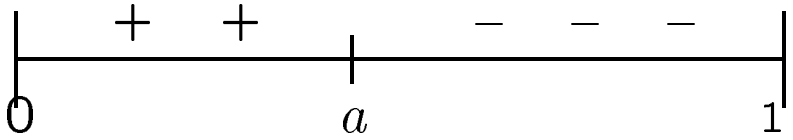
Goal: bound # training examples needed

- PAC bound: if target $c \in H$, then sufficient to see only $\frac{1}{\varepsilon}[\ln(|H|) + \ln(\frac{1}{\delta})]$ examples.
- Hoeffding: if target $c \notin H$, replace ε by $2\varepsilon^2$.

Not so good for "initial subintervals."



Problem: $|H|$ is infinite, but not that many "really different" subintervals. (E.g., if m points, then only $m + 1$ different ways to classify.)

How to measure complexity?

One useful way $\rightarrow$ "Vapnik-Chervonenkis dimension."

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

$C = \{$disjunctions of subsets of n features$\}$?

$C = \{$conjunctions of subsets of 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]: $C[m] = O(m^{VCdim(C)})$.

Also, can almost replace $\log |H|$ with VC-dimension in PAC bounds.

> (NOT OBVIOUS, since although only $C[m]$ "different" concepts in C with respect to the m examples, *which ones they are* depends on data.)

Note: always have $VCdim(C) \leq \log_2(|H|)$.

Upper,lower bound theorems

- **Theorem 1:** If C is "well behaved" then for any distribution D, target concept in C, if you get:

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

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

  This is called "uniform convergence."

- **Theorem 2:** If $|C| > 2$, then for any learning alg A, there exists a distribution D, and a target concept in C such that if A sees less than

  $$max\left[\frac{1}{\varepsilon}\log(1/\delta), \frac{VCdim(C)-1}{32\varepsilon}\right] \text{ examples,}$$

  then with prob $\geq 1-\delta$, its error is $\geq \varepsilon$.

References:

Pollard, *Convergence of Stochastic Processes*.

Blumer, Ehrenfeucht, Haussler, Warmuth: "Learnability and the V.C. dimension." *JACM* 6(4), 1989.