Lecture-08: Growth functions and VC-dimension

1 Growth function

Rademacher complexity can be bounded in terms of the growth function. For any hypothesis $h \in H$ and an m-sized unlabeled sample set $S = \{x_1, \dots, x_m\} \subseteq \mathcal{X}$, we denote the range $h_S \triangleq (h(x_1), \dots, h(x_m)) \in \mathcal{Y}^m$. For different hypothesis $h \in H$ and a fixed m-sized sample set S, we get a set of m-length $\mathcal{Y}$-valued sequences $\{h_S : h \in H\}$.

Definition 1.1 (Dichotomy). Given a hypothesis set H, a **dichotomy** of a set S is one of the possible ways of labeling the points of S using a hypothesis in H.

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

$$\Pi_H(m) \triangleq \max_{S \subseteq \mathcal{X} : |S| = m} |\{h_S : h \in H\}|.$$

Remark 1. Following is true for growth function.

(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 is a purely combinatorial measure, and unlike Rademacher complexity, it doesn't depend on the unknown distribution D.

Lemma 1.3 (Massart's lemma). *Let $A \subset \mathbb{R}^m$ be a finite set with $r = \max_{x \in A} \|x\|_2$, then*

$$\mathbb{E}\left[\frac{1}{m} \sup_{x \in A} \langle \sigma, x \rangle\right] \leqslant \frac{r\sqrt{2 \ln |A|}}{m},$$

where $\sigma : \Omega \to \{-1, 1\}^m$ is an independent Rademacher random vector.

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

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

From the independence of Rademacher random vector σ, the application of Hoeffding lemma to random variables $-tx_i \leqslant t\sigma_i x_i \leqslant tx_i$, and the definition of r, we get

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

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 1.4. *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 $S = (x_1, \dots, x_m) \in \mathcal{X}^m$, we denote

$$G|_S \triangleq \{g_S = (g(x_1), \dots, g(x_m))) : g \in G\}.$$

Since $g \in G$ takes values in $\{-1, 1\}$, the norm of these vectors is bounded by $\sqrt{m}$. Applying Massart's lemma to the set G, we get

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

By definition, we have $|G|_S| \leqslant \Pi_G(m)$, and hence the result follows. □

Corollary 1.5 (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

$$P\left\{\left|R(h) - \hat{R}(h)\right| > \varepsilon\right\} \leqslant 4\Pi_H(2m)e^{-\frac{m\varepsilon^2}{8}}.$$

The generalization bound obtained from this bound differs from Corollary 1.5 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 set S of $m \in \mathbb{N}$ points is said to be **shattered** by a hypothesis set H when H realizes all possible dichotomies of S, 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,

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

Remark 4. By definition, if $\text{VCdim}(H) = d$, 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 $\text{VCdim}(H)$, it suffices to show that a set S of cardinality d can be shattered by H. To give an upper bound, we need to prove that no set S of cardinality $d+1$ can be shattered by H, which 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 $S = \{x_1, x_2\}$. Further, there is no sample $S = \{x_1, x_2, x_3\}$ such that $x_1 \leqslant x_2 \leqslant x_3$ and $h_S = (1, -1, 1)$. That is, no set of three points can be shattered, and hence $\text{VCdim}(H) = 2$.

Example 2.4 (Hyperplanes in $\mathbb{R}^2$). Consider a hypothesis set H of separating hyperplanes in $\mathbb{R}^2$

$$H \triangleq \left\{h \in \{-1,1\}^{\mathbb{R}^2} : h = \operatorname{sign}(w_1x_1 + w_2x_2 + b), w \in \mathbb{R}^2, b \in \mathbb{R}\right\}.$$

Lower bound: $\operatorname{VCdim}(H) \geqslant 3$**:** Any three non-collinear points in $\mathbb{R}^2$ can be shattered. To obtain the first three dichotomies, we choose a hyperplane that has two points on one side and the third point on the opposite side. Fourth dichotomy has all three points on the same side of the hyperplane. Rest four dichotomies can be obtained by permutation of signs.
Upper bound: $\operatorname{VCdim}(H) < 4$**:** Four points cannot be shattered by considering two cases:

(i) The four points lie on the convex hull defined by the four points. A positive labeling for one diagonal pair and a negative labeling for the other diagonal pair cannot be realized

(ii) Three of the four points lie on the convex hull and the remaining point is internal. A labeling which is positive for the points on the convex hull and negative for the interior point cannot be realized.

Theorem 2.5 (Radon). *Any set $X \subset \mathbb{R}^d$ with $|X| = d+2$ can be partitioned into two subsets X_1 and X_2 such that the convex hulls of X_1 and X_2 intersect.*

Proof. Let $X = \{x_1, \dots, x_d + 2\} \subset \mathbb{R}^d$. The following is a system of $d+1$ linear equations in $\alpha \in \mathbb{R}^{d+2}$

$$\sum_{i=1}^{d+2} \alpha_i x_i = 0, \qquad \sum_{i=1}^{d+2} \alpha_i = 0,$$

since first equality lead to d equations, one for each component. The number of unknowns α is larger than the number of equations, therefore the system admits a non-zero solution $\beta \in \mathbb{R}^{d+2}$ such that $\sum_{i=1}^{d+2} \beta_i = 0$. We find the non-empty sets

$$I_+ \triangleq \{i \in [d+2] : \beta_i > 0\}, \qquad I_- \triangleq \{i \in [d+2] : \beta_i < 0\}.$$

Thus, we can find partition of the set X as $X_1 = \{x_i \in X : i \in I_+\}$ and $X_2 = \{x_i \in X : i \in I_-\}$. We define $b \triangleq \sum_{i \in I_+} \beta_i = \sum_{i \in I_-} -\beta_i$, then we have

$$y = \sum_{i \in I_+} \frac{\beta_i}{b} x_i = \sum_{i \in I_-} \frac{-\beta_i}{b} x_i,$$

where $\sum_{i \in I_+} \frac{\beta_i}{b} = \sum_{i \in I_-} \frac{-\beta_i}{b} = 1$ and $\frac{\beta_i}{b} > 0$ for all $i \in I_+$ and $\frac{-\beta_i}{b} > 0$ for all $i \in I_-$. Thus, we have found an element y in the convex hull of both sets X_1 and X_2. □

Example 2.6 (Hyperplanes). Consider the hypothesis set H to be the set of separating hyperplanes in $\mathbb{R}^d$, i.e.

$$H \triangleq \left\{h \in \{-1,1\}^{\mathbb{R}^d} : h(x) = \operatorname{sign}(\langle w, x \rangle + b), w \in \mathbb{R}^d, b \in \mathbb{R}\right\}.$$

Lower bound: $\operatorname{VCdim}(H) \geqslant d+1$**:**
We take a sample $S = \{0, e_1, \dots, e_d\} \subseteq \mathbb{R}^d$. Let $y = (y_0, \dots, y_d) \in \{-1,1\}^{d+1}$, then we will find $w \in \mathbb{R}^d, b \in \mathbb{R}$ such that $h(x_i) = y_i$ for each $i \in \{0, \dots, d\}$. We choose $w = (y_1, \dots, y_d)$ and $b = \frac{y_0}{2}$, then

$$\operatorname{sign}(\langle w, x_i \rangle + b) = \operatorname{sign}\left(y_i \mathbb{1}_{\{i \neq 0\}} + \frac{y_0}{2}\right) = y_i \text{ for each } i \in \{0, \dots, d\}.$$

Upper bound: $\operatorname{VCdim}(H) < d+2$**:**
To obtain an upper bound, it suffices to show that no set of $(d+2)$ points can be shattered by separating hyperplanes. From the Radon theorem, for any set X of $(d+2)$ points there exists a partition X_1, X_2 such that their convex hulls intersect. Therefore, there are no hyperplanes in $\mathbb{R}^d$ that separate X_1 and X_2. We define the $y \in \{-1,1\}^{d+2}$ such that for each $i \in [d+2]$

$$y_i = \mathbb{1}_{X_1}(x_i) - \mathbb{1}_{X_2}(x_i).$$

This dichotomy can't be achieved by any separating hyperplane in $\mathbb{R}^d$.

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

Theorem 2.7 (Sauer's). *Let H be a hypothesis set with* $\text{VCdim}(H) = d$. *Then, for all $m \in N$, we have*

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

Proof. The proof is by induction on $m+d$. The statement clearly holds for $m = 1$ and $d = 0$ or $d = 1$. 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 $\Pi_H(2) < 4$ and $\Pi_H(1) = 2$, and the upper bound of $1+m = 2$ holds.

Now, assume that it holds for $(m-1, d-1)$ and $(m-1, d)$. Fix a set $S = \{x_1, \dots, x_m\}$ with $\Pi_H(m)$ dichotomies and let $G = H|_S$ be the set of concepts H induced by restriction to S.

Consider the subsample $S' = \{x_1, \dots, x_{m-1}\} \subset S$ 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 $\text{VCdim}(G_1) \leqslant \text{VCdim}(G) \leqslant d$, 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,

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

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 2.8. *Let H be a hypothesis set with* $\text{VCdim}(H) = d$, *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 $(\frac{m}{d})^{d-i} \geqslant 1$. Further, the summation of positive terms over $i \in \{0, \dots, d\}$ can be upper bounded by summation over $i \in \{0, \dots, m\}$. Therefore,

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

The last equality follows from the Binomial theorem and the following inequality from the fact that $1 + x \leqslant e^x$ for all $x \in \mathbb{R}$. □

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

(i) either $\text{VCdim}(H) = d < \infty$, in which case $\Pi_H(m) = O(m^d)$,

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

Corollary 2.9 (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 $1 - \delta$*

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

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

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