

ECE695 Lecture notes 7

Shai Ben David

December 2, 2003

Goal. Prove that that $VC \dim(H)$ controls the sample size needed to guarantee that

$$\forall P \forall h \in H |Er^S(h) - Er^P(h)|$$

is small. Towards proving upper bounds on the sample complexity we considered a different problem: ϵ -nets, ϵ -approximation. We proved: if H has $VC \dim = d (< \infty)$ ($H \subseteq 2^X$) then for every probability distribution P over X , any $\epsilon \geq \delta > 0$, if a sample S consists of m i.i.d. P - generated points then with probability $> 1 - \delta$ S is an ϵ -net once:

$$m > \frac{c}{\epsilon} (d + \log \frac{1}{\delta})$$

A similar theorem holds for ϵ -approximation, only now we need

$$m > \frac{c}{\epsilon^2} (d + \log \frac{1}{\delta})$$

Recall: For a given $H \subseteq 2^X$ and P probability distribution over X , $S \subseteq X$ is an ϵ -approximation with respect to H, P if

$$\forall h \in H \left| \frac{|S \cap h|}{|S|} - P(h) \right| < \epsilon.$$

Theorem. (Basic theorem of Statistical Learning Theory): Let $H \subseteq 2^X$ have a finite $VC \dim d$, then for every probability distribution P over $X \times \{0, 1\}$, $\forall \epsilon > 0, \delta > 0$, if S is an i.i.d. P sample of size bigger than $\frac{c}{\epsilon} (d + \log \frac{1}{\delta})$, then with probability $> 1 - \delta, \forall h \in H$

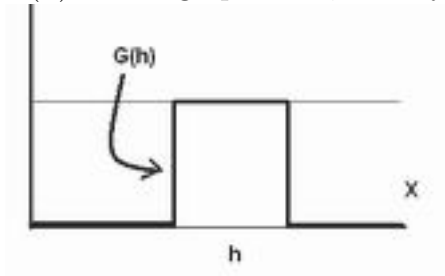
$$|E^S(h) - E^P(h)| < \epsilon.$$

Alternatively, for every m , if S is an m -size P -sample then

$$\forall h \in H |E^S(h) - E^P(h)| < c \sqrt{\frac{d + \log \frac{1}{\delta}}{m}}$$

where c is a constant.

Proof. Given a class $H \subseteq 2^X$ we define a class $G(H) \subseteq 2^{X \times \{0,1\}}$. For every $h \in H$ let $G(h)$ be the graph of h , namely $G(h) = \{(x, y) | h(x) = y\}$.



$$G(H) =^{def} \{G(h) | h \in H\}$$

$$G(H) \subseteq 2^{X \times \{0,1\}}$$

Claim. If a sample S is an $(G(H)^C, P)\epsilon$ -approximation for $G(H)^C =^{def} \{X \times \{0,1\} \setminus G(h) | h \in H\}$ then

$$\forall h \in H \quad |E^S(h) - E^P(h)| < \epsilon.$$

Proof.

$$E^S(h) = \frac{|\{(x, y) \in S | h(x) \neq y\}|}{|S|} = \frac{|S \cap G(h)^C|}{|S|}$$

$$E^P(h) = P(\{(x, y) : h(x) \neq y\}) = P(G(h)^C)$$

To complete the proof of the learning theorem, we need to show that

$$\forall H \subseteq 2^X \quad VC \dim(H) = VC \dim(G(H)^C)$$

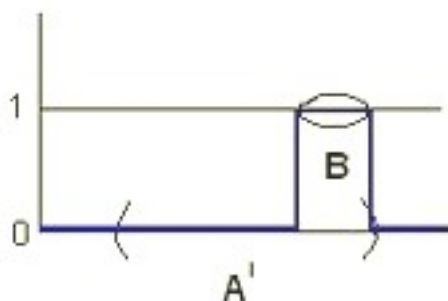
Claim. $\forall H \subseteq 2^Y$, let $H^C = \{Y \setminus h | h \in H\}$. Then we have

$$VC \dim(H) = VC \dim(H^C).$$

Proof. A set $A \subseteq Y$ is shattered by H iff it is shattered by H^C .

Claim. $\forall H \subseteq 2^X$, defining $G(H) \in 2^{X \times \{0,1\}}$ as we did, we have $VC \dim(H) = VC \dim(G(H))$.

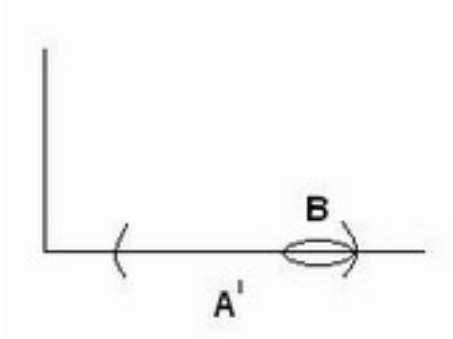
Proof. (a) Let $A \subseteq X$ be shattered by H . Then $G(H)$ shatters $A \times \{1\}$. For every $B \subseteq A \times \{1\} \exists B^1 \subseteq A$ such that $B = B^1 \times \{1\}$. So we pick h for which $A \cap h = B^1$ and get $A \times \{1\} \cap G(h) = B$



(b) Let $A \in X \times \{0,1\}$ be shattered by $G(H)$. We should find some $A^1 \in X$ shattered by H and $|A^1| = |A|$. Pick

$$A^1 = \{x \in X | \exists y (x, y) \in A\}$$

(i) If $G(H)$ shatters A , then H shatters A^1 .



Given $B^1 \in A^1$ let $B = \{(x, y) \in A | x \in B^1\}$ if $G(h) \cap A = B$ then $h \cap A^1 = B^1$.

(ii) We want to show that $|A^1| = |A|$. It suffices to see that if A is shattered by some $G(H)$ then there is no x for which both $(x, 0)$ is in A and $(x, 1)$ is in A ; this is true since $\nexists h$ such that it could not include both $(x, 0)$ and $(x, 1)$ \square

Big Picture

Let us fix a set of potential predicting rules H :

$$\forall h \in H \quad E^P(h) \leq E^S(h) + c \sqrt{\frac{d + \log \frac{1}{\delta}}{m}}$$

where the first term of the sum represents the training error and the second is the generalization error. \square

