VC Dimension and PAC Learnability

Instructor: *Jess Sorrell* Scribe: *Jess Sorrell*

0.1 VC Dimension

What about infinite hypothesis classes? It turns out we can still PAC learn them, so long as they have finite *Vapnik-Chervonenkis (VC) dimension.*

Definition 0.1 (Shattering)**.** A hypothesis class $\mathcal{H}$ is said to *shatter* a set of points $C = \{x_1, \dots, x_m\} \subset \mathcal{X}$ if for every possible labeling $\{y_1, \dots, y_m\} \in \{0,1\}^m$, there exists a hypothesis $h \in \mathcal{H}$ such that $h(x_i) = y_i$ for all $i \in [m]$. That is, $\mathcal{H}$ can realize any of the 2^m possible dichotomies of C.

For example, consider the class of linear classifiers in $\mathbb{R}^2$. A linear classifier in $\mathbb{R}^d$ is a function $h : \mathbb{R}^d \to \{0,1\}$ of the form $h(x) = \mathbb{1}[w \cdot x + b \geq 0]$ for some $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. The decision boundary is the hyperplane $\{x : w \cdot x + b = 0\}$. In $\mathbb{R}^2$, this is a line. Any set of 3 non-collinear points can be shattered.

Proof. Let the three non-collinear points be $x_1, x_2, x_3 \in \mathbb{R}^2$. We must show that for any of the $2^3 = 8$ possible labelings, we can find a line that correctly classifies the points.

- **All points have the same label (e.g., (0,0,0) or (1,1,1)):** We can draw a line that does not pass through any point, leaving all points on one side. This line classifies all points as 0 (or 1, by flipping the decision rule).

- **One point has a different label from the other two (e.g., (1,0,0)):** Suppose x_1 is labeled 1 and x_2, x_3 are labeled 0. Since the points are non-collinear, they form a triangle. We can draw a line that separates the vertex x_1 from the side formed by the line segment connecting x_2 and x_3. This line correctly classifies the points. The same logic applies to the labelings (0,1,0), (0,0,1), and their complements (0,1,1), (1,0,1), (1,1,0).

Since all 8 dichotomies can be realized, the set of 3 non-collinear points is shattered. □

Definition 0.2 (VC Dimension [Vapnik and Chervonenkis, 1971])**.** The *VC dimension* of a hypothesis class $\mathcal{H}$, denoted $VC - dim(\mathcal{H})$, is the size of the largest set of points that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter arbitrarily large sets of points, its VC dimension is infinite.

Example: Threshold Functions The class of threshold functions on the real line, $\mathcal{H} = \{h_t : t \in \mathbb{R}\}$ where $h_t(x) = \mathbb{1}[x \geq t]$, has VC dimension 1.

Proof. To show the VC dimension is 1, we must show that there is a set of size 1 that can be shattered, and no set of size 2 can be shattered.

VC-dim($\mathcal{H}$) $\geq$ 1: Consider any single point set $C = \{x\}$. We can achieve the labeling $y = 1$ by choosing $t \leq x$, and the labeling $y = 0$ by choosing $t > x$. Since both labelings are possible, any set of size 1 can be shattered.

VC-dim($\mathcal{H}$) $<$ 2: Consider any set of two points $C = \{x_1, x_2\}$ with $x_1 < x_2$. There are $2^2 = 4$ possible labelings: $(0,0), (0,1), (1,1), (1,0)$.

- $(0,0)$: Choose $t > x_2$. Then $h_t(x_1) = 0$ and $h_t(x_2) = 0$.
- $(0,1)$: Choose t such that $x_1 < t \leq x_2$. Then $h_t(x_1) = 0$ and $h_t(x_2) = 1$.
- $(1,1)$: Choose $t \leq x_1$. Then $h_t(x_1) = 1$ and $h_t(x_2) = 1$.
- $(1,0)$: This would require $h_t(x_1) = 1$ and $h_t(x_2) = 0$. This implies $x_1 \geq t$ and $x_2 < t$, which means $x_1 > x_2$. This contradicts our assumption that $x_1 < x_2$.

Since the labeling $(1,0)$ cannot be realized, no set of size 2 can be shattered.

Therefore, the VC dimension of the class of threshold functions is 1. □

For example, the class of linear classifiers (hyperplanes) in $\mathbb{R}^d$ has VC dimension $d + 1$.

The key result connecting VC dimension to PAC learnability is the following theorem, which shows that finite VC dimension is a necessary and sufficient condition for PAC learnability.

Definition 0.3 (Agnostic PAC Learning)**.** A hypothesis class $\mathcal{H}$ is agnostically PAC learnable if there exists a learning algorithm $\mathcal{L}$ and a polynomial function $m_0(\cdot, \cdot)$ such that for all distributions D over $\mathcal{X} \times \mathcal{Y}$, for all $\varepsilon, \delta \in (0,1)$, if $m \geq m_0(1/\varepsilon, 1/\delta)$, the algorithm $\mathcal{L}$, given m i.i.d. samples S from D, returns a hypothesis $h_S \in \mathcal{H}$ such that with probability at least $1 - \delta$:

$$R_D(h_S) \leq \min_{h^* \in \mathcal{H}} R_D(h^*) + \varepsilon$$

Theorem 0.4 (Fundamental Theorem of PAC Learning)**.** *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$. Then $\mathcal{H}$ is PAC learnable if and only if it has finite VC dimension. Furthermore, if $VC - dim(\mathcal{H}) = d < \infty$, for any ERM algorithm $\mathcal{L}$, the sample complexity for the realizable case is given by*

$$m_0(\varepsilon, \delta) \in O\left(\frac{d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta}}{\varepsilon}\right).$$

For the agnostic case, the complexity is $m_0(\varepsilon, \delta) \in O\left(\frac{d + \log(1/\delta)}{\varepsilon^2}\right)$.

Definition 0.5 (Growth Function)**.** The *growth function* of a hypothesis class $\mathcal{H}$, denoted $\Pi_{\mathcal{H}}(m)$, is the maximum number of distinct ways that $\mathcal{H}$ can classify a set of m points.

$$\Pi_{\mathcal{H}}(m) = \max_{C \subset \mathcal{X}, |C|=m} |\{(h(x_1), \dots, h(x_m)) : h \in \mathcal{H}\}|$$

Lemma 0.6 (Sauer's Lemma)**.** *Let $\mathcal{H}$ be a hypothesis class with $VC - dim(\mathcal{H}) = d$. Then, for any m,*

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

If $m \geq d$, this can be further bounded: $\sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d$.

Before we begin the proof of today's main theorem, let's give a high level sketch of the proof approach:

1. Note that since $R_D(h^*) = 0$, $R_S(h^*) = 0$ for any sample S, and so ERM returns must return a hypothesis h_S with $R_S(h_S) = 0$. Therefore we want to bound the probability of drawing a sample S such that there exists $h \in \mathcal{H}$ where $R_D(h) > \varepsilon$, but $R_S(h) = 0$.

2. Introduce another random sample $T \sim D^m$ (sometimes called a "ghost sample"). Since T is random and independent of S, any bad hypothesis should have error at least $\varepsilon/2$ on T with good probability.

3. Relate the probability that S is "bad" for some hypothesis h to the probability that S is bad for h and T is "representative" for h.

4. Nothing special about S and T, they're both sampled i.i.d. from D, probability of S being bad while T representative can be bounded by probability that we draw a sample U of size $2m$ such that there exists some hypothesis h with $R_U(h) > \varepsilon/4$, but that a random partition of U gives one partition S for which $R_S(h) = 0$ (balls and bins argument). Union bound over the $\Pi_{\mathcal{H}}(2m)$ possible labelings of U.

5. Put everything together

Proof of Theorem 0.3 (Realizable Case). Ok let's start!

Step 1. Define the "bad" set B.

Let

$$B = \{S \subset \mathcal{X} : |S| = m \text{ and } \exists h \in \mathcal{H} \text{ such that } R_D(h) > \varepsilon, \text{ but } R_S(h) = 0\}$$

be the set of samples that are "bad" in the sense that there exists a bad hypothesis $h \in \mathcal{H}$ with zero error on S. We want to bound $\Pr_{S \sim D^m}[S \in B]$.

Step 2: Define the "bad" set B' with ghost sample T.

Now let

$$B' = \{(S,T) \subset \mathcal{X} : |S| = |T| = m, \exists h \in \mathcal{H} \text{ s.t. } R_D(h) > \varepsilon \wedge R_S(h) = 0 \wedge R_T(h) > \varepsilon/2\}$$

be the set of pairs of samples such that there is a "bad" hypothesis h with 0 error on S, but error at least half its expectation on T.

Step 3: Show that $\Pr_{S\sim D^m}[S \in B] \leq 2 \Pr_{S,T\sim D^{2m}}[(S,T) \in B']$.

$$\begin{aligned}
\Pr_{S,T\sim D^{2m}}[(S,T) \in B'] &= \mathop{\mathbb{E}}_{S,T\sim D^{2m}}[\mathbb{1}[(S,T) \in B']] \\
&= \mathop{\mathbb{E}}_{S\sim D^m}[\mathop{\mathbb{E}}_{T\sim D^m}[\mathbb{1}[(S,T) \in B']]] && \text{by independence} \\
&= \mathop{\mathbb{E}}_{S\sim D^m}[\mathbb{1}[S \in B] \cdot \mathop{\mathbb{E}}_{T\sim D^m}[\mathbb{1}[(S,T) \in B']]]
\end{aligned}$$

Note that if $S \notin B$, then $\mathbb{E}_{T\sim D^m}[\mathbb{1}[(S,T) \in B']] = 0$. However, whenever $S \in B$ because of some hypothesis h_S, we know by definition of the set B that $R_D(h_S) > \varepsilon$. Therefore $\mathbb{E}_{T\sim D^m}[R_T(h_S)] > \varepsilon$. Then

$$\begin{aligned}
\Pr_{T\sim D^m}[R_T(h_S) > \varepsilon/2] &= 1 - \Pr_{T\sim D^m}[R_T(h_S) \leq \varepsilon/2] \\
&\geq 1 - \Pr_{T\sim D^m}[R_T(h_S) - R_D(h_S) \leq -\varepsilon/2]
\end{aligned}$$

Let $\rho = R_D(h_S) > \varepsilon$. Then

$$\begin{aligned}
\Pr_{T\sim D^m}[R_T(h_S) - R_D(h_S) \leq -\varepsilon/2] &\leq \Pr_{T\sim D^m}[R_T(h_S) - R_D(h_S) \leq -\rho/2] \\
&\leq e^{-\frac{\rho m}{8}} \\
&\leq e^{-\frac{\varepsilon m}{8}} \\
&\leq 1/2
\end{aligned}$$

by multiplicative Chernoff bounds so long as $m \geq \frac{8\ln(1/2)}{\varepsilon}$. Therefore

$$\Pr_{T\sim D^m}[R_T(h_S) > \varepsilon/2] \geq 1/2,$$

which in turn implies that

$$\begin{aligned}
\Pr_{S,T\sim D^{2m}}[(S,T) \in B'] &= \mathop{\mathbb{E}}_{S\sim D^m}[\mathbb{1}[S \in B] \cdot \mathop{\mathbb{E}}_{T\sim D^m}[\mathbb{1}[(S,T) \in B']]] \\
&\geq \mathop{\mathbb{E}}_{S\sim D^m}[\mathbb{1}[S \in B] \cdot 1/2] \\
&= 1/2 \Pr_{S\sim D^m}[S \in B]
\end{aligned}$$

and so

$$\Pr_{S\sim D^m}[S \in B] \leq 2 \Pr_{S,T\sim D^{2m}}[(S,T) \in B'].$$

Step 4: Show that $\Pr_{S,T\sim D^{2m}}[(S,T)\in B'] \le e^{-\varepsilon m/4}\Pi_{\mathcal{H}}(2m)$.

We know that

$$\begin{aligned}
\Pr_{S,T\sim D^{2m}}[(S,T)\in B'] &= \mathop{\mathbb{E}}_{S,T\sim D^{2m}}[\mathbb{1}[(S,T)\in B']] \\
&= \mathop{\mathbb{E}}_{S,T\sim D^{2m}}[\max_{h\in\mathcal{H}} \mathbb{1}[R_D(h)>\varepsilon]\cdot \mathbb{1}[R_S(h)=0]\cdot \mathbb{1}[R_T(h)>\varepsilon/2]] \\
&\le \mathop{\mathbb{E}}_{S,T\sim D^{2m}}[\max_{h\in\mathcal{H}} \mathbb{1}[R_S(h)=0]\cdot \mathbb{1}[R_T(h)>\varepsilon/2]] \\
&\le \mathop{\mathbb{E}}_{S,T\sim D^{2m}}[\max_{h\in\mathcal{H}} \mathbb{1}[R_S(h)=0]\cdot \mathbb{1}[R_{S,T}(h)>\varepsilon/4]] \\
&\le \mathop{\mathbb{E}}_{S,T\sim D^{2m}}[\sum_{h\in\mathcal{H}_{S,T}} \mathbb{1}[R_S(h)=0]\cdot \mathbb{1}[R_{S,T}(h)>\varepsilon/4]] \\
&= \mathop{\mathbb{E}}_{U\sim D^{2m}}[\sum_{h\in\mathcal{H}_U} \mathbb{1}[R_{S'}(h)=0]\cdot \mathbb{1}[R_U(h)>\varepsilon/4]]
\end{aligned}$$

for any randomly selected subset $S'\subset U$ of size m, since the elements of S,T are i.i.d.

Let $S'\leftarrow \mathcal{U}(U)^m$ denote the process of randomly subsampling m elements from U *without replacement*.

$$\begin{aligned}
\Pr_{S,T\sim D^{2m}}[(S,T)\in B'] &\le \mathop{\mathbb{E}}_{U\sim D^{2m}}[\sum_{h\in\mathcal{H}_U} \mathbb{1}[R_{S'}(h)=0]\cdot \mathbb{1}[R_U(h)>\varepsilon/4]] \\
&\le \mathop{\mathbb{E}}_{U\sim D^{2m}}\mathop{\mathbb{E}}_{S'\leftarrow\mathcal{U}(U)}[\sum_{h\in\mathcal{H}_U} \mathbb{1}[R_{S'}(h)=0]\cdot \mathbb{1}[R_U(h)>\varepsilon/4]] \\
&= \mathop{\mathbb{E}}_{U\sim D^{2m}}[\sum_{h\in\mathcal{H}_U} \mathbb{1}[R_U(h)>\varepsilon/4]\mathop{\mathbb{E}}_{S'\leftarrow\mathcal{U}(U)} \mathbb{1}[R_{S'}(h)=0]] \\
&\le \mathop{\mathbb{E}}_{U\sim D^{2m}}[\sum_{h\in\mathcal{H}_U} \mathbb{1}[R_U(h)>\varepsilon/4](1-\varepsilon/4)^m)] \\
&\le \mathop{\mathbb{E}}_{U\sim D^{2m}}[\sum_{h\in\mathcal{H}_U} e^{-\varepsilon m/4}] \\
&\le e^{-\varepsilon m/4}\mathop{\mathbb{E}}_{U\sim D^{2m}}[|\mathcal{H}_U|] \\
&\le e^{-\varepsilon m/4}\Pi_{\mathcal{H}}(2m) && \text{by definition of } \Pi_{\mathcal{H}} \\
&\le e^{-\varepsilon m/4}(2em/d)^d && \text{from Sauer's lemma}
\end{aligned}$$

Step 5: Put everything together.

We want

$$\begin{aligned}
\Pr_{S\sim D^m}[S\in B] \le 2\Pr_{S,T\sim D^{2m}}[(S,T)\in B'] \le 2e^{-\varepsilon m/4}(2em/d)^d \le \delta. \\
\Rightarrow \ln 2 - \varepsilon m/4 + d\ln(2em/d) \le \ln(\delta) \\
\Rightarrow \varepsilon m/4 \ge \ln 2 + d\ln(2em/d) + \ln(1/\delta)
\end{aligned}$$

$$\Rightarrow m \geq \frac{4}{\varepsilon}(\ln 2 + d\ln(2em/d) + \ln(1/\delta))$$

Taking $m \in O(\frac{d\log(1/\varepsilon)}{\varepsilon} + \frac{\ln(1/\delta)}{\varepsilon})$ satisfies this inequality.

□

Proof of Sauer's Lemma. We prove this by induction on $m+d$. Let $\Pi_{\mathcal{H}}(C) = \{(h(x_1), \ldots, h(x_m)) : h \in \mathcal{H}\}$ for a set $C = \{x_1, \ldots, x_m\}$.

Base cases: If $m = 1$, $\Pi_{\mathcal{H}}(1) \leq 2 = \binom{1}{0} + \binom{1}{1}$ (if $d \geq 1$). If $d = 0$, $\mathcal{H}$ contains one hypothesis, so $\Pi_{\mathcal{H}}(m) = 1 = \binom{m}{0}$.

Inductive step: Assume the lemma holds for all pairs (m', d') with $m' + d' < m + d$. Let $C = \{x_1, \ldots, x_m\}$ be a set of m points. Let $C' = \{x_1, \ldots, x_{m-1}\}$.

Let $\Pi_{\mathcal{H}}(C')$ be the set of dichotomies on C'. Let $S \subseteq \Pi_{\mathcal{H}}(C')$ be the set of dichotomies on C' that can be extended to two different labelings for x_m. So

$$S = \{h_0(x_1), \ldots, h_0(x_{m-1}) : h_0 \in \mathcal{H} \wedge \exists h_1 \in \mathcal{H} \text{ s.t. } h_0(x_i) = h_1(x_i) \forall i \in [m-1] \wedge h_0(x_m) \neq h_1(x_m)\}$$

That is, for each $s \in S$, there exist $h_0, h_1 \in \mathcal{H}$ such that they agree on C' (producing s) but $h_0(x_m) = 0$ and $h_1(x_m) = 1$.

The total number of dichotomies on C can be counted as:

$$|\Pi_{\mathcal{H}}(C)| = |\Pi_{\mathcal{H}}(C')| + |S|$$

Now, consider the hypothesis class $\mathcal{H}_S$ which gives rise to the dichotomies in S. Any set shattered by $\mathcal{H}_S$ can be extended to shatter that set plus x_m using the original class $\mathcal{H}$. If $\mathcal{H}_S$ shatters a set $C'' \subset C'$ of size k, then $\mathcal{H}$ shatters $C'' \cup \{x_m\}$ of size $k + 1$. Since $VC(\mathcal{H}) = d$, we must have $k + 1 \leq d$, so $k \leq d - 1$. Thus, $VC(\mathcal{H}_S) \leq d - 1$.

By the induction hypothesis for $(m - 1, d)$ and $(m - 1, d - 1)$:

$$|\Pi_{\mathcal{H}}(C')| \leq \sum_{i=0}^{d} \binom{m-1}{i}$$

$$|S| \leq \Pi_{\mathcal{H}_S}(m-1) \leq \sum_{i=0}^{d-1} \binom{m-1}{i}$$

Combining these:

$$\begin{aligned}
|\Pi_{\mathcal{H}}(C)| &\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{j=1}^{d} \binom{m-1}{j-1} \\
&= \binom{m-1}{0} + \sum_{i=1}^{d} \left(\binom{m-1}{i} + \binom{m-1}{i-1} \right)
\end{aligned}$$

$$= \binom{m}{0} + \sum_{i=1}^{d} \binom{m}{i} \quad \text{(by Pascal's identity)}$$

$$= \sum_{i=0}^{d} \binom{m}{i}$$

Since this holds for any set C of size m, it holds for the maximum, $\Pi_{\mathcal{H}}(m)$. $\square$

VC Dimension of Various Models

- Linear classifiers in d dimensions: $d + 1$
- Neural networks with ReLU activation functions and finite precision weights: $O(WL \log W)$ where W is the number of weights and L is the number of layers [Bartlett et al., 2019]
- Neural networks with periodic activation functions: ∞

References

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Vladimir N Vapnik and Alexey Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16 (2):264–280, 1971.