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). Given a hypothesis set H, a **dichotomy** of a sample $x \in \mathcal{X}^m$ is one of the possible ways of labeling the points of sample x using a hypothesis $h \in H$, and denoted by $h_x \triangleq (h(x_1), \dots, h(x_m)) \in \mathcal{Y}^m$.

Definition 1.2 (Dichotomy set). For hypothesis set H, the set of dichotomies of sample $x \in \mathcal{X}^m$, is the set of m-length $\mathcal{Y}$-valued sequences $H_x \triangleq \{h_x : h \in H\} \subseteq \mathcal{Y}^m$.

Definition 1.3 (Growth function). For a hypothesis set H, the **growth function** $\Pi_H : \mathbb{N} \to \mathbb{N}$ 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.
(b) It is the maximum number of dichotomies for m points using hypotheses in H.
(c) It is a measure of richness of the hypothesis set H.
(d) It doesn't depend on the unknown distribution D, unlike Rademacher complexity.

Lemma 1.4 (Massart). *Consider a finite set $A \subset \mathbb{R}^m$ with $r \triangleq \max_{x \in A} \|x\|_2$, and independent Rademacher random vector $\sigma : \Omega \to \{-1,1\}^m$. Then, we have $\mathbb{E}\left[\frac{1}{m}\sup_{x \in A} \langle \sigma, x \rangle\right] \leqslant \frac{r\sqrt{2\ln|A|}}{m}$.*

Proof. For any $t > 0$, using Jensen's inequality for the convex function $f(x) = e^{tx}$, rearranging terms, and bounding the supremum of positive numbers by its sum, we obtain

$$\exp\left(t\mathbb{E}\left[\sup_{x \in A} \langle \sigma, x \rangle\right]\right) \leqslant \mathbb{E}\left[\exp\left(t \sup_{x \in A} \langle \sigma, x \rangle\right)\right] = \mathbb{E}\left[\sup_{x \in A} e^{t\langle \sigma, x \rangle}\right] \leqslant \mathbb{E}\left[\sum_{x \in A} e^{t\langle \sigma, x \rangle}\right].$$

From the independence of Rademacher random vector σ, the application of Hoeffding lemma to independent random vector $t\sigma \circ x$ such that $-t|x_i| \leqslant t\sigma_i x_i \leqslant t|x_i|$, and the definition of r, we get

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

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

Corollary 1.5. *Let $G \subset \{-1,1\}^{\mathcal{X}}$ be a family of functions, then $\mathcal{R}_m(G) \leqslant \sqrt{\frac{2\ln \Pi_G(m)}{m}}$.*

Proof. For a fixed sample $x = (x_1, \dots, x_m) \in \mathcal{X}^m$, we denote $g_x \triangleq (g(x_1), \dots, g(x_m))) \in \mathcal{Y}^m$ for any $g \in G$. Therefore, we can write the restriction of G to sample x, as $G_x \triangleq \{g_x : g \in G\}$. Since $g \in G$ takes values in $\{-1,1\}$, the norm of these vectors is $\sqrt{m}$. Applying Massart's lemma to the restricted set G_x, we get

$$\mathcal{R}_m(G) = \mathbb{E}_x \hat{\mathcal{R}}_x(G) = \mathbb{E}_x \mathbb{E}_\sigma \left[\sup_{g \in G} \frac{1}{m} \langle \sigma, g_x \rangle\right] = \mathbb{E}_{\sigma,x}\left[\sup_{u \in G_x} \frac{1}{m} \langle \sigma, u \rangle\right] \leqslant \mathbb{E}\left[\sqrt{\frac{2\ln|G_x|}{m}}\right].$$

By definition, we have $|G_x| \leqslant \Pi_G(m)$, and hence the result follows. ☐

Corollary 1.6 (Growth function generalization bound). *Let $H \subset \mathcal{Y}^{\mathcal{X}}$ be a family of functions where $\mathcal{Y} = \{-1,1\}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, for any hypothesis $h \in H$*

$$R(h) \leqslant \hat{R}(h) + \sqrt{\frac{2\ln \Pi_H(m)}{m}} + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

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). A sample $x \in \mathcal{X}^m$ is said to be **shattered** by a hypothesis set H when H realizes all possible dichotomies of x, that is when $\Pi_H(m) = 2^m$.

Definition 2.2 (VC-dimension). The **VC-dimension** of a hypothesis set H is the size of the largest set that can be fully shattered by H. That is,

$$\text{VC-dim}(H) \triangleq \max\{m \in \mathbb{N} : \Pi_H(m) = 2^m\}.$$

Remark 4. By definition, if VC-dim$(H) = d$, there exists a set of size d that can be fully shattered. This does not imply that all sets of size d or less are fully shattered, in fact, this is typically not the case.

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

Example 2.3 (Intervals on the real line). Consider a hypothesis set H of separating intervals on real line

$$H \triangleq \left\{h \in \{-1,1\}^{\mathbb{R}} : h = \mathbb{1}_{[a,b]} - \mathbb{1}_{[a,b]^c}, a, b \in \mathbb{R}\right\}.$$

Then $d \geqslant 2$, since $(1,1), (-1,-1), (1,-1), (-1,1)$ can all be realized by $x \in \mathbb{R}^2$. Further, there is no sample $x \in \mathbb{R}^3$ such that $x_1 < x_2 < x_3$ and $h_S = (1,-1,1)$. That is, no set of three points can be shattered, and hence 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). *Let $H \subseteq \{-1,1\}^{\mathcal{X}}$ have* VC-dim$(H) = d$*. Then, we have $\Pi_H(d) \leqslant \sum_{i=0}^{d}\binom{m}{i}$, for all $m \in N$.*

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. If $d = 1$, then $\Pi_H(2) < 4, \Pi_H(1) = 2$, and the upper bound of $1 + m = 2$ holds. Therefore, the statement holds true for the pairs $(m,d) = (1,1)$ and $(m,d-1) = (1,0)$.

We assume that the inductive hypothesis is 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, $|H_x| = |\{h_x : h \in H\}| = \Pi_H(m)$. We can partition the hypothesis set H by the vectors $h_x \in H_x$, by defining equivalence classes $H(h_x) \triangleq \{g \in H : g_x = h_x\}$. Consider the subsample $x' = (x_1, \dots, x_{m-1})$, and the corresponding set of dichotomies $H_{x'}$. For each $h_x \in H_x$, there is a projection

and denote projection operator $\pi : \mathbb{R}^S \to \mathbb{R}^{S'}$. We consider the two family of functions

$$G_1 = H|_{S'} = \{\pi \circ g : g \in G\}, \qquad G_2 = \left\{g' \in G_1 : \left|\pi^{-1} \circ g'\right| = 2\right\}.$$

It follows that there exists functions $g_1, g_2 \in G$ such that $g_1|_{S'} = g_2|_{S'}$. In particular, $g_1(x_m) \neq g_2(x_m)$ but they agree on all other points $S' \subset S$. It follows that $|G| = |G_1| + |G_2|$.

Since $G_1 \subset G$, it follows that $\text{VC-dim}(G_1) \leqslant \text{VC-dim}(G) \leqslant d$, then by the definition of growth function and induction hypothesis,

$$|G_1| \leqslant \Pi_{G_1}(m-1) \leqslant \sum_{i=0}^{d} \binom{m-1}{i}.$$

Further, by definition of G_2, if a set $Z \subseteq S'$ is shattered by G_2, then the set $Z \cup \{x_m\}$ is shattered by G. Therefore,

$$\text{VC-dim}(G_2) \leqslant \text{VC-dim}(G) - 1 = d - 1.$$

From the definition of growth function and induction hypothesis,

$$|G_2| \leqslant \Pi_{G_2}(m-1) \leqslant \sum_{i=0}^{d-1} \binom{m-1}{i}.$$

Since $|G| = |G_1| + |G_2|$, we have

$$|G| \leqslant \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} = \sum_{i=0}^{d} \left(\binom{m-1}{i} + \binom{m-1}{i-1} \right) = \sum_{i=0}^{d} \binom{m}{i}.$$

Hence, the result holds for (m, d). □

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

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

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

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

Since the summation of positive terms over $i \in \{0, \ldots, d\}$ can be upper bounded by summation over $i \in \{0, \ldots, m\}$, we get $(\frac{m}{d})^d \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^i \leqslant (\frac{m}{d})^d \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i$. From the 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 $\text{VC-dim}(H) = d < \infty$, in which case $\Pi_H(m) = O(m^d)$,

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

Corollary 2.6 (VC-dimension generalization bounds). *Let $H \subset \{-1, 1\}^{\mathcal{X}}$ be a family of functions with VC-dimension d. Then, for any $\delta > 0$, with probability at least $1 - \delta$*

$$R(h) \leqslant \hat{R}(h) + \sqrt{\frac{2d \ln \frac{em}{d}}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}, \text{ for all } h \in H.$$

Remark 8. (i) Generalization risk is of the form $R(h) \leqslant \hat{R}(h) + O\left(\sqrt{\frac{\ln(m/d)}{m/d}}\right)$, which implies that the ratio $\frac{m}{d}$ is important.

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

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