12, 13, 25 September 2021

CS 492(F), Comp. Learning Theory – Rademacher Complexity and VC-Dimension (Ch3)

1. Motivation:

(1) In the previous chapter, we learnt two results on uniform generalisation bounds for finite hypothesis sets, which let us state the guarantees of learning algorithms formally. However, these results stop applying as soon as the hypothesis set $\mathcal{H}$ becomes infinite. In fact, this limitation cannot be fixed by naive tricks, because the conclusions of those results crucially rely on the finiteness of $\mathcal{H}$:

$$\mathbb{P}\left[\forall h \in \mathcal{H}.\ R(h) \le \hat{R}_S(h) + \ldots \log|\mathcal{H}| \ldots + \ldots\right] \ge 1-\delta.$$

but $\log|\mathcal{H}| = \infty$ for an infinite $\mathcal{H}$.

(2) What should we do to handle infinite $\mathcal{H}$'s? The answer from this chapter is to use an appropriate notion of complexity of a hypothesis set and try to get a uniform generalisation bound using that notion, instead of $\log|\mathcal{H}|$.

(3) Rademacher complexity, growth function, and VC-dimensions are such notions. They all measure quantitatively the complexity of $\mathcal{H}$. We will study uniform generalisation bounds using these notions.

(4) The results of this chapter are not tied to specific learning algorithms. They are fundamental, and they are used repeatedly in the book to analyse concrete ML algorithms.

2. Setting

(1) We assume general input and output spaces $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$. Usually, $\mathcal{Y}$ is finite. For instance, $\mathcal{Y} = \{-1, 1\}$ or $\mathcal{Y} = \{0, 1\}$.

(2) $\mathcal{H} \subseteq [\mathcal{X} \to \mathcal{Y}]$ is an assumed (nonempty) set of hypotheses.

(3) $L : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ is an assumed (loss) function. Intuitively, it describes a loss. One example L is $(y_1, y_2) \mapsto \mathbb{1}\{y_1 \neq y_2\}$.

(4) $\mathcal{G} \overset{\text{def.}}{=} \{ (x, y) \mapsto L(h(x), y) \in [\mathcal{X} \times \mathcal{Y} \to \mathbb{R}] \mid h \in \mathcal{H} \}$.

Note the close relationship between $\mathcal{H}$ and $\mathcal{G}$. We will later show that the complexity of $\mathcal{H}$ is closely related to that of $\mathcal{G}$.

3. Rademacher Complexity.

(1) $\mathcal{U}$ ... input space. $\mathcal{V} \subseteq \mathbb{R}$ ... output space.

$\mathcal{F} \subseteq [\mathcal{U} \to \mathcal{V}]$ ... nonempty set of functions.

Def 3.1 Let $S \in \mathcal{U}^m$, that is, $S = (u_1, u_2, \cdots, u_m)$.

The empirical Rademacher complexity of $\mathcal{F}$ with respect to S is defined as follows:

$$\widehat{R}_S(\mathcal{F}) \overset{\text{def.}}{=} \mathop{\mathbb{E}}_{\vec{\sigma} \sim (\text{uniform}\{-1,1\})^m} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \right].$$

Note that $\sum_{i=1}^{m} \sigma_i f(x_i) = \langle (\sigma_i)_{i \in [m]}, (f(x_i))_{i \in [m]} \rangle$ (i.e., inner product) which gets maximized when $(f(x_i))_{i \in [m]}$ and $(\sigma_i)_{i \in [m]}$ (among $(f(x_i))_{i \in [m]}$'s with $\|(f(x_i))_{i \in [m]}\| = c$ for some fixed c) point to the same direction as m-dimensional vectors.

Thus, $\hat{R}_S(\mathcal{F})$ measures the complexity of $\mathcal{F}$ by measuring how well $\mathcal{F}$ expresses different directions when we consider vectors $(f(x_i))_{i\in[m]}$ for $f \in \mathcal{F}$.

Def 3.2. Let $D \in Pr(\mathcal{U})$ and $m \in \mathbb{N}$. The **Rademacher complexity** of $\mathcal{F}$ with respect to D and m is:

$$R_m(\mathcal{F}) \overset{\text{def.}}{=} \mathbb{E}_{S\sim D^m}[\hat{R}_S(\mathcal{F})],$$

that is, the expectation of the empirical Rademacher complexity over sample $S \sim D^m$.

(2) Example.

① $\mathcal{F} \subseteq [\mathbb{R} \to \{+1,-1\}]$

$\mathcal{F} = \{ u \mapsto \mathbb{1}_{\{u\in[l,r]\}} - \mathbb{1}_{\{u\notin[l,r]\}} \mid l \le r \in 3 \}$.

② $S_1 = (3)$ $\quad S_2 = (3,-2)$ $\quad S_3 = (3,-2,5)$.

$\hat{R}_{S_1}(\mathcal{F}) = 1$ $\quad \hat{R}_{S_2} = 1$ $\quad \hat{R}_{S_3} = \frac{1}{8}\times 1\times 1 + \frac{1}{8}\times\frac{2}{3}\times 1 = \frac{13}{24}$

$S_4 = (3,3)$

$\hat{R}_{S_4}(\mathcal{F}) = \frac{1}{4}\times 1\times 2 + \frac{1}{4}\times 0\times 2 = \frac{1}{2}$.

③ If D is the Dirac distribution on 3 (so that $(u_1,\dots,u_m)\sim D^m$ implies $u_1 = \dots = u_m = 3$), then $R_2(\mathcal{F}) = \frac{1}{2}$. But if D is the uniform distribution on $[0,1]$ (so that $(u_1,u_2)\sim D^2$ implies $u_1 \neq u_2$ with prob. 1), then $R_2(\mathcal{F}) = 1$.

④ Note that R_m depends on the distribution D, and $\hat{R}_S$ depends on the sample S.

(3) Generalization bound.

important boundedness requirement

Thm 3.3. $\mathcal{F} \subseteq [\mathcal{U} \to [0,1]]$. That is, $\mathcal{D} = [0,1]$.

Rademacher complexity.

Let $\delta > 0$, $m \in \mathbb{N}$ and D be a distribution on $\mathcal{U}$. Then,

$$\mathbb{P}_{S \sim D^m}\left[\forall f \in \mathcal{F}.\ \mathbb{E}_{u \sim D}[f(u)] \le \mathbb{E}_{u \sim \hat{D}_S}[f(u)] + 2R_m(\mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right] \ge 1-\delta.$$

↳ empirical distribution over S, i.e., uniform distribution over S.

and $$\mathbb{P}_{S \sim D^m}\left[\forall f \in \mathcal{F}.\ \mathbb{E}_{u \sim D}[f(u)] \le \mathbb{E}_{u \sim \hat{D}_S}[f(u)] + 2\hat{R}_S(\mathcal{F}) + 3\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right] \ge 1-\delta.$$

↑ empirical Rademacher complexity.

which doesn't refer to a specific $f \in \mathcal{F}$

① $R_m(\mathcal{F})$ in the upper bound is what lets us have a uniform bound, i.e., a bound that works for all $f \in \mathcal{F}$.

② The proof uses McDiarmid's inequality

Thm D.8 [McDiarmid's inequality]

$X_1, \dots, X_m \in \mathcal{X}$ for $m \ge 1$ .... independent random vars.

$c_1, \dots, c_m > 0$.

$f : \mathcal{X}^m \to \mathbb{R}$ s.t. $|f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_m) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots x_m)| \le c_i$

for all $x_1, \dots, x_m, x_i', i$.

Then, $$\mathbb{P}\left[f(X_1, \dots X_m) - \mathbb{E}_{X_1, \dots X_m}[f(X_1, \dots, X_m)] \ge \varepsilon\right] \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^m c_i^2}\right)$$

and $$\mathbb{P}\left[f(X_1, \dots, X_m) - \mathbb{E}_{X_1, \dots X_m}[f(X_1, \dots, X_m)] \le -\varepsilon\right] \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^m c_i^2}\right)$$

* More general than Hoeffding's inequality.

Proof of Thm 3.3. Fix $\delta > 0$, $m \in \mathbb{N}$.

Define $\Phi : \mathcal{U}^m \to [-1,1]$ by

$$\Phi(S) \overset{\text{def.}}{=} \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{u \sim D}[f(u)] - \mathbb{E}_{u \sim \hat{D}_S}[f(u)] \right).$$

For the first bound in the thm, it suffices to prove that:

with prob. $\geq 1-\delta$ over $S \sim D^m$, (with probability at least $1-\delta$)

$$\Phi(S) \leq 2\mathcal{R}_m(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

We will show: (i) $\Phi(S) \leq \mathbb{E}_{S \sim D^m}[\Phi(S)] + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$ w.p. $\geq 1-\delta$.

(ii) $\mathbb{E}_{S \sim D^m}[\Phi(S)] \leq 2\mathcal{R}_m(\mathcal{F})$.

Why does (i) hold? Because of McDiarmid's inequality, as shown below:

$$\left| \Phi(u_1, \dots, u_i, \dots, u_m) - \Phi(u_1, \dots, u_i', \dots, u_m) \right|$$

$$\leq \left| \sup_{f \in \mathcal{F}} \left(\left(\mathbb{E}_{u \sim D}[f(u)] - \mathbb{E}_{u \sim \hat{D}_{(u_1, \dots, u_i, \dots, u_m)}}[f(u)] \right) - \left(\mathbb{E}_{u \sim D}[f(u)] - \mathbb{E}_{u' \sim \hat{D}_{(u_1, \dots, u_i', \dots, u_m)}}[f(u')] \right) \right) \right|$$

($\hat{D}_{(u_1, \dots, u_i, \dots, u_m)} = \hat{D}_S$, $\hat{D}_{(u_1, \dots, u_i', \dots, u_m)} = \hat{D}_{S'}$)

$$= \left| \sup_{f \in \mathcal{F}} \frac{1}{m} \left(f(u_i') - f(u_i) \right) \right| \leq \frac{1}{m}$$

Thus, by McDiarmid's ineq., for any $\varepsilon > 0$,

$$\mathbb{P}_{S \sim D^m} \left[\Phi(S) - \mathbb{E}_{S \sim D^m}[\Phi(S)] \leq \varepsilon \right] \geq 1 - \exp\left(- \frac{2\varepsilon^2}{m \cdot \left(\frac{1}{m}\right)^2} \right)$$

$$= 1 - \exp(-2m\varepsilon^2).$$

So, if $\varepsilon = \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$, we can get (i).

Called Rademacher variables.

What about (ii)? Because we can exploit symmetries of $\vec{\sigma}$ in the def'n of $\mathcal{R}_m(\mathcal{F})$, as shown in the following calculation:

$$\mathop{\mathbb{E}}_{S\sim D^m}[\Phi(S)] = \mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{f\in F}\left(\mathop{\mathbb{E}}_{u\sim D}[f(u)] - \mathop{\mathbb{E}}_{u\sim \hat{D}_S}[f(u)]\right)\right]$$

$$= \mathop{\mathbb{E}}_{S}\left[\sup_{f}\left(\mathop{\mathbb{E}}_{S'\sim D^m}\left[\mathop{\mathbb{E}}_{u'\sim \hat{D}_{S'}}[f(u')]\right] - \mathop{\mathbb{E}}_{u\sim \hat{D}_S}[f(u)]\right)\right]$$

$$= \mathop{\mathbb{E}}_{S}\left[\sup_{f}\mathop{\mathbb{E}}_{S'}\left[\mathop{\mathbb{E}}_{u'\sim \hat{D}_{S'}}[f(u')] - \mathop{\mathbb{E}}_{u\sim \hat{D}_S}[f(u)]\right]\right]$$

$$\le \mathop{\mathbb{E}}_{S,S'}\left[\sup_{f}\left(\mathop{\mathbb{E}}_{u'\sim \hat{D}_{S'}}[f(u')] - \mathop{\mathbb{E}}_{u\sim \hat{D}_S}[f(u)]\right)\right].$$

flipping parts of S and S' based on $\vec{\sigma}$ doesn't change the dist. of (S,S')

$$= \mathop{\mathbb{E}}_{\substack{S,S'\\ \vec{\sigma}\sim \text{uniform}\{\pm 1\}^m}}\left[\sup_{f}\frac{1}{m}\sum_{i=1}^m \sigma_i\left(f(u_i') - f(u_i)\right)\right].$$

$$\le \mathop{\mathbb{E}}_{S',\vec{\sigma}}\left[\sup_{f}\frac{1}{m}\sum_{i=1}^m \sigma_i f(u_i')\right] + \mathop{\mathbb{E}}_{S,\vec{\sigma}}\left[\sup_{f}\frac{1}{m}\sum_{i=1}^m -\sigma_i f(u_i)\right]$$

$(S,\vec{\sigma})$ and $(S,-\vec{\sigma})$ have the same distributions.

$$= \text{"} \qquad + \mathop{\mathbb{E}}_{S,\vec{\sigma}}\left[\sup\frac{1}{m}\sum_{i=1}^m \sigma_i f(u_i)\right]$$

$$= 2R_m(F).$$

the argument for

For the second claim of the thm, we repeat (i) but for $\frac{\delta}{2}$, instead of δ. This gives us:

$$\mathop{\mathbb{P}}_{S\sim D^m}\left[\Phi(S) \le \mathop{\mathbb{E}}_{S\sim D^m}[\Phi(S)] + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \ge 1-\frac{\delta}{2}. \quad \text{(iii)}$$

By (ii), we have $\mathop{\mathbb{E}}_{S\sim D^m}[\Phi(S)] \le 2R_m(F)$. (iv)

Now, if we apply McDiarmid's inequality to the function

$$S=(u_1,\dots,u_m) \longmapsto \widehat{R}_S(F)$$

then by essentially the same argument as the one for (i), we get:

$$\mathop{\mathbb{P}}_{S\sim D^m}\left[R_m(F) \le \widehat{R}_S(F) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \ge 1-\frac{\delta}{2}. \quad \text{(v)}$$

the claim of the theorem follows from (iii), (iv) and (v) via union bound.

D.

4. Rademacher Complexity in Our Setup.

(i) How does the generalization bound (in Thm 3.3) apply in our setup?

In our setup,

$Y = \{-1, 1\}$. $\quad \mathcal{H} \subseteq [\mathcal{X} \to Y]$ ... hypo. set.

$Z = \mathcal{X} \times Y \quad \mathcal{F} = \{ (x,y) \mapsto \mathbb{1}\{h(x) \neq y\} \mid h \in \mathcal{H} \}$.

Then, for $f(x,y) = \mathbb{1}\{h(x) \neq y\}$ and $D \in Pr(Z) = Pr(\mathcal{X} \times Y)$

$$\mathbb{E}_{(x,y)\sim D}[f(x,y)] = \mathbb{E}_{(x,y)\sim D}[\mathbb{1}\{h(x) \neq y\}] = R(h)$$

$$\mathbb{E}_{(x,y)\sim \hat{D}_S}[f(x,y)] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq y_i\} = \hat{R}_S(h).$$

$(S = ((x_1,y_1), \cdots, (x_m,y_m)))$

Thus, the bounds in Thm 3.3 become:

(vi) $$\mathbb{P}_{S\sim D^m}\left[\forall h \in \mathcal{H}.\ R(h) \le \hat{R}_S(h) + 2R_m(\mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right] \ge 1-\delta.$$

and (vii) $$\mathbb{P}_{S\sim D^m}\left[\forall h \in \mathcal{H}.\ R(h) \le \hat{R}_S(h) + 2\hat{R}_S(\mathcal{F}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \ge 1-\delta.$$

To get meaningful bounds, we should rephrase $R_m(\mathcal{F})$ and $\hat{R}_S(\mathcal{F})$ in terms of $R_m(\mathcal{H})$ and $\hat{R}_{S_{\mathcal{X}}}(\mathcal{H})$ (projection of S to $\mathcal{X}$-part. That is, $S_{\mathcal{X}} = (x_1, \cdots, x_m)$.)

(ii)

Lemma 3.4 For all $S = ((x_1,y_1), \cdots, (x_m,y_m)) \in (\mathcal{X}\times Y)^m$,

$$2\cdot\hat{R}_S(\mathcal{F}) = \hat{R}_{S_{\mathcal{X}}}(\mathcal{H})$$

where $S_{\mathcal{X}} = (x_1, \cdots, x_m)$.

Proof $\hat{R}_S(\mathcal{F}) = \mathbb{E}_{\vec{\sigma}\sim(\text{Uniform }\{-1,1\})^m}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i,y_i)\right]$

$= \mathbb{E}_{\vec{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i \mathbb{1}\{h(x_i)\neq y_i\}\right]$.

$= \mathbb{E}_{\vec{\sigma}}\left[\sup_{h} \frac{1}{m}\sum_{i=1}^m \sigma_i \frac{1-h(x_i)y_i}{2}\right]$

$= \mathbb{E}_{\vec{\sigma}}\left[\frac{1}{m}\sum_{i=1}^m \frac{\sigma_i}{2}\right] + \frac{1}{2}\mathbb{E}_{\vec{\sigma}}\left[\sup_{h} \frac{1}{m}\sum_{i=1}^m (-y_i\sigma_i)h(x_i)\right]$

$= 0 + \frac{1}{2}\mathbb{E}_{\vec{\sigma}}\left[\sup_{h} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)\right] = \frac{1}{2}\hat{R}_{S_{\mathcal{X}}}(\mathcal{H})$ □.

$(\sigma_1,\dots,\sigma_m)$ and $(-y_1\sigma_1,\dots,-y_m\sigma_m)$ have the same distribution.

$\frac{1}{2}$ here is the reason that we can remove 2 from $2\hat{R}_S(\mathcal{F})$

Thm 3.5 Let $D \in \mathcal{R}(\mathcal{X}\times\mathcal{Y})$

Then, for all $\delta > 0$,

$$\mathbb{P}_{S\sim D^m}\left[\forall h\in\mathcal{H}.\ R(h) \le \hat{R}_S(h) + \mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right] \ge 1-\delta$$

and

$$\mathbb{P}_{S\sim D^m}\left[\forall h\in\mathcal{H}.\ R(h) \le \hat{R}_S(h) + \hat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \ge 1-\delta.$$

$R(h)$ depend on D.

Proof. By Lemma 3.4, $2\hat{R}_S(\mathcal{F}) = \hat{R}_{S_{\mathcal{X}}}(\mathcal{H})$.

Taking the expectation over S gives: $2\mathcal{R}_m(\mathcal{F}) = \mathcal{R}_m(\mathcal{H})$.

Replacing $\mathcal{F}$-related parts in (vi) and (vii) using the above equations gives the conclusion of the theorem. □.

5. Growth Function.

(1) Computing $\mathcal{R}_S(\mathcal{H})$ and $\mathcal{R}_m(\mathcal{H})$ is hard in most cases, partly because their dependence on sample S and distribution D. Actually, this difficulty comes from the fact that they carry a lot of information about the problem. (which is a good thing.).

(2) The growth function and VC dimension are complexity measures of a hypothesis set $\mathcal{H}$, which are less informative than Rademacher complexity but easier to compute as a result. They describe how expressive $\mathcal{H}$ is.

Def 3.6 The growth function $\Pi_{\mathcal{H}}: \mathbb{N} \to \mathbb{N}$ for a hypo. set $\mathcal{H}$ is defined by $\Pi_{\mathcal{H}}(m) = \max_{(x_1,\dots,x_m) \in \mathcal{X}^m} |\{(h(x_1),\dots,h(x_m)) \mid h \in \mathcal{H}\}|$

① We continue to use the setup in 4. So, $\mathcal{Y} = \{-1, 1\}$.

② Note that $\Pi_{\mathcal{H}}(m) \le 2^m$ (since $\mathcal{Y} = \{-1,1\}$).

③ Large $\Pi_{\mathcal{H}}(m)$ means very expressive

Thm 3.7 [Massart's lemma]. Let $A \subseteq \mathbb{R}^m$ be a finite nonempty set. with $r = \max_{y \in A} \|y\|_2$. Then,

$$\mathbb{E}_{\vec{\sigma} \sim (\text{uniform}\{-1,1\})^m}\left[\frac{1}{m} \sup_{y \in A} \sum_{i=1}^m \sigma_i y_i\right] \le \frac{r\sqrt{2\log|A|}}{m}.$$

Pf. We use Cor D.11

Cor D.11. [Maximum inequality]

$(Y_{ij})_{(i,j) \in [m] \times [n]} \in \mathbb{R}^{m \times n}$ ... random variables s.t.

(i) $\mathbb{E}[Y_{ij}] = 0$, $Y_{ij} \in [-r_{ij}, r_{ij}]$.

(ii) $Y_{1j}, Y_{2j}, \dots, Y_{mj}$ are independent for every $j \in [n]$.

Let $Z_j = \sum_{i=1}^m Y_{ij}$.

Then, $\mathbb{E}[\max_{j \in [n]} Z_j] \le r\sqrt{2\log n}$, $r = \sup_j \sqrt{\sum_{i=1}^m r_{ij}^2}$

$(Y_{iy})_{(i,y)\in[m]\times A}$ where $Y_{iy} = \sigma_i y_i$ (random because σ_i is Co.)

Then Y_{iy} is mean zero, and $(Y_{iy})_i$ is indep. for each $y \in A$.

Also, $-|y_i| \le Y_{iy} \le |y_i|$. Thus, by Cor D.11,

$$\mathbb{E}\left[\max_{y\in A} \sum_{i=1}^m Y_{iy}\right] \le \sup_{y\in A} \|y\|_2 \cdot \sqrt{2\log|A|}$$

Thus

$$\mathop{\mathbb{E}}_{\vec\sigma}\left[\max_{y\in A} \frac{1}{m}\sum_{i=1}^m \sigma_i y_i\right] \le \frac{1}{m} r \cdot \sqrt{2\log|A|}. \qquad \square$$

Cor 3.8. Let $\mathcal{H}$ be a family of measurable functions from $\mathcal{X}$ to $\{-1,1\}$. Consider $D \in P_1(\mathcal{X})$. Then,

$$R_m(\mathcal{H}) \le \sqrt{\frac{2\log \Pi_{\mathcal{H}}(m)}{m}}$$

Proof: for each $S = (x_1,\dots,x_m) \in \mathcal{X}^m$,

$$\mathcal{H}|_S \overset{\text{def}}{=} \{(h(x_1),\dots,h(x_m)) \mid h\in\mathcal{H}\}.$$

Then,

$$R_m(\mathcal{H}) = \mathop{\mathbb{E}}_{S\sim D^m}\left[\mathop{\mathbb{E}}_{\vec\sigma\sim(\text{uniform}\{\pm1\})^m}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum \sigma_i h(x_i)\right]\right]$$

$$= \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\vec\sigma}\left[\sup_{\vec y\in\mathcal{H}|_S} \frac{1}{m}\sum \sigma_i y_i\right]\right] \le \mathop{\mathbb{E}}_{S}\left[\frac{1}{m}\sqrt{m}\sqrt{2\log|\mathcal{H}|_S|}\right]$$

$$\le \mathop{\mathbb{E}}_{S}\left[\sqrt{\frac{2\log\Pi_{\mathcal{H}}(m)}{m}}\right] = \sqrt{\frac{2\log\Pi_{\mathcal{H}}(m)}{m}}.$$

$\square$

Thm 3.9. For all $D \in P_1(\mathcal{X}\times\{\pm1\})$ and $\delta > 0$,

$$\mathop{\mathbb{P}}_{S\sim D^m}\left[\forall h\in\mathcal{H}.\ R(h) \le R_S(h) + \sqrt{\frac{2\log\Pi_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right] \ge 1-\delta.$$

Proof. Immediate consequence of Thm 3.5 and Cor 3.8. □

6. VC Dimension

(1) The growth fn of $\mathcal{H}$ is easier to deal with than the Rademacher complexity of $\mathcal{H}$, partly due to its independence on the distribution D on $\mathcal{X}$ and sample $S \in \mathcal{X}^m$.

But sometimes it is still difficult to compute. Can we get a measure on the complexity / expressiveness of $\mathcal{H}$ that is easier to handle?

(2) The VC dimension of $\mathcal{H}$ is such a measure. It is more abstract (i.e., less informative) than the growth fn. $\Pi_{\mathcal{H}}$ by not being dependent on m, the size of sample. In so doing, it becomes a quantity that is easier to compute. Note the usual high-level strategy of using a less-informative but easier-to-handle concept

(3). $\mathcal{Y} = \{-1, +1\}$. $\mathcal{H} \subseteq [\mathcal{X} \to \mathcal{Y}]$ by $\mathcal{H}$.

Def. A (finite) subset $X \subseteq \mathcal{X}$ is shattered if for every subset $X_0 \subseteq X$, there exists a hypo. $h \in \mathcal{H}$ s.t. $X_0 = \{x \in X \mid h(x) = +1\}$.

Def 3.10. The VC dimension of a hypothesis set $\mathcal{H}$ is the size of the largest (finite) set that is shattered by $\mathcal{H}$:

$$\text{VCdim}(\mathcal{H}) = \sup\{m \mid \Pi_{\mathcal{H}}(m) = 2^m\}.$$

(1) Note that the largest finite set might not exist. To cover such a case, we use sup here. In that case, $\text{VCdim}(\mathcal{H}) = \infty$.

② $VCDim(\mathcal{H}) = d$ doesn't imply that every subset of $\mathcal{X}$ with size d is shattered. Instead, it means some subset is shattered.

③ $VCDim(\mathcal{H}) = d$ implies that for every $d' \le d$, there is an $\mathcal{H}$-shattered subset of $\mathcal{X}$ with size d'. Exercise: Prove this.

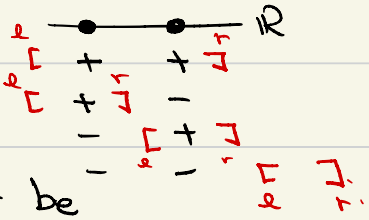
④ If $\mathcal{H}$ consists of diverse fns, for a given set of labelled inputs $(x_1, y_1), \cdots, (x_m, y_m)$, $\mathcal{H}$ is likely to have a hypothesis h that has zero error for this set. The VC dimension uses this informal consequence to measure the complexity of $\mathcal{H}$.

$VCDim(\mathcal{H}) = d$ means that for some size-d subset $X_0 \subseteq \mathcal{X}$, however we label inputs in X_0, we can achieve 0 error with some $h \in \mathcal{H}$.

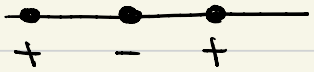
(4) Examples.

① Intervals on the real line.

$\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{ x \mapsto$ if $l \le x \le u$ then 1 else $-1 \mid l, u \in \mathbb{R} \}$.

$VCDim(\mathcal{H}) = 2$. Why? 

So, shattered.

But: 

cannot be classified correctly by any $h \in \mathcal{H}$.

② Hyperplanes in $\mathbb{R}^d$.

$\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{ x \mapsto \text{sgn}(\langle w, x \rangle + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R} \}$

$\text{sgn}(y) = \begin{cases} +1 & \text{if } y \ge 0 \\ -1 & \text{if } y < 0 \end{cases}$

$VCDim(\mathcal{H}) = d+1$

(i) Why $VCDim(\mathcal{H}) \ge d+1$?

$x_0 = (0, 0, \cdots 0) \in \mathbb{R}^d$

$x_1 = (1, 0 \cdots, 0)$

$\vdots$

$x_d = (0, 0, \cdots 0, 1)$

Then, for every $(y_0, \cdots, y_d) \in \{-1, +1\}^{d+1}$, pick the following $h \in \mathcal{H}$:

$h(x) = \mathrm{sgn}(\langle y, x\rangle + \frac{y_0}{2})$ where $y = (y_1, \dots, y_d)$.

Then, $h(x_0) = \mathrm{sgn}(\frac{y_0}{2}) = y_0$. and $h(x_i) = \mathrm{sgn}(y_i + \frac{y_0}{2}) = y_i$ $(i \geq 1)$.

(ii) why $\mathrm{VCDim}(\mathcal{H}) \leq d+1$? Because of Radon's thm.

Thm 3.13. [Radon's thm] Any $X \subseteq \mathbb{R}^d$ with $|X| = d+2$.

can be partitioned into X_1 and X_2 s.t. the convex hulls of X_1 and X_2 overlap.

Note that if we label every X_1 with $+1$ and X_2 with -1, the X with this labelling cannot be classified correctly by any $h \in \mathcal{H}$.

Thus, the thm implies that $\mathrm{VCDim}(\mathcal{H}) < d+2$.

Proof Let $X = \{x_1, \dots, x_{d+2}\} \subseteq \mathbb{R}^d$ and consider.

the following system of $(d+1)$ linear equations over $d+2$ variables $\alpha_1, \dots, \alpha_{d+2}$:

d equations ... 1 equation

$$\sum_{i=1}^{d+2} \alpha_i x_i = \vec{0} \ (\in \mathbb{R}^d) \qquad \sum_{i=1}^{d+2} \alpha_i = 0.$$

Then, $\exists$ a solution $(\beta_i)_{i \in [d+2]}$ s.t. $\beta_i \neq 0$ for some $i \in [d+2]$

Let $J_1 = \{j \mid \beta_j > 0\}$ and $J_2 = \{j \mid \beta_j \leq 0\}$.

Since $\sum_{j=1}^{d+2} \beta_j = 0$ and $\beta_i \neq 0$ for some i, $J_1 \neq \emptyset$ and $J_2 \neq \emptyset$.

Let $X_1 = \{x_j \mid j \in J_1\}$ and $X_2 = \{x_j \mid j \in J_2\}$.

Define $\beta = \sum_{j \in J_1} \beta_j$ Then, Since $\sum_{i=1}^{d+2} \beta_i x_i = \vec{0}$.

$$\sum_{j \in J_1} \frac{\beta_j}{\beta} x_j = \sum_{j \in J_2} -\frac{\beta_j}{\beta} x_j.$$

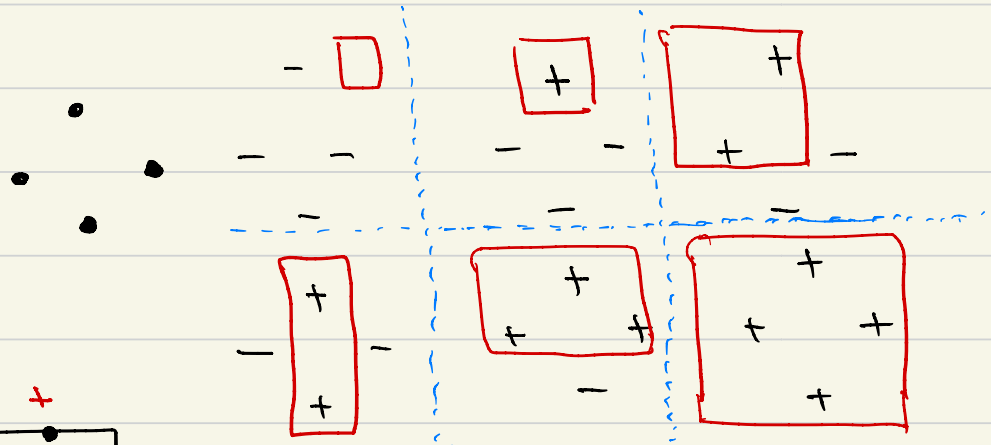
This means that $y = \sum_{j \in J_1} \frac{\beta_j}{\beta} x_j$ belongs to the convex hull of X_1 and also to that of X_2. Q.

③ Axis-aligned rectangle.

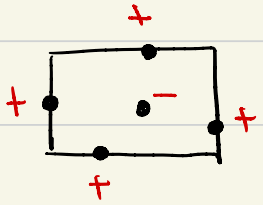
$\mathcal{X} = \mathbb{R}^2$, $\mathcal{H} = \{ x \mapsto$ if $x \in [a,r] \times [b,t]$ then 1 else $-1 \mid a,r,b,t \in \mathbb{R} \}$.

VCDim$(\mathcal{H}) = 4$.

Why ≥ 4?



Why < 5?



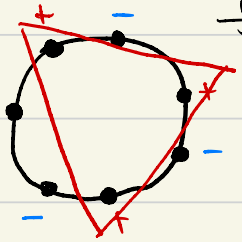
We can do this kind of labelling which no axis-aligned rectangle can get right.

④ Convex Polygon with d vertices. $\mathcal{X} = \mathbb{R}^2$. $\mathcal{H}$ = fns generated by convex d-gons.

VCDim$(\mathcal{H}) = 2d+1$.

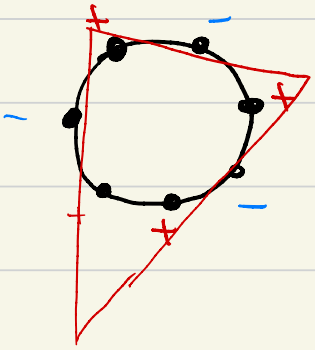
Why $\geq 2d+1$?

equally spaced points →



fewer positive points.

Start from a triangle touching positive points and extend it.



fewer negative points.

Start from a triangle touching negative points and shrink it.

⑤ Sine functions. $\mathcal{X} = \mathbb{R}$. $\mathcal{H} = \{ x \mapsto -\text{sgn}(\sin(wx)) \mid w \in \mathbb{R} \}$.

VCDim$(\mathcal{H}) = \infty$.

(5) Generalization bound with VC dim.

① Assume that $VCDim(\mathcal{H}) = d$. Then,

$$\Pi_{\mathcal{H}}(m) \le \left(\frac{em}{d}\right)^d \quad \text{(Cor. 3.18)}.$$

Thus, we can derive a uniform generalization bound involving d (VC-dim.) from the one based on the growth fn., which we recall below.

$$\mathbb{P}_{S\sim D^m}\left[\ \forall h\in\mathcal{H}.\ R(h) \le \hat{R}_S(h) + \sqrt{\frac{2\log \Pi_{\mathcal{G}}(m)}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{m}}\ \right] \ge 1-\delta.$$

where D is a dist. over $\mathcal{X}\times\mathcal{Y}$. That is, we can get:

$$\mathbb{P}\left[\ \forall h\in\mathcal{H}.\ R(h) \le \hat{R}_S(h) + \sqrt{\frac{2d\log^{em/d}}{m}} + \sqrt{\frac{\log 1/\delta}{m}}\ \right] \ge 1-\delta.$$

(Cor 3.19)

② Here is more detailed and more formal explanation.

Thm 3.17 [Sauer's Lemma]

$VCDim(\mathcal{H}) = d$ implies that $\Pi_{\mathcal{H}}(m) \le \sum_{i=0}^{d}\binom{m}{i}$ for all m.

Proof. Induction on $m+d$.

Base case: $d=0$. Then, $|\mathcal{H}|=1$. Thus, $\Pi_{\mathcal{H}}(m)=1$, and the thm holds.

$m=1$ and $d>0$. Then, $\sum_{i=0}^{d}\binom{m}{i} = \sum_{i=0}^{d}\binom{1}{i} \ge 2$. But $\Pi_{\mathcal{H}}(1) \le 2$.

So, the thm holds in this case as well.

Inductive case: $d>0$ and $m>1$.

for $S_0 \subseteq \mathcal{X}$, $\mathcal{H}|_{S_0} \overset{def.}{=} \{\, h|_{S_0} \in [S_0 \to \mathcal{Y}] \mid h\in\mathcal{H}\}$.

Pick $S\subseteq\mathcal{X}$ with $|S|=m$ s.t. $\Pi_{\mathcal{H}}(m) = |\mathcal{H}|_S|$.

Let $(x_1,\cdots,x_m)$ be the enumeration of elements of S.

Let $S' = \{x_1, \dots, x_{m-1}\}$. (i.e., $S' = S \setminus \{x_m\}$).

and $\mathcal{H}' = \{ h \in \mathcal{H}|_{S'} \mid \exists h_1, h_2 \in \mathcal{H}|_S$ s.t.

$h = h_1|_{S'} = h_2|_{S'}$ and

$h_1(x_m) = +1$ and $h_2(x_m) = -1 \}$.

Then,

$$|\mathcal{H}'| + |\mathcal{H}|_{S'}| = |\mathcal{H}|_S| = \Pi_{\mathcal{H}}(m)$$

Also, $VCDim(\mathcal{H}|_{S'}) \le VCDim(\mathcal{H}) = d$. and

$VCDim(\mathcal{H}') \le VCDim(\mathcal{H}|_S) - 1 \le VCDim(\mathcal{H}) - 1 = d-1$.

Thus,

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

$$= \Pi_{\mathcal{H}'}(m-1) + \Pi_{\mathcal{H}|_{S'}}(m-1)$$

Ind. hypo $\rightarrow$ $$\le \sum_{i=0}^{d-1} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i} = \sum_{i=0}^{d} \binom{m-1}{i-1} + \binom{m-1}{i}$$

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

□

Cor 3.18. $VCDim(\mathcal{H}) = d$ implies that

$$\Pi_{\mathcal{H}}(m) = \left(\frac{em}{d}\right)^d = O(m^d)$$

for all $m \ge d$.

Proof

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

$$\le \sum_{i=0}^{d} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i}$$

$$\le \sum_{i=0}^{m} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} = \left(\frac{m}{d}\right)^d \left(1 + \frac{d}{m}\right)^m$$

$$\le \left(\frac{m}{d}\right)^d e^{m \cdot \frac{d}{m}} = \left(\frac{em}{d}\right)^d$$

□

Cor 3.19. VCDim$(\mathcal{H}) = d$. $\delta > 0$. $D \in \Pr(X \times Y)$.

$\Rightarrow$ $\mathbb{P}_{S \sim D^m}\left[\forall h \in \mathcal{H}.\ R(h) \le \hat{R}_S(h) + \sqrt{\frac{d \log(em/d)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}\right] \ge 1 - \delta$.

Proof

By Cor. 3.18 and Cor 3.9. D.

$$R(h) \le \hat{R}_S(h) + O\left(\sqrt{\frac{\log m/d}{m/d}}\right).$$

So, the smaller d, the tighter the bound. Related to Occam's razor principle.