Computational Learning Theory

10-701/15-781, Recitation

March 25, 2010

Ni Lao

What's Computational Learning Theory?

- Laws about whether we can perform learning successfully or not
  - Instead of relying purely on empirical knowledge, our skills in probability can help
- Often in the form of the following question
  - With a family of models *H* of certain complexity, how many training samples *R* is needed in order to learn a model *h* with reasonable training time and sufficient accuracy on future data?
- Major components
  - Model complexity
    - Num. of parameters? Size of hypothesis space? VC-dimension?
  - Sample complexity
  - Error rate
  - Time complexity

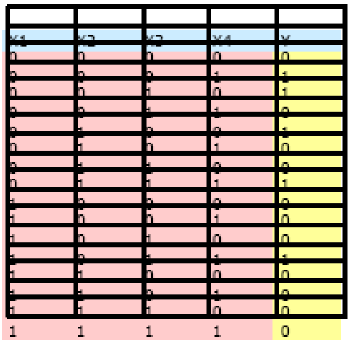
What We Have Learnt in Class

- For categorical inputs
  - PAC Learning
    - (Probably Approximately Correct Learning)
    - All inputs and outputs are binary→ easy to measure $|H|$
    - Data is noiseless → easy to analyze

- For continuous inputs
  - VC dimension
    - a hypothesis family H can shatter a set of points $x_1, x_2 .. x_r$, iff for every possible label $y_1, y_2 .. y_r$ (2^r of them), there exists some hypothesis h in H that can gets zero training error
    - VC(H) is the maximum number of points that can be shattered by H

Example: PAC Learning of Boolean Functions

- Chose number of samples R such that with probability less than δ we'll select a bad hypothesis (which makes mistakes more than fraction ε of the time)

Machine	Example Hypothesis	$\lvert H \rvert$	R required to PAC-learn
And-positive-literals	X3 ^ X7 ^ X8	2^m	$\frac{0.69}{\varepsilon}\left(m+\log_2\frac{1}{\delta}\right)$
And-literals	X3 ^ ~X7	3^m	$\frac{0.69}{\varepsilon}\left((\log_2 3)m+\log_2\frac{1}{\delta}\right)$
Lookup Table		2^{2^m}	$\frac{0.69}{\varepsilon}\left(2^m+\log_2\frac{1}{\delta}\right)$
And-lits or And-lits	(X1 ^ X5) v (X2 ^ ~X7 ^ X8)	$(3^m)^2 = 3^{2m}$	$\frac{0.69}{\varepsilon}\left((2\log_2 3)m+\log_2\frac{1}{\delta}\right)$

Disjunctive Normal Form (DNF)

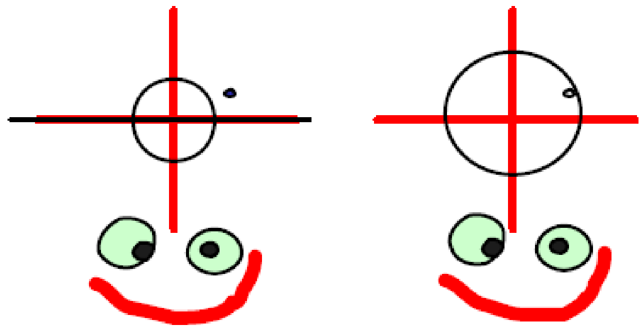
$R > a \log_2(\lvert H \rvert) + b$

2001, Andrew W. Moore

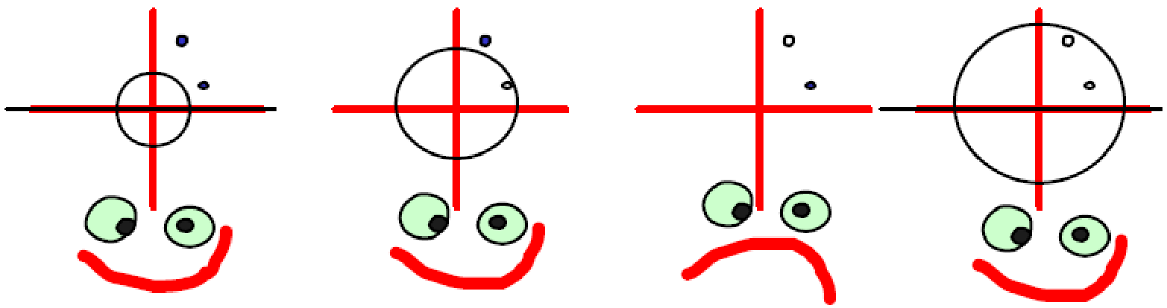
Example: VCd of Circle Hypothesis

- H={f(x,b) = sign(x.x-b)}, VC(H)=?

- N=1



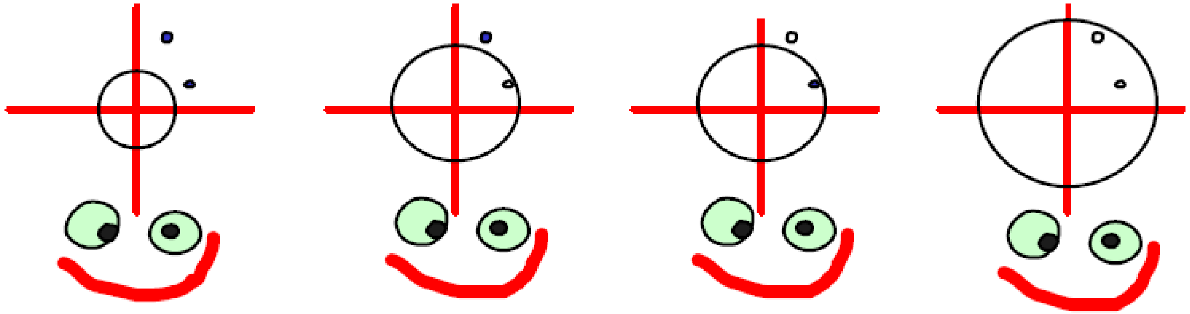
- N=2



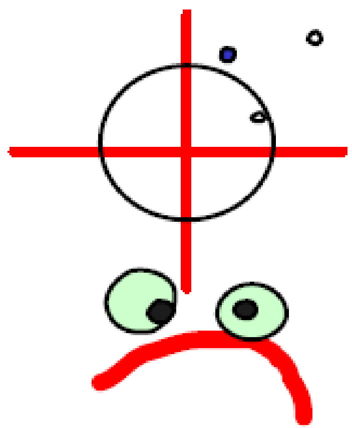
Example: VCd of Circle Hypothesis

- $H=\{f(x,a,b) = \mathrm{sign}(ax.x-b)\}$, $VC(H)=?$

- N=2



- N=3



Often
VC(H)= No. Parameter

Homework 4

- VCd of Gaussian Bayes Models
  - Practice your VCd finding skills, in two class classification problems

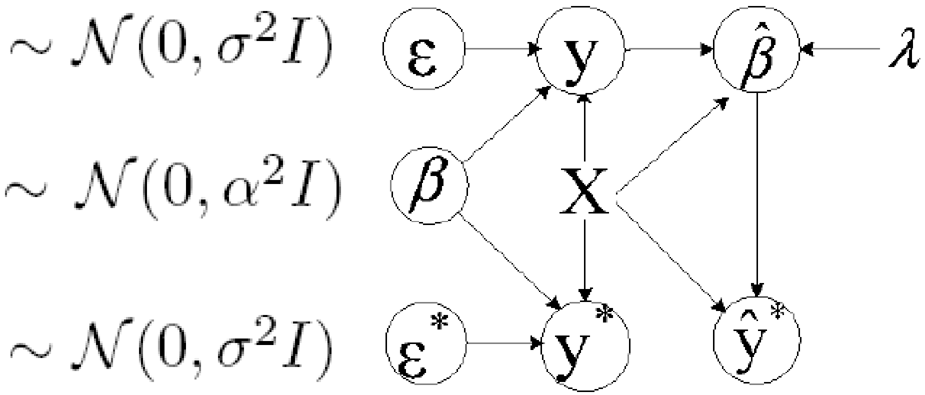
ID	**a**	**b**	**c**	**d**	**e**	**f**
No. features	1	1	2	2	2	2
Shared Covariance Matrix?	Y	N	Y	Y	N	N
Naive Bayes?	-	-	Y	N	Y	N
No. parameters						
VC dimension						

Policy: number of parameters is 0.5pt each. VC dimension is 2pt each, and you get $\max(0, 2 - d_{best} + d_{your})$pt, where d_{best} is the best bound I know, and d_{your} is your answer. You need to convince me in order to get credit for the VC dimension, but you need not give a formal proof.

Hint: think about what kind of decision boundary we get in each of the models.

Homework 4

- Linear Regression Model
  - express the average risk R(λ)/n for linear regression (λ=0) as a function of #features p and #samples n



- Result from hw3 (slightly revised)

$$R(\lambda) = E[e(\lambda)^T e(\lambda)]$$

- $$= \sum_{i=1..p} \left[\left(\frac{\lambda d_i}{d_i^2 + \lambda} \right)^2 \alpha^2 + \left(\frac{d_i^2}{d_i^2 + \lambda} \right)^2 \sigma^2 \right] + \sum_{i=1..n} \sigma^2$$

Summary of Model Selection Methods

- VC dimension (Structural Risk Minimization)
  - Very conservative

- AIC (Akaike Information Criterion)
  - Asymptotically the same as Leave-one-out CV

- BIC (Bayesian Information Criterion)
  - Asymptotically the same as a carefully chosen k-fold CV

- (CV) Cross-validation
  - The ultimate weapon used by most people who apply ML techniques

- The End
- Thanks