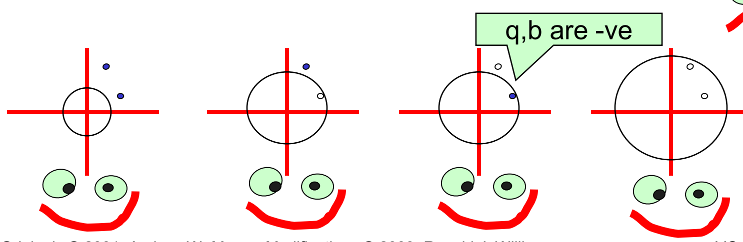
Given machine ***f***, the VC-dimension *h* is

> The maximum number of points that can be arranged so that ***f*** shatters them.

Example: What's VC dimension of f(x,q,b) = sgn(qx•x-b)



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



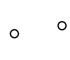
VC dim of separating line

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

> The maximum number of points that can be arranged so that ***f*** shatters them.

Example: For 2-d inputs, what's VC-dim of f(x,**w**,b) = sgn(**w**•x+b)?

Well, can f shatter these three points?



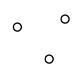
VC dim of line machine

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

> The maximum number of points that can be arranged so that ***f*** shatters them.

Example: For 2-d inputs, what's VC-dim of f(x,**w**,b) = sgn(**w**•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.

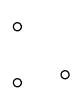
VC dim of line machine

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

> The maximum number of points that can be arranged so that ***f*** shatters them.

Example: For 2-d inputs, what's VC-dim of f(x,**w**,b) = sgn(**w**•x+b)?

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



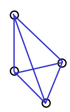
VC dim of line machine

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

The maximum number of points that can be arranged so that ***f*** shatters them.

Example: For 2-d inputs, what's VC-dim of f(x,**w**,b) = sgn(**w**•x+b)?

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



Can always draw six lines between pairs of four points.

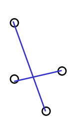
VC dim of line machine

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

The maximum number of points that can be arranged so that ***f*** shatters them.

Example: For 2-d inputs, what's VC-dim of f(x,**w**,b) = sgn(**w**•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.

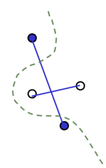
VC dim of line machine

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

> The maximum number of points that can be arranged so that ***f*** shatters them.

Example: For 2-d inputs, what's VC-dim of f(x,**w**,b) = sgn(**w**•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

VC dim of linear classifiers in m dimensions

If input space is m-dimensional and if **f** is *sgn(**w**•**x**-b)*, what is the VC-dimension?

Proof that *h* >= *m+1*: Show that there is a set of *m+1* points that can be shattered.

Can you guess how to construct such a set?

VC dim of linear classifiers in m dimensions

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

Proof that $h >= m+1$: Show that there is a set of $m+1$ points that can be shattered. Define m+1 input points thus:

$\mathbf{x}_0 = (0,0,0,\ldots,0)$
$\mathbf{x}_1 = (1,0,0,\ldots,0)$
$\mathbf{x}_2 = (0,1,0,\ldots,0)$
:
$\mathbf{x}_m = (0,0,0,\ldots,1)$ So $x_k[j] = 1$ if $k=j$ and 0 otherwise

Let $y_0, y_1, y_2, \ldots y_m$, be any one of the 2^{m+1} combinations of class labels (±1)

Guess how we can define $w_1, w_2, \ldots w_m$ and b to ensure $\mathrm{sgn}(\mathbf{w} \cdot \mathbf{x}_k + b) = y_k$ for all k. Note: $\mathrm{sgn}(\mathbf{w} \cdot \mathbf{x}_k + b) = \mathrm{sgn}\left(b + \sum_{j=1}^{m} w_j x_k[j]\right)$

VC dim of linear classifiers in m dimensions

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

Proof that $h >= m+1$: Show that there is a set of $m+1$ points that can be shattered. Define m+1 input points thus:

$\mathbf{x}_0 = (0,0,0,\ldots,0)$
$\mathbf{x}_1 = (1,0,0,\ldots,0)$
$\mathbf{x}_2 = (0,1,0,\ldots,0)$
:
$\mathbf{x}_m = (0,0,0,\ldots,1)$ So $x_k[j] = 1$ if $k=j$ and 0 otherwise

Let $y_0, y_1, y_2, \ldots y_m$, be any one of the 2^{m+1} combinations of class labels (±1)

Guess how we can define $w_1, w_2, \ldots w_m$ and b to ensure $\mathrm{sgn}(\mathbf{w} \cdot \mathbf{x}_k + b) = y_k$ for all k. Note: $\mathrm{sgn}(\mathbf{w} \cdot \mathbf{x}_k + b) = \mathrm{sgn}\left(b + \sum_{j=1}^{m} w_j x_k[j]\right)$

Answer: $b=y_0/2$ and $w_k = y_k$ for all k.

VC dim of linear classifiers in m dimensions

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

Proof that $h >= m+1$: Show that there is a set of $m+1$ points that can be shattered. Define m+1 input points thus:

$\mathbf{x}_0 = (0,0,0,\dots,0)$
$\mathbf{x}_1 = (1,0,0,\dots,0)$
$\mathbf{x}_2 = (0,1,0,\dots,0)$
:
$\mathbf{x}_m = (0,0,0,\dots,1)$ So $x_k[j] = 1$ if $k=j$ and 0 otherwise

Let $y_0, y_1, y_2, \dots y_m$, be any one of the 2^{m+1} combinations of class labels (±1)

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

Another answer: $b=y_0$ and $w_k = y_k - y_0$ for all k.

This is solution to the system of m+1 linear equations in m+1 unknowns obtained by dropping the sgn function.

VC dim of linear classifiers in m dimensions

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

- Now we know that h >= m+1
- In fact, h = m+1
- Proof that h < m+2 is a little more difficult
  - requires showing that no set of m+2 points can be shattered

Finite Hypothesis Spaces

- What's the relation to our earlier TESTERR analysis for finite hypothesis spaces?
- Suppose there are H hypotheses.
- There are 2^n different labellings of n points.
- Thus if VC-dim = h, there must be at least 2^h different hypotheses in the hypothesis space.
- Thus $2^h \le H$.
- Therefore VC-dimension satisfies

  $$h \le \log_2 H$$

  for any hypothesis space of size H.
- Can plug this into TESTERR bound formulas where appropriate.

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.*

Structural Risk Minimization

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

$$\mathrm{TESTERR}(\alpha) \le \mathrm{TRAINERR}(\alpha) + \sqrt{\frac{h(\ln(2R/h)+1) - \ln(\delta/4)}{R}}$$

i	f_i	TRAINERR	VC-Confidence	Probable upper bound on TESTERR	Choice
1	f_1				
2	f_2				
3	f_3				⌦
4	f_4				
5	f_5				
6	f_6				

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 to choose the best learning machine

Alternatives to VC-dim-based model selection

- What could we do instead?
  1. Cross-validation

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

Alternatives to VC-dim-based model selection

- What could we do instead?
  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

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

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				

Alternatives to VC-dim-based model selection

- What could we do instead?
    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.

*Another million caveats

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

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				

Which model selection method is best?

1. (CV) Cross-validation
2. AIC (Akaike Information Criterion)
3. BIC (Bayesian Information Criterion)
4. (SRMVC) Structural Risk Minimization 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.

Extra Comments

- Beware: that second “VC-confidence” term is usually very very conservative (at least hundreds of times larger than the empirical overfitting effect).

  Why?
  - Because the analysis is distribution-free
  - Because the analysis considers other worst-case aspects that easily may not apply to the particular hypothesis we’ve learned
- An excellent tutorial on VC-dimension and Support Vector Machines:

  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

Extra Comments

- Beware: that second “VC-confidence” term is usually very very conservative (at least hundreds of times larger than the empirical overfitting effect).

  Why?
  - Because the analysis is distribution-free
  - Because the analysis considers other worst-case aspects that easily may not apply to the particular hypothesis we’ve learned
- An excellent tutorial on VC-dimension and Support Vector Machines:

  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

Coming Attraction

What you should know

- The definition of a learning machine: $\boldsymbol{f}(\boldsymbol{x},\alpha)$
- 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