

1 Learning in the unrealizable case: Agnostic PAC

Recall from last time that:

- For every distribution $p(x, y)$ and sample S_n it holds that:

$$e_p(h_S^{ERM}) - e_p^* \leq 2 \sup_{h \in \mathcal{H}} |e_S(h) - e_p(h)| \quad (1)$$

- For a finite hypothesis class it holds that:

$$\mathbb{P}_{S_n \sim p} \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_p(h)| \geq \epsilon \right] \leq 2|\mathcal{H}|e^{-2n\epsilon^2}$$

- For a finite hypothesis class it holds that:

$$\mathbb{P}_{S_n \sim p} \left[|e_p(h_S^{ERM}) - e_p^*| > \epsilon \right] \leq 2|\mathcal{H}|e^{-\frac{n\epsilon^2}{2}}$$

There are a couple of things we can conclude from this:

- For every $\delta > 0, n$ we have that with probability at least $1 - \delta$ it holds that:

$$e_p(h) \leq e_S(h) + \sqrt{\frac{1}{2n} \ln \frac{2|\mathcal{H}|}{\delta}} \quad (2)$$

The factor $\ln \frac{2|\mathcal{H}|}{\delta}$ can be improved to $\ln \frac{|\mathcal{H}|}{\delta}$ by using a one-sided inequality.

- For every $\delta, \epsilon > 0$ we have that for all $n \geq \frac{2}{\epsilon^2} \ln \frac{2|\mathcal{H}|}{\delta}$ it holds that:

$$\mathbb{P}_{S_n \sim p} \left[|e_p(h_S^{ERM}) - e_p^*| \geq \epsilon \right] \leq \delta \quad (3)$$

- For every $\delta, \epsilon > 0$ we have that with probability at least $1 - \delta$ it holds that:

$$e_p(h_S^{ERM}) \leq e_p^* + \sqrt{\frac{2}{n} \ln \frac{2|\mathcal{H}|}{\delta}} \quad (4)$$

Proof of 1: We know that:

$$\mathbb{P}_{S_n \sim p} \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_p(h)| \geq \epsilon \right] \leq 2|\mathcal{H}|e^{-2n\epsilon^2}$$

to have confidence δ we require

$$\mathbb{P}_{S_n \sim p} \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_p(h)| \geq \epsilon \right] \leq \delta$$

which will hold as long as:

$$2|\mathcal{H}|e^{-2n\epsilon^2} \leq \delta \quad (5)$$

This minimum ϵ for which this will hold is:

$$\epsilon(n, \delta) = \sqrt{\frac{1}{2n} \ln \frac{2|\mathcal{H}|}{\delta}} \quad (6)$$

and thus $e_p(h)$ is guaranteed to be in $[e_S(h) - \epsilon(n, \delta), e_S(h) + \epsilon(n, \delta)]$.

Proof of 2: We know that:

$$\mathbb{P}_{S_n \sim p} \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_p(h)| \geq \epsilon \right] \leq 2|\mathcal{H}|e^{-2n\epsilon^2}$$

We would like a confidence of δ which will hold as long as

$$2|\mathcal{H}|e^{-2n\epsilon^2} \leq \delta \quad (7)$$

This will hold for all $n \geq \frac{2}{\epsilon^2} \ln \frac{2|\mathcal{H}|}{\delta}$.

Proof of 3: Similar to proof of 1.

What happens for infinite hypothesis classes, e.g., linear separators. To get an intuition, we can see what happens with quantization. Consider the example of a perceptron where each parameter can be a k bit number. So the number of different hypotheses is $|\mathcal{H}| = (2^k)^{d+1}$, so $\ln |H| = (d+1)k \ln 2$.

2 Infinite Hypothesis Classes - The VC Dimension

Define $\Pi_{\mathcal{H}}(S)$ to be the set of dichotomies that \mathcal{H} induces on a given sample S . Formally:

$$\Pi_{\mathcal{H}}(S) = \{[h(x_1), \dots, h(x_n)] : h \in \mathcal{H}\} \quad (8)$$

Note $\Pi_{\mathcal{H}}(S) \subseteq \{0, 1\}^n$.

Denote by $\Pi_{\mathcal{H}}(n)$ the maximum number of dichotomies induced by a sample of size n :

$$\Pi_{\mathcal{H}}(n) = \max_{S: |S|=n} |\Pi_{\mathcal{H}}(S)| \quad (9)$$

$\Pi_{\mathcal{H}} : \mathbb{N} \rightarrow \mathbb{N}$ is called the **growth function**.

The key importance of this definition is the following uniform convergence results:

$$\mathbb{P}_{S_n \sim p} \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_p(h)| \geq \epsilon \right] \leq 4\Pi_{\mathcal{H}}(2n)e^{-\frac{n\epsilon^2}{8}}$$

Note how similar this is to

$$\mathbb{P}_{S_n \sim p} \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_p(h)| \geq \epsilon \right] \leq 2|\mathcal{H}|e^{-2n\epsilon^2}$$

However it is a pretty useless inequality if $\Pi_{\mathcal{H}}(n)$ grows exponentially fast with n .

If $|\Pi_{\mathcal{H}}(n)| = 2^n$ for a sample of size n we say that S is **shattered** by \mathcal{H} .

Definition: The VC dimension is the maximum n such that there exists a sample of size n that is shattered by \mathcal{H} . We'll denote it by $\text{VCdim}(\mathcal{H})$.

How can we prove that a $\text{VCdim}(\mathcal{H})$ has VC dimension k ?

- Find a sample S_k of size k that can be shattered by \mathcal{H} (i.e., $\Pi_{\mathcal{H}}(S_k) = 2^k$). This will show that $\text{VCdim}\mathcal{H} \geq k$.
- Show that any sample S_n with $n > k$ cannot be shattered. This will show that $\text{VCdim}\mathcal{H} \leq k$. Note that it is enough to show that $n + 1$ cannot be shattered, since shattering any larger sample will necessarily involve shattering a sub-sample of size $n + 1$.

Examples:

- Threshold functions on the line. Hypotheses of the type: $h(x) = \text{sgn}(wx - b)$ for any a, b . Any two non-identical points can be shattered, but not three since $(+, -, +)$ is not realizable. Thus $\text{VCdim}(\mathcal{H}) = 2$.
- Axis aligned rectangles. There is a set of four points which can be shattered (show it). So $\text{VCdim}(\mathcal{H}) \geq 4$. For five points there must be a point that is not the right/left/top/bottom most. There are thus two points such that any rectangle that includes them also includes this point, so it cannot have a negative labeling when they have a positive one.
- Linear separators: $h(x) = \text{sgn}(w \cdot x - b)$. Where $x, w \in \mathbb{R}^d$ and b is a scalar. Denote this class by \mathcal{H}_d . Cover showed that

$$\Pi_{\mathcal{H}_d}(n) = 2 \sum_{i=0}^d \binom{n-1}{i} \quad (10)$$

For $n \leq d + 1$ (so $d \geq n - 1$) we get

$$2 \sum_{i=0}^d \binom{n-1}{i} = 2 \sum_{i=0}^{n-1} \binom{n-1}{i} = 2^n \quad (11)$$

so all dichotomies are realized. If $n = d + 2$ we have

$$\Pi_{\mathcal{H}_d}(n) = 2 \sum_{i=0}^{n-2} \binom{n-1}{i} = 2 \sum_{i=0}^{n-1} \binom{n-1}{i} - 2 \binom{n-1}{n-1} = 2^n - 2 \quad (12)$$

so they are not shattered.

We can also give a direct proof of the following. First, show that there exist $d + 1$ points that can be shattered. Choose $x_i = e_i$ the

3 THE FUNCTION $\Pi_{\mathcal{H}}(N)$ GROWS POLYNOMIALLY IN N FOR FINITE $VCDIM(H)$

standard basis and $x_{d+1} = 0$ the all zeros vector. Say we want to realize a dichotomy s_1, \dots, s_{d+1} and assume $s_{d+1} = 1$. Choose $\mathbf{w} = [2s_1, \dots, 2s_d]$ and $b = -1$. Then $\text{sgn}(\mathbf{w}\mathbf{x}_i - b) = s_i$. This shows that a set of size $d + 1$ can be shattered.

Next, we want to prove that a set of $d + 2$ points cannot be shattered. Given such a set $\mathbf{x}_1, \dots, \mathbf{x}_m$ (where $m \geq d + 2$) define $\mathbf{v}_i = [\mathbf{x}_i, -1]$. These are m points in \mathbb{R}^{d+1} and hence are linearly dependent. i.e., there exist a_1, \dots, a_m (not all of them zero) such that $\mathbf{v}_m = \sum_{k=1}^{m-1} a_k \mathbf{v}_k$. Assume by contradiction that this set can be shattered. So any dichotomy can be realized. Consider the dichotomy $s = [\text{sgn}(a_1), \dots, \text{sgn}(a_{m-1}), -1]$ and denote by \mathbf{w}, b the parameters that realize it. Then lets see what is its sign on the m^{th} example:

$$[\mathbf{w}, b]^T \mathbf{v}_m = \sum_{k=1}^{m-1} a_k [\mathbf{w}, b]^T \mathbf{v}_k \quad (13)$$

Since $[\mathbf{w}, b]^T \mathbf{v}_k$ is assumed to have the sign of a_k the above sum will be non-negative. Thus the dichotomy s above cannot be realized.

- You might think that the VC dimension is just the number of parameters that define the class. This turns out to be wrong even for a very simple case. Consider hypotheses of the type $h(x) = \text{sgn}[\sin(\alpha x)]$. Then for any n it is possible find a set of points x_1, \dots, x_n that can be shattered. For example we can take $x_i = 2\pi 10^{-i}$. Then for a set of dichotomies δ_i we can set

$$\alpha(\delta) = \frac{1}{2} \left(\sum_{j=1}^l (1 - \delta_j) 10^j + 1 \right) \quad (14)$$

which yields

$$\sin(\alpha x_i) = \sin \left(\pi \left(\sum_{j=1}^l (1 - \delta_j) 10^{j-i} + 10^{-i} \right) \right) \quad (15)$$

3 The function $\Pi_{\mathcal{H}}(n)$ grows polynomially in n for finite $VCDim(H)$

To get a useful uniform convergence bound we need $\Pi_{\mathcal{H}}(n)$ to grow slowly enough we n . It turns out that this grows is polynomial in n as long as the VC dimension of H is finite.

Sauer's (and Shelah's) Lemma: Denote $d = VCDim(H)$. Then:

$$\Pi_{\mathcal{H}}(n) \leq \sum_{i=0}^d \binom{n}{i} \quad (16)$$

See proof in e.g. Kearns and Vazirani or Shai's lecture notes.

But how fast does this function grow?

Lemma: For $n > d$ it holds that:

$$\Pi_{\mathcal{H}}(n) \leq \left(\frac{en}{d}\right)^d \quad (17)$$

Proof:

$$\begin{aligned} \left(\frac{d}{n}\right)^d \sum_{i=0}^d \binom{n}{i} &\leq \sum_{i=0}^d \left(\frac{d}{n}\right)^i \binom{n}{i} && \text{Since } d < n \\ &\leq \sum_{i=0}^d \left(\frac{d}{n}\right)^i \binom{n}{i} && \text{Since } d < n \\ &\leq \sum_{i=0}^n \left(\frac{d}{n}\right)^i \binom{n}{i} && \text{Since } d < n \\ &= \left(1 + \frac{d}{n}\right)^n \leq e^d \end{aligned}$$

4 Generalization Bounds

Recall the main theorem saying that:

$$\mathbb{P}_{S_n \sim p} \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_p(h)| \geq \epsilon \right] \leq 4\Pi_{\mathcal{H}}(2n) e^{-\frac{n\epsilon^2}{8}}$$

Several results follow from this:

- **Confidence bound on $e_p(h)$:** Lets say we have n samples and we want a bound on $e_p(h)$ that holds with probability at least $1 - \delta$ over the samples S_n drawn from p . Given n, δ we want the smallest ϵ such that the uniform convergence bound holds with confidence δ . To get confidence δ we need:

$$4\Pi_{\mathcal{H}}(2n) e^{-\frac{n\epsilon^2}{8}} \leq \delta \quad (18)$$

We know that $\Pi_{\mathcal{H}}(2n) \leq \left(\frac{2en}{d}\right)^d$ so if we have

$$4 \left(\frac{2en}{d}\right)^d e^{-\frac{n\epsilon^2}{8}} \leq \delta$$

This will also imply confidence δ . The minimum ϵ for which this holds is when we have equality.

$$\epsilon = \sqrt{\frac{8}{n} \left[d \ln \left(\frac{2en}{d} \right) + \ln \frac{4}{\delta} \right]}$$

From which we get the following $1 - \delta$ confidence bound on $e_p(h)$:

$$e_S(h) - \sqrt{\frac{8}{n} \left[d \ln \left(\frac{2en}{d} \right) + \ln \frac{4}{\delta} \right]} \leq e_p(h) \leq e_S(h) + \sqrt{\frac{8}{n} \left[d \ln \left(\frac{2en}{d} \right) + \ln \frac{4}{\delta} \right]} \quad (19)$$

- **Sample complexity bounds:** Similar to the finite hypothesis class case, bounds can be obtained on the n needed to obtain a given ϵ, δ accuracy and confidence. It can be shown that for a given $\epsilon, \delta > 0$ the following number of samples suffices to get ϵ, δ learning:

$$n \geq \frac{64}{\epsilon^2} \left(2d \ln \frac{12}{\epsilon} + \ln \frac{4}{\delta} \right) \quad (20)$$

From this we can conclude that classes \mathcal{H} with finite VCdim are PAC learnable. In fact, it can be shown that if a class has infinite VCdim then it is not PAC learnable. This means that there exist ϵ_0, δ_0 such that there is a distribution p such that the algorithm fails to return a required hypothesis regardless of the number of samples.

- **Generalization error of ERM:** The difference between the generalization error of the ERM and the best hypothesis in the class can also be obtained. Namely, for every n, δ with confidence δ .

To do this we use the lemma linking the maximum difference between the training and generalization curves to the ERM generalization error.

$$\begin{aligned} \mathbb{P}_{S_n \sim p} [e_p(h_S^{ERM}) - e_p^* \geq \epsilon] &\leq \mathbb{P}_{S_n \sim p} [2 \sup_{h \in \mathcal{H}} |e_S(h) - e_p(h)| \geq \epsilon] \\ &= \mathbb{P}_{S_n \sim p} \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_p(h)| \geq \frac{\epsilon}{2} \right] \\ &\leq 4 \left(\frac{2en}{d} \right)^d e^{-\frac{n\epsilon^2}{32}} = \delta \end{aligned}$$

And the minimum ϵ for which this holds is:

$$\epsilon = \sqrt{\frac{32}{n} \left[d \ln \left(\frac{2en}{d} \right) + \ln \frac{4}{\delta} \right]}$$