Agnostic Learning

Thus far, $c \in H$ is assumed to achieve zero training error of hypotheses. In this setting, $c \in H$ is not assumed. So, agnostic learner L selects hypothesis h^* with minimum training error

How many training examples suffice to guarantee that $error_Q(h^*) < error_D(h^*) + \epsilon$ with probability of at least $1 - \delta$?

1. $P(error_Q(h) \geq error_D(h) + \epsilon) \leq \exp(-2|D|\epsilon^2)$, by Hoeffding's inequality
2. $P(\exists h \in H \;\; error_Q(h) \geq error_D(h) + \epsilon) \leq |H| \exp(-2|D|\epsilon^2)$, by union bound
3. To determine the no. $|D|$ of training examples required to reduce this probability to be at most δ, $|H| \exp(-2|D|\epsilon^2) \leq \delta$.
4. Then, $|D| \geq (1/(2\epsilon^2)) (\ln |H| + \ln (1/\delta))$.

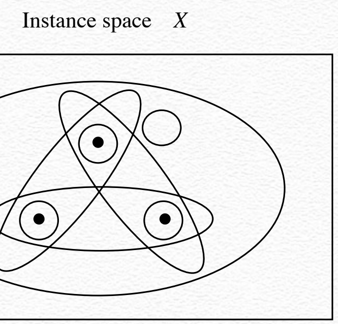
Shattering a Set of Instances

Definition. A **dichotomy** $(Y, S \setminus Y)$ of a set S is a partition of S into two disjoint subsets $Y \in 2^S$ and $S \setminus Y$.

Definition. A hypothesis $h \in H$ is **consistent** with a dichotomy $(Y, S \setminus Y)$ of a set S of input instances iff

$$(\forall x \in Y \;\; h(x) = 1) \wedge (\forall x \in S \setminus Y \;\; h(x) = 0) \; .$$

Definition. A set of input instances $S \subseteq X$ is **shattered** by hypothesis space H iff for every dichotomy of S, there exists some hypothesis in H that is consistent with this dichotomy.



Vapnik-Chervonenkis (VC) Dimension

Can H shatter a large subset S of X? Larger S implies more expressive H.

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 $VC(H) = \infty$.

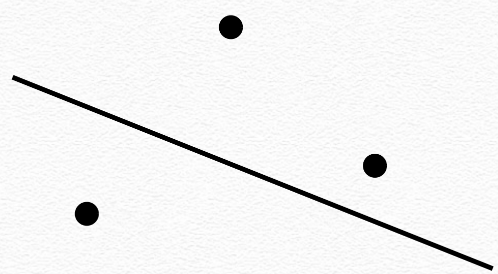
Proposition 1. For any finite H, $VC(H) \leq \log_2 |H|$.

Implication. Sample complexity can be potentially reduced if the $\ln |H|$ term can be replaced by $VC(H)$. See Theorem 3 later.

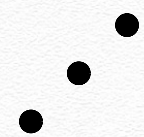
Vapnik-Chervonenkis (VC) Dimension

Example 1. $X = \mathbb{R}$, H = real intervals (a, b) where $a, b \in \mathbb{R}$.

Example 2. $X = x, y$ plane, H = linear decision surfaces (recall hypothesis space of perceptron unit with 2 inputs)



(*a*)



(*b*)

How many training examples will ϵ-exhaust $VS_{H,D}$ using VC Dimension?

Theorem 3 (Blumer et al. 1989). Let $0 < \epsilon, \delta \leq 1$. If D is a set of independent random examples of some target concept c s.t. $|D| \geq (1/\epsilon)\,(8\ VC(H) \log_2 (13/\epsilon) + 4 \log_2 (2/\delta))$, then the probability that $VS_{H,D}$ is ϵ-**exhausted** (w.r.t. c) is at least $1 - \delta$:

$$P(\forall h \in VS_{H,D}\ \ error_Q(h) < \epsilon) \geq 1 - \delta\ .$$