Lecture 5

* HW 1 posted. Due in 3 weeks.
* More details regarding presentations in ≤ next week.
* Scribing?
* Last time : Concentration inequalities.

* Today : VC dimension

Vapnik-Chervonenkis Dimension

Recap: For finite $\mathcal{H}$,

→ PAC learn with $O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}\right)$ samples

→ agnostic-PAC learning with $O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon^2}\right)$ samples

How about infinite $\mathcal{H}$?

→ Discretization: Linear classifier in $\mathbb{R}^d$

32 bit system → $(2^{32})^d$ possible classifiers

→ VC dimension gives a nice way to handle ∞ classes.

Shattering

Def. (Restriction & Shattering)

The restriction of class $\mathcal{H}$ to a set of examples $C = \{c_1, \ldots, c_n\} \in \mathcal{X}$ is a subset of $\{0,1\}^{|C|}$ given by $\mathcal{H}_C = \{(h(c_1), \ldots, h(c_n)) : \forall h \in \mathcal{H}\}$

We say that $\mathcal{H}$ shatters C if $|\mathcal{H}_C| = 2^{|C|}$.

Shattering => all possible labellings are realized.

Corollary of No free lunch theorem.

Let $\mathcal{H}$ be a hypothesis class & assume there exists a set $C \subseteq X$ of size $2n$ s.t. $\mathcal{H}$ shatters C. Then $\exists$ a distribution D over $X \times \{0,1\}$ & a predictor $h^* \in \mathcal{H}$ s.t. $R(h^*) = 0$, but for any learning algorithm A, $\Pr_{S \sim D^n}[R(A(S)) \geq \frac{1}{8}] \geq \frac{1}{7}$.

$\mathcal{H}$ shatters a set of size $2n$ =>
cannot learn with n examples.

But if C is s.t. $|\mathcal{H}_C| << 2^{|C|}$ maybe can do something?

Idea: for any distribution supported on C, real hypothesis space under consideration is $\mathcal{H}_C$.

$\mathcal{H}_C$ is finite!

If $\mathcal{H}_C$ is small, then maybe can learn.

Def (VC dimension)

The VC dimension of a hypothesis class $\mathcal{H}$, denoted by $VCdim(\mathcal{H})$ is the size of the largest set $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitary size, then $VCdim(\mathcal{H}) = \infty$.

Examples

To show $VCdim(\mathcal{H}) = d$.

1. There exists some set C of size d that can be shattered by $\mathcal{H}$.
2. No set of size $d+1$ is shattered by $\mathcal{H}$.

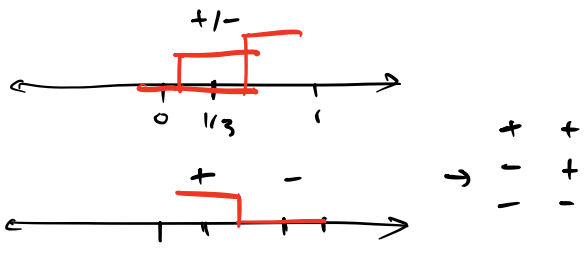
Example 1

Threshold functions:

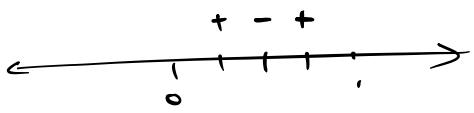
Let $\mathcal{X} = [0,1]$, $\mathcal{H} = \{ h_b(x) = \mathbb{1}(x \geq b), b \in [0,1] \}$

$\mathcal{H}$: set of thresholds in $\mathbb{R}$.

$VCdim(\mathcal{H}) = 1$



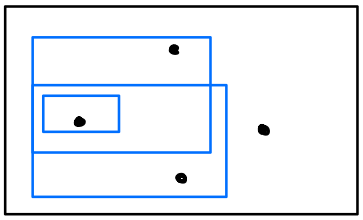
If we also allow reverse i.e. $\mathbb{1}(x < b)$, $VCdim(\mathcal{H}) = 2$.
(previous $\mathcal{H}$ U new stuff)

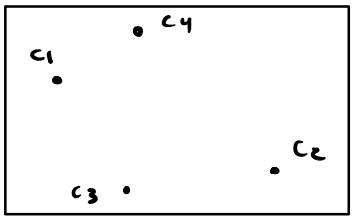


Example 2

Axis-aligned rectangles

$$\mathcal{X} = \mathbb{R}^2, \quad \mathcal{H}_{a_1,a_2,b_1,b_2}(x_1, x_2) = \mathbb{1}\begin{pmatrix} a_1 \le x_1 \le a_2 \ \& \\ b_1 \le x_2 \le b_2 \end{pmatrix}$$





- $\begin{cases} c_1: \min x_1 \\ c_2: \max x_1 \\ c_3: \min x_2 \\ c_4: \max x_2 \end{cases}$
- $\{c_5$: any other (not unique minimizer)

Cannot realize this label.

Example 3

Finite classes.

For any finite hypothesis class $\mathcal{H}$, have $VCdim(\mathcal{H}) \le \log(|\mathcal{H}|)$. This is because for any set C, $|\mathcal{H}_C| \le |\mathcal{H}|$. Therefore, if $2^{|C|} > |\mathcal{H}|$, then we cannot shatter C.

VC Theorem

Let $\mathcal{H}$ be a hypothesis class with $VCdim(\mathcal{H}) = d < \infty$.
Then there is a absolute constant $c > 0$ s.t.
$\mathcal{H}$ has uniform convergence property with,

$$n_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq c \cdot \frac{d \log(d/\epsilon) + \log(1/\delta)}{\epsilon^2}$$

Corollary.

$\mathcal{H}$ is agnostically-PAC learnable with
$O\left(\frac{d \log(d/\epsilon) + \log(1/\delta)}{\epsilon^2}\right)$ samples.

Note

1) Possible to show $n_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq c \cdot \frac{d + \log(1/\delta)}{\epsilon^2}$

2) This is for binary classification with 0/1 loss.
Will see how to generalize to other losses / beyond binary classification later.

→ for $d = \log(|\mathcal{H}|)$

$n_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq O\left(\frac{\log(|\mathcal{H}|/\delta)}{\epsilon^2}\right)$ Sample complexity

Proof

Outline: 1) for any set $C \subseteq X$, $\mathcal{H}_C$: "Effective size" of $\mathcal{H}$
$|\mathcal{H}_C| \simeq |C|^d$.

2) small "effective size" good for union bound to get UC.

Step 1: Polynomial growth of $\mathcal{H}_C$.

Definition (Growth function)

The growth function of $\mathcal{H}$, $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as

$$\tau_{\mathcal{H}}(n) = \max_{C \subseteq X, |C| = n} |\mathcal{H}_C|.$$

If $VCdim(\mathcal{H}) = d$, then $\tau_{\mathcal{H}}(n) = 2^n \quad \forall\ n \leq d$.

Sauer's lemma gives a good upper bound $\forall\ n \geq d$.

Lemma: (Sauer's Lemma)

$\forall n$, $VCdim(\mathcal{H}) = d$, $\tau_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}$.

For $n > d+1$, this implies $\tau_{\mathcal{H}}(n) \leq \left(\frac{ne}{d}\right)^d$.

Proof: We instead show stronger inequality.

For any $C = \{c_1, \dots, c_n\}$ & any $\mathcal{H}$.

$$|\mathcal{H}_C| \leq |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|. \quad -(1)$$

This is sufficient since if $VCdim(\mathcal{H}) = d$, $\mathcal{H}$ cannot shatter any set B of size $|B| > d$. There are $\binom{n}{i}$ subsets of size, we get our bound.

Prove (1) by induction.

Base case, $n = 1$

Either 1) $|\mathcal{H}_C| = 1$, LHS = RHS since one labelling, shatters $\{\phi\}$.

2) $|\mathcal{H}_C| = 2$, two labellings shatter $\{\{\phi\}, \{c_1\}\}$

Induction step.

Assume that (1) holds for all sets of size $k < n$.

Let $C = \{c_1, \dots, c_n\}$ & $C' = \{c_2, \dots, c_n\}$

Define $Y_0 = \{(y_2, \dots y_n) : (0, y_2, \dots, y_n) \in \mathcal{H}_C$ OR $(1, y_2, \dots, y_n) \in \mathcal{H}_C\}$

$Y_1 = \{(y_2, \dots, y_n) : (0, y_2, \dots, y_n) \in \mathcal{H}_C$ AND $(1, y_2, \dots, y_n) \in \mathcal{H}_C\}$

Claim : $|\mathcal{H}_C| = |Y_0| + |Y_1|$

True because $(y_2, \ldots, y_n)$ is counted once in Y_0, but again in Y_1 if can be shattered.

By induction we get,

$$|Y_0| \leq |\{B \subseteq C' : \mathcal{H} \text{ shatters } B\}|$$

$$= |\{B \subseteq C : c_1 \notin B \text{ AND } \mathcal{H} \text{ shatters } B\}|.$$

For Y_1, define $\mathcal{H}' \subseteq \mathcal{H}$ to be

$$\mathcal{H}' = \{h \in \mathcal{H}, \text{ if } h' \in \mathcal{H} \text{ s.t. } ((1-h'(c_1), h'(c_2), \ldots, h'(c_n)) = (h(c_1), h(c_2), \ldots, h(c_n))\}$$

$\mathcal{H}'$: all hypothesis for which the hypothesis that agrees everywhere in C except c_1 is also in $\mathcal{H}$.

Note : 1) If $\mathcal{H}'$ shatters $B \subseteq C'$ then it also shatters $B \cup \{c_1\}$.

2) $Y_1 = \mathcal{H}'_{C'}$.

$$|Y_1| = |\mathcal{H}'_{C'}| \leq |\{B \subseteq C' : \mathcal{H}' \text{ shatters } B\}| \quad (\text{By (1)})$$

$$= |\{B \subseteq C' : \mathcal{H}' \text{ shatters } B \cup \{c_1\}\}|$$

$$= |\{B \subseteq C : c_1 \in B \text{ AND } \mathcal{H}' \text{ shatters } B\}|$$

$$\leq |\{B \subseteq C : c_1 \in B \text{ AND } \mathcal{H} \text{ shatters } B\}|$$

$$\therefore |\mathcal{H}_C| = |Y_0| + |Y_1|$$

$$\leq |\{B \subseteq C : c_1 \notin B \text{ AND } \mathcal{H} \text{ shatters } B\}|$$
$$+ |\{B \subseteq C : c_1 \in B \text{ AND } \mathcal{H} \text{ shatters } B\}|$$
$$= |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|,$$

which completes our proof. ●

Step 2: Symmetrization

Lemma: For a class $\mathcal{H}$ with growth function $T_{\mathcal{H}}$, $\mathbb{E}_{S \sim D^n} \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)| \leq \sqrt{\frac{2\log(2T_{\mathcal{H}}(2n))}{n}}$.

Once we have this, we can Markov's inequality to get a high probability statement.

$$\Pr\left[\sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)| > t\right] \leq \frac{\sqrt{\frac{2\log(2T_{\mathcal{H}}(2n))}{n}}}{t}.$$

But we'll use McDiarmid's to get something better.