Lecture-09: Growth functions and VC-dimension

1 Growth function

Rademacher complexity can be bounded in terms of the growth function.

Definition 1.1 (Dichotomy). A *dichotomy* of an unlabeled sample $x \in \mathcal{X}^m$ using a hypothesis $h \in H \subseteq \mathcal{Y}^{\mathcal{X}}$ is the generated label sequence $h_x \triangleq (h(x_1),\dots,h(x_m)) \in \mathcal{Y}^m$. For a hypothesis set $H \subseteq \mathcal{Y}^{\mathcal{X}}$, the set of dichotomies of sample $x \in \mathcal{X}^m$, is the set of m-length label sequences $H_x \triangleq \{h_x : h \in H\} \subseteq \mathcal{Y}^m$.

Definition 1.2 (Growth function). For a hypothesis set H, the *growth function* $\Pi_H : \mathbb{Z}_+ \to \mathbb{Z}_+$ is defined as

$$\Pi_H(m) \triangleq \max_{x \in \mathcal{X}^m} |H_x| = \max_{x \in \mathcal{X}^m} |\{h_x : h \in H\}|.$$

Remark 1. Growth function is a purely combinatorial measure, and the following holds true for it.
(a) It is the maximum number of distinct ways in which m points can be classified using hypotheses in H. Note that it is maximum and not supremum, since there are finitely many elements in each set H_x. Specifically, $|H_x| \leqslant |\mathcal{Y}|^m$.
(b) It is a measure of richness of the hypothesis set H.
(c) It doesn't depend on the unknown distribution D, unlike Rademacher complexity.

Lemma 1.3 (Massart). *Consider a finite set $A \subset \mathbb{R}^m$ with $r \triangleq \max_{u \in A} \|u\|_2$, and independent Rademacher random vector $\sigma : \Omega \to \{-1,1\}^m$. Then, we have $\mathbb{E}[\frac{1}{m}\sup_{u \in A} \langle \sigma, u \rangle] \leqslant \frac{r}{m}\sqrt{2\ln|A|}$.*

Proof. Fix $t > 0$. Applying Jensen's inequality to the convex function $f(x) = e^{tx}$, rearranging terms, upper bounding the supremum of positive numbers by its sum, and linearity of expectation, we obtain

$$e^{t\mathbb{E}\sup_{x \in A}\langle \sigma, x\rangle} \leqslant \mathbb{E}e^{t\sup_{x \in A}\langle \sigma, x\rangle} = \mathbb{E}\sup_{x \in A} e^{t\langle \sigma, x\rangle} \leqslant \mathbb{E}\sum_{x \in A} e^{t\langle \sigma, x\rangle} = \sum_{x \in A} \mathbb{E}e^{t\langle \sigma, x\rangle}.$$

From the independence of Rademacher random vector σ, the application of Hoeffding lemma for each product term where $-t|x_i| \leqslant t\sigma_i x_i \leqslant t|x_i|$ for all $i \in [m]$, and the definition of r, we get

$$e^{t\mathbb{E}\sup_{x \in A}\langle \sigma, x\rangle} \leqslant \sum_{x \in A} \mathbb{E}[e^{t\langle \sigma, x\rangle}] \leqslant \sum_{x \in A}\prod_{i=1}^m \mathbb{E}[e^{t\sigma_i x_i}] \leqslant \sum_{x \in A}\prod_{i=1}^m e^{\frac{4t^2x_i^2}{8}} \leqslant \sum_{x \in A} e^{\frac{t^2}{2}\|x\|_2^2} \leqslant |A| e^{\frac{t^2r^2}{2}}.$$

Taking the natural log of both sides and dividing by t, we get $\mathbb{E}\sup_{x \in A}\langle \sigma, x\rangle \leqslant \frac{1}{t}\ln|A| + \frac{tr^2}{2}$. The upper bound is minimized by taking $t^* = \frac{1}{r}\sqrt{2\ln|A|}$. We get the result by dividing the both sides of this minimized upper bound by m. □

Corollary 1.4. *For binary labels $\mathcal{Y} \triangleq \{-1,1\}$ and hypothesis set $H \subset \mathcal{Y}^{\mathcal{X}}$, we have $\mathcal{R}_m(H) \leqslant \sqrt{\frac{2}{m}\ln\Pi_H(m)}$.*

Proof. Recall that $h_x \triangleq (h(x_1),\dots,h(x_m))) \in \mathcal{Y}^m$ for any unlabeled sample $x \in \mathcal{X}^m$ and hypothesis $h \in H$. For a fixed sample x and hypothesis set H, we denote the dichotomy set by $H_x \triangleq \{h_x : h \in H\} \subseteq \mathcal{Y}^m$. Any vector $y \in \mathcal{Y}^m$ has norm $\|y\|_2 = \sqrt{m}$. Applying Massart's lemma to the finite set H_x, we get

$$\mathcal{R}_m(H) = \mathbb{E}_x\hat{\mathcal{R}}_x(H) = \mathbb{E}_x\mathbb{E}_\sigma\sup_{h \in H}\frac{1}{m}\langle \sigma, h_x\rangle = \mathbb{E}_x\mathbb{E}_\sigma\sup_{u \in H_x}\frac{1}{m}\langle \sigma, u\rangle \leqslant \mathbb{E}\sqrt{\frac{2}{m}\ln|H_x|}.$$

By definition, we have $|H_x| \leqslant \Pi_H(m)$, and hence the result follows. □

Corollary 1.5 (Growth function generalization bound). *Consider hypothesis set $H \subset \mathcal{Y}^{\mathcal{X}}$ for binary labels $\mathcal{Y} \triangleq \{-1,1\}$. Then, for any $\delta > 0$*

$$P\Big(\bigcap_{h \in H}\left\{R(h) \leqslant \hat{R}(h) + \sqrt{\frac{2}{m}\ln\Pi_H(m)} + \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}\right\}\Big) \geqslant 1 - \delta.$$

Remark 2. Growth function bounds can be also derived directly without using Rademacher complexity bounds. The resulting bound is $P\left\{\left|R(h)-\hat{R}(h)\right|>\epsilon\right\} \leqslant 4\Pi_H(2m)e^{-\frac{m\epsilon^2}{8}}$. The generalization bound obtained from this bound differs from Corollary **??** only in constants.

Remark 3. The computation of the growth function may not be always convenient since, by definition, it requires computing $\Pi_H(m)$ for all $m \in \mathbb{N}$.

2 Vapnik-Chervonenkis (VC) dimension

The VC-dimension is also a purely combinatorial notion but it is often easier to compute than the growth function or the Rademacher Complexity. We will consider the target space $\mathcal{Y} = \{-1,1\}$ in the following.

Definition 2.1 (Shattering). An unlabeled sample $x \in \mathcal{X}^m$ is said to be *shattered* by a hypothesis set $H \subseteq \mathcal{Y}^{\mathcal{X}}$ when this set realizes all possible dichotomies of x, that is when $|H_x| = |\mathcal{Y}|^m$.

Definition 2.2 (VC-dimension). The *VC-dimension* of a hypothesis set H is the size of the largest unlabeled sample that can be fully shattered by H. That is, $\text{VC-dim}(H) \triangleq \max\{m \in \mathbb{Z}_+ : \Pi_H(m) = 2^m\}$.

Remark 4. By definition $\text{VC-dim}(H) = d$ implies that there exists an unlabeled sample $x \in \mathcal{X}^d$ of size d that can be fully shattered, i.e. $|H_x| = |\mathcal{Y}|^d$. This does not imply that all unlabeled samples of size d or less are fully shattered. In fact, this is typically not the case. It is easy to see that if no unlabeled samples of size m are fully shattered, then no unlabeled samples of size $m+1$ can be fully shattered.

Remark 5. To compute the VC-dim-dimension we will typically show a lower bound for its value and then a matching upper bound. To show a lower bound d for $\text{VC-dim}(H)$, it suffices to show that a sample $x \in \mathcal{X}^d$ can be shattered by hypothesis set H. To show an upper bound, we need to prove that no sample $x \in \mathcal{X}^{d+1}$ can be shattered by hypothesis set H. This step is typically more difficult.

Example 2.3 (Intervals on the real line). For binary labels $\mathcal{Y} \triangleq \{-1,1\}$ and input space $\mathcal{X} = \mathbb{R}$, consider a hypothesis set $H \subseteq \mathcal{Y}^{\mathbb{R}}$ of separating intervals on real line $\mathbb{R}$ defined as

$$H \triangleq \left\{x \mapsto \mathbb{1}_{[a,b]}(x) - \mathbb{1}_{[a,b]^c}(x) : a,b \in \mathbb{R}\right\} \subseteq \mathcal{Y}^{\mathbb{R}}.$$

We observe that for $d = 2$, possible dichotomies are $\mathcal{Y}^d = \{(-1,-1),(-1,1),(1,-1),(1,1)\}$. Let $x \in \mathbb{R}^d$, then we can find $a,b \in \mathbb{R}$ such that corresponding $h^{a,b} \in H$ achieves any dichotomy in $\mathcal{Y}^d$. To show this, we can assume that $x_1 < x_2$ without any loss of generality, and observe that for any $h^{a,b} \in H$

$$h_x^{a,b} = \begin{cases} (-1,-1), & x_2 < a \text{ or } x_1 > b, \\ (-1,1), & x_1 < a < x_2 < b, \\ (1,-1), & a < x_1 < b < x_2, \\ (1,1), & a < x_1 < x_2 < b. \end{cases}$$

Further, for any sample $x \in \mathbb{R}^3$ such that $x_1 < x_2 < x_3$ there is no $a,b \in \mathbb{R}$ such that $h_x^{a,b} = (1,-1,1)$. That is, no set of three points can be shattered, and hence $\text{VC-dim}(H) = 2$.

Remark 6. The VC-dimension of any vector space of dimension $r < \infty$ can be shown to be at most r.

Theorem 2.4 (Sauer). *Consider hypothesis set $H \subseteq \mathcal{Y}^{\mathcal{X}}$ for binary labels $\mathcal{Y} \triangleq \{-1,1\}$ with* $\text{VC-dim}(H) = d$. *Then, we have $\Pi_H(m) \leqslant \sum_{i=0}^{d} \binom{m}{i}$ for all $m \in \mathbb{N}$.*

Proof. The proof is by induction on the pair (m,d). If $d = 0$, then $\Pi_H(1) < 2$ for all points $x \in \mathcal{X}$, which implies H consists of single function, and therefore the upper bound of unity holds for $m = 1$. If $d = 1$, then $\Pi_H(2) < 4$ and $\Pi_H(1) = 2$, and the upper bound of $1 + m = 2$ holds for $m = 1$. Therefore, the statement holds true for the pairs $(m,d) = (1,1)$ and $(m,d-1) = (1,0)$.

We assume that the inductive hypothesis holds true for $(m-1,d-1)$ and $(m-1,d)$. Let $x \in \mathcal{X}^m$ be the sample with $\Pi_H(m)$ dichotomies. That is, $|H_x| = \Pi_H(m)$. We define $G \triangleq \{g \in H : g_x \in H_x\}$, and hence $\text{VC-dim}(G) \leqslant \text{VC-dim}(H) = d$. Further, we observe that $G_x = H_x$ and hence $\pi_G(m) = \pi_H(m)$. Consider the subsample $x' = (x_1,\ldots,x_{m-1})$, the corresponding dichotomy set $H_{x'}$, and define $G^1 \triangleq$

$\{g \in H : g_{x'} \in H_{x'}\}$. It follows that VC-dim$(G^1) \leqslant$ VC-dim$(H) = d$ and together with induction hypothesis, we obtain

$$|H_{x'}| = \left|G^1_{x'}\right| \leqslant \pi_{G^1}(m-1) \leqslant \sum_{i=0}^{d} \binom{m-1}{i}.$$

We define projection operator $\pi : \mathcal{Y}^m \to \mathcal{Y}^{m-1}$ for all $y \in \mathcal{Y}^{m-1}$ as $\pi(y) = (y_1,\dots,y_{m-1})$. For each $v \in H_{x'}$, we have $\pi^{-1}(v) = \{(v,-1),(v,1)\}$ and $\pi^{-1}(v) \cap H_x \neq \emptyset$, and hence we can find a set $H^1_x \subseteq H_x$ that is bijective to $H_{x'}$ and $\pi(H^1_x) = H_{x'}$. We define $H^2_x \triangleq H_x \setminus H^1_x$ such that for each $v \in H^2_x$ we have $\pi^{-1} \circ \pi(v) = \{(v,-1),(v,1)\}$. We can find a bijection from H^2_x to the set $H^2_{x'} \triangleq \{v \in H_{x'} : \pi^{-1}(v) = \{(v,-1),(v,1)\}\}$. We define hypothesis set $G^2 \triangleq \{g \in H : g_{x'} \in H^2_{x'}\}$ such that $G^2_{x'} = H^2_{x'}$. From the definition of G^2, we have if $\left|G^2_y\right| = \pi_{G_2}(k) = 2^k$ for $y \in \mathcal{Y}^k$ and $k \leqslant m-1$, then $\left|G^2_{y,x_m}\right| = 2\pi_{G_2}(k) = 2^{k+1}$ and hence VC-dim$(G_2) + 1 \leqslant$ VC-dim(G). Together with the induction hypothesis, it follows that

$$\left|G^2_{x'}\right| = \left|H^2_{x'}\right| \leqslant \pi_{G_2}(m-1) \leqslant \sum_{i=0}^{d-1} \binom{m-1}{i}.$$

Since $\pi_H(m) = |H_x| = \left|H^1_x\right| + \left|H^2_x\right|$, we obtain the result for (m,d). □

Corollary 2.5. *Let H be a hypothesis set with* VC-dim$(H) = d$*, then*

$$\Pi_H(m) \leqslant \left(\frac{em}{d}\right)^d = O(m^d), \textit{ for all } m \geqslant d.$$

Proof. For $m \geqslant d$ and $0 \leqslant i \leqslant d$, we have $(\frac{m}{d})^{d-i} \geqslant 1$. Therefore,

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

From Binomial theorem, we get $\sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i = \left(1 + \frac{d}{m}\right)^m$. Since $1 + x \leqslant e^x$ for all $x \in \mathbb{R}$, we get $\left(1 + \frac{d}{m}\right)^m \leqslant e^d$, and hence the result follows. □

Remark 7. The growth function only exhibits two types of behavior,

(i) either VC-dim$(H) = d < \infty$, in which case $\Pi_H(m) = O(m^d)$,

(ii) or VC-dim$(H) = \infty$, in which case $\Pi_H(m) = 2^m$ for all $m \in \mathbb{N}$.

Corollary 2.6 (VC-dimension generalization bounds). *Consider hypothesis set $H \subset \mathcal{Y}^{\mathcal{X}}$ for binary labels $\mathcal{Y} \triangleq \{-1,1\}$ with VC-dimension d. Then, for any $\delta > 0$*

$$P\Big(\bigcap_{h\in H} \left\{R(h) \leqslant \hat{R}(h) + \sqrt{\frac{2d}{m}\ln\frac{em}{d}} + \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}\right\}\Big) \geqslant 1 - \delta.$$

Remark 8. With high probability, we observe the following for the generalization risk $R(h)$.

(i) Generalization risk is of the form $R(h) \leqslant \hat{R}(h) + O\left(\sqrt{\frac{\ln(m/d)}{m/d}}\right)$, signifying the importance of the ratio $\frac{m}{d}$.

(ii) Without the intermediate step of Rademacher complexity, a direct bound on generalization risk can be obtained as

$$\hat{R}(h) + \sqrt{\frac{1}{m}(8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta})}.$$