VC-dimension definition

Defn: A set of points S is **shattered** by a concept class C if there are concepts in C that split S in all of the $2^{|S|}$ possible ways.

In other words, all possible ways of classifying points in S are acheivable using concepts in C.

E.g., any 3 non-collinear points can be shattered by linear threshold functions in 2-D.

VC-dimension definition, contd.

The **VC-dimension** of a concept class C is the size of the largest set of points that can be shattered by C.

> So, if the VC-dimension is d, that means *there exists* a set of d points that can be shattered, but there is *no* set of $d + 1$ points that can be shattered.

E.g., VC-dim(linear threshold fns in 2-D) = 3.

Examples

What's the VC-dimension of:

$C = \{$intervals of real line$\}$?

$C = \{$axis-parallel boxes in 2-D$\}$?

Examples, contd

$C = \{$monotone disjunctions (n features)$\}$?

$C = \{$all functions on n features$\}$?

How does VC-dimension measure complexity?

Define $C[m]$ = maximum number of ways to split m points using concepts in C. E.g., for $m \leq VCdim(C)$, $C[m] = 2^m$.

Theorem [VC]: $C[m] = O(m^{VCdim(C)})$.

I.e., only $C[m]$ "different" concepts in C with respect to the m examples.

Suggests maybe can replace $\log |C|$ by $\log C[m]$ in PAC bounds, but NOT OBVIOUS, since *which concepts they are* depends on the data.

Upper,lower bound theorems

- **Theorem 1:** Given class C, for any distribution D, target concept in C, if you get:

  $$\frac{k}{\epsilon}[VCdim(C)\log(1/\epsilon) + \log(1/\delta)] \text{ examples,}$$

  then with prob $\geq 1 - \delta$, **all** consistent $h \in C$ have error $\leq \epsilon$. ($k \approx 8$)

- **Theorem 2:** If $|C| > 2$, then for any learning alg A, there exists a distribution D, and distribution on target concepts in C such that expected error of A is greater than ϵ if A sees less than

  $$\frac{VCdim(C)-1}{8\epsilon}$$

  examples.

- Does this mean you can't learn classes of infinite VC-dimension? No: since allowed poly dependence on $size(c)$.

- Does this mean "complicated hypotheses are bad"? No. Just means you need a lot of data to learn complicated things.