

Intro. to machine learning (CSI 5325)

Lecture 16: Learning theory

Greg Hamerly

Spring 2008

Some content from Tom Mitchell.

- 1 Computational learning theory
- 2 Probably approximately correct (PAC) learning
- 3 Vapnik-Chervonenkis dimension

Computational learning theory

What general laws constrain inductive learning?

We seek theory to relate:

- Probability of successful learning
- Number of training examples
- Complexity of hypothesis space
- Accuracy to which target concept is approximated
- Manner in which training examples presented

PAC Learning

Consider a class C of possible target concepts defined over a set of instances X of length n , and a learner L using hypothesis space H .

Definition: C is **PAC-learnable** by L using H if

- for all $c \in C$, distributions \mathcal{D} over X , ϵ such that $0 < \epsilon < 1/2$, and δ such that $0 < \delta < 1/2$,
- learner L will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $\text{error}_{\mathcal{D}}(h) \leq \epsilon$, in time that is polynomial in $1/\epsilon$, $1/\delta$, n and $\text{size}(c)$.

Unbiased learning

Given: sample space X , class of concepts C , hypothesis space $H = C$, and n boolean features for each example.

- $|C| = 2^{|X|}$
- $|X| = 2^n$
- $|H| = |C| = 2^{2^n}$

Therefore, sample complexity is:

$$m \geq \frac{1}{\epsilon} (2^n \ln 2 + \ln(1/\delta))$$

Thus an unbiased hypothesis space is *not* PAC-learnable, since it has an exponential dependence on n .

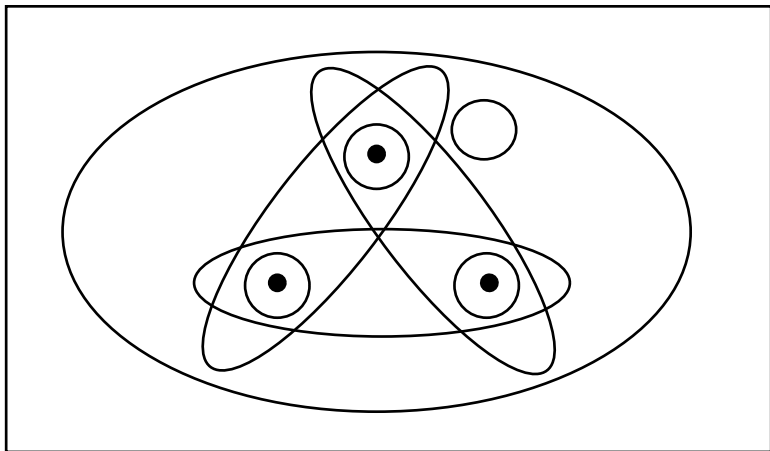
Shattering a set of instances

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.

Three instances shattered

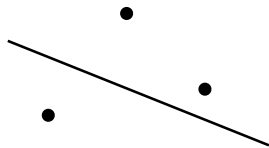
Instance space X



The Vapnik-Chervonenkis 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 $VC(H) \equiv \infty$.

VC dimension of linear decision surfaces

*(a)**(b)*

Sample Complexity from VC Dimension

How many randomly drawn examples suffice to ϵ -exhaust $VS_{H,D}$ with probability at least $(1 - \delta)$?

$$m \geq \frac{1}{\epsilon} (4 \log_2(2/\delta) + 8VC(H) \log_2(13/\epsilon))$$

Compare VC and PAC sample complexity

PAC (assuming consistent learner):

$$m \geq \frac{1}{\epsilon} (\ln(1/\delta) + \ln |H|)$$

VC:

$$m \geq \frac{1}{\epsilon} (4 \log_2(2/\delta) + 8VC(H) \log_2(13/\epsilon))$$

Compare VC dimension and hypothesis space size

$$VC(H) \leq \log_2 |H|$$

Why?

- assume $VC(H) = d$
- H requires at least 2^d distinct hypotheses to shatter d instances
- thus, $2^d \leq |H|$, and $d = VC(H) \leq \log_2 |H|$

Examples of VC dimension

	VC dimension
hyperplane (perceptron) in d dimensions	$d + 1$
axis-aligned rectangle in the $2-d$ plane	4
convex k -sided polygon in $2-d$ plane	$2k + 1$
CNF of up to n boolean literals	n
k -term DNF over n variables	$\leq nk \log_2 3$
multi-layer perceptron	?

VC Dimension of multi-layer perceptron network

Given a multilayer perceptron G with:

- n input nodes
- $s \geq 2$ internal (non-input) nodes
- at most r inputs to any internal node
- concept class C over \mathbb{R}^r with $VC(C) = d$

Then:

$$VC(G) \leq 2ds \log(es)$$

where $e = 2.71828 \dots$. Note: $d = r + 1$ for linear perceptron.

Limitations:

- applies only to thresholded linear perceptrons;
- does not apply to sigmoid-based networks trained with backprop – those have VC dimension at least this big (why?)
- doesn't account for backprop's training method – small-to-large weights

VC dimension and number of parameters

Note that the earlier bound of PAC learning depended on $|H|$, the size of the hypothesis space.

Further, $|H|$ typically depends on the number of parameters (d) in the model.

- often $|H|$ has an exponential relationship with d
- this leads to the 'curse of dimensionality', which is most evident in bias-free learning
 - we saw this in the sample complexity of the n -term bias-free boolean learner

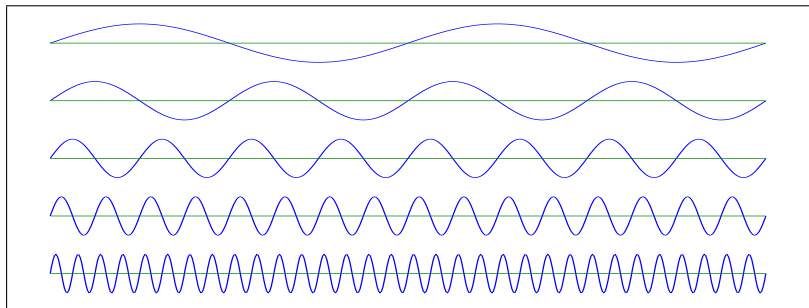
However, the VC dimension often may not depend so directly on d .

Classes with infinite VC dimension

The following family of classifiers has an infinite VC dimension. . .

$$h(x) = \sin(\alpha x) \geq 0$$

. . . but a finite number of parameters (only one!).



Infinite hypothesis spaces and VC dimension

Note that just because a hypothesis space is infinite does not imply that the VC dimension is infinite

- e.g. the interval classifier $h(x) = a < x < b$
 - $|H| = \infty$, but $VC = 2$

However, $VC = \infty$ does imply that $|H| = \infty$

Both of these results are implied by $VC(H) \leq \log_2 |H|$