Lecture-11: 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 $h_x \triangleq (h(x_1),\ldots,h(x_m)) \in \mathcal{Y}^m$. 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 $\mathbb{E}\sup_{x\in A}\langle\sigma,x\rangle \leqslant \frac{1}{t}\ln|A| + \frac{tr^2}{2}$ for all $t > 0$. It follows that $\mathbb{E}\sup_{x\in A}\langle\sigma,x\rangle \leqslant \inf_{t>0}\frac{1}{t}\ln|A| + \frac{tr^2}{2}$ and the upper bound is minimized for $t^* \triangleq \frac{1}{r}\sqrt{2\ln|A|}$. We get the result by dividing the both sides of this minimized upper bound by m. □

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

Proof. We fix an unlabeled sample $x \in \mathcal{X}^m$ and hypothesis $h \in H$. Recall that $h_x \triangleq (h(x_1),\ldots,h(x_m))) \in \mathcal{Y}^m$ and 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 $\|y\|_2 = \sqrt{m}$. In particular, any vector $h_x \in H_x$ has norm $\sqrt{m}$. Applying Massart's lemma to the finite set H_x,

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

The result follows from the monotonicity of log and the definition of growth function. □

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 $P\Big(\cup_{h\in H} \big\{ \big|R(h) - \hat{R}(h)\big| > \epsilon \big\} \Big) \leqslant 4\Pi_H(2m)e^{-m\epsilon^2/8}$. Taking $\delta \geqslant 4\Pi_H(2m)e^{-m\epsilon^2/8}$, we get $\sqrt{\frac{8}{m}\ln\frac{4}{\delta} + \frac{8}{m}\ln\Pi_H(2m)} \leqslant \epsilon$. That is, the generalization bound is

$$P\Big(\bigcap_{h\in H} \left\{ R(h) \leqslant \hat{R}(h) + \sqrt{\frac{8}{m}\Big(\ln\frac{4}{\delta} + \ln\Pi_H(2m)\Big)} \right\} \Big) \geqslant 1 - \delta.$$

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 $\mathcal{Y} \triangleq \{-1,1\}$ 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.

Remark 5. We observe that if a sample $x \in \mathcal{X}^{m+1}$ can be fully shattered, i.e. $|H_x| = |\mathcal{Y}|^{m+1}$. That is, for each $y \in \mathcal{Y}^{m+1}$ there exists $h^y \in H$ such that $h^y_x = y$. We take $x' \in \mathcal{X}^m$ such that $x'_i = x_i$ for $i \in [m]$, then $h^y_{x'} = (y_1, \ldots, y_m)$. That is, a subset of size m can also be fully shattered. It follows that if no unlabeled samples of size m are fully shattered, then no unlabeled samples of size $m+1$ can be fully shattered.

Remark 6. To compute the VC-dimension of a hypothesis set, 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

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

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 $h^{a,b} \in H$

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

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 $h^{a,b}_x = (1,-1,1)$. That is, no set of three points can be shattered, and hence $\text{VC-dim}(H) = 2$.

Remark 7. The VC-dimension for hyperplanes in any vector space of dimension $d < \infty$ can be shown to be at most $d + 1$.

Definition 2.4. We define $\Phi_d : \mathbb{Z}_+ \to \mathbb{Z}_+$ as $\Phi_d(m) \triangleq \sum_{i=0}^{d} \binom{m}{i}$ for each $m, d \in \mathbb{Z}_+$.

Lemma 2.5. *For any $d, m \in \mathbb{Z}_+$ the following properties hold for Φ_d.*
(a) $\Phi_0(m) = \Phi_d(0) = 1$.
(b) $\Phi_{d-1}(m-1) + \Phi_d(m-1) = \Phi_d(m)$.
(c) $\Phi_d(m) \leqslant (em/d)^d$ *for* $m \geqslant d$.

Proof. Recall that $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$ for each $m, d \in \mathbb{Z}_+$.
(a) Follows from the definition.
(b) Recall that $\binom{m-1}{i-1} + \binom{m-1}{i} = \binom{m}{i}\left(\frac{i}{m} + \frac{m-i}{m}\right) = \binom{m}{i}$. Summing both sides over $i \in \{0, \ldots, d\}$ and from the definition of $\Phi_d(m)$, we obtain the result.
(c) For $m \geqslant d$ and $0 \leqslant i \leqslant d$, we have $\left(\frac{m}{d}\right)^{d-i} \geqslant 1$. Therefore,

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

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.

□

Theorem 2.6 (Sauer). *Consider hypothesis set $H \subseteq \mathcal{Y}^{\mathcal{X}}$ for binary labels $\mathcal{Y} \triangleq \{-1, 1\}$ with* VC-dim$(H) = d$. *Then, we have $\Pi_H(m) \leqslant \Phi_d(m)$ for all $m \in \mathbb{Z}_+$.*

Proof. The proof is by induction on the pair (m, d). We show the base case for pairs $(m, 0)$ and $(0, d)$. In the inductive step, we show the lemma holds for any m, d with $m + d = k$ for some constant k assuming that it holds for all m, d with $m + d < k$.
(a) *Base case.* For any pair $(m, 0)$ and $(0, d)$, we have $\Phi_0(m) = \Phi_d(0) = 1$. When VC-dimension for hypothesis set H is $d = 0$, it means $1 \leqslant |H_x| \leqslant \sup_{x \in \mathcal{X}} |H_x| < 2$ and hence $\Pi_H(1) = 1 \leqslant 1$. This implies that $|H_x| = 1$ for all points $x \in \mathcal{X}$, which implies that all hypotheses $h \in H$ are a single constant. It follows that $\Pi_H(m) = \sup_{x \in \mathcal{X}^m} |H_x| = 1$ for all $m \geqslant 1$. If $m = 0$, then $\Pi_H(0) = \sup_{x \in \emptyset} |H_x| = 0 \leqslant 1$.
(b) *Inductive case.* Consider a pair (m, d) such that VC-dim$(H) = d$, and we assume that the inductive hypothesis holds true for $(m-1, d-1)$ and $(m-1, d)$. Let $x \in \mathcal{X}^m$ be a sample with $|H_x| = \Pi_H(m)$ dichotomies. For each $y \in H_x$, we find some $h^y \in H$ such that $h^y_x = y$ and define $G \triangleq \{h^y \in H : y \in H_x\} \subseteq H$ and hence VC-dim$(G) \leqslant$ VC-dim$(H) = d$. Consider the subsample $x' \triangleq (x_2, \ldots, x_m)$ and the corresponding dichotomy set $H_{x'} = \{h_{x'} : h \in H\} = \{g_{x'} : g \in G\}$. For each $y' \in H_{x'} \subseteq \mathcal{Y}^{m-1}$, there exists a $g^{y'} \in G$ such that $g^{y'}_{x'} = y'$, and we define

$$G^1 \triangleq \left\{ g^{y'} : y' \in H_{x'} \right\}, \qquad G^2 \triangleq G \setminus G^1.$$

For each $g^2 \in G$ there exists a unique $g^1 \in G^1$ such that $g^1_{x'} = g^2_{x'}$ and $g^1(x_1) \neq g^2(x_2)$.
  (i) We observe that VC-dim$(G^1) \leqslant$ VC-dim$(G) \leqslant d$ and hence $\left|G^1\right| \leqslant \Pi_{G^1}(m-1) \leqslant \Phi_d(m-1)$.
  (ii) Further, if G^2 shatters $(x_2, \ldots, x_d)$ then G shatters $(x_1, \ldots, x_d)$ since for each $g^2 \in G^2$ there exists $g^1 \in G_1 = G \setminus G^1$ such that $g^1(x_1) \neq g^2(x_1)$ and hence VC-dim$(G^2) \leqslant$ VC-dim$(G) - 1 \leqslant d - 1$. It follows that $\left|G^2\right| \leqslant \Pi_{G^2}(m-1) \leqslant \Phi_{d-1}(m-1)$.

Combining the two results and the fact that $\Pi_H(m) = |H_x| = |G| = \left|G^1\right| + \left|G^2\right|$, we obtain $\Pi_H(m) \leqslant \Phi_d(m-1) + \Phi_{d-1}(m-1) = \Phi_d(m)$.

□

Corollary 2.7. *Let H be a hypothesis set with* VC-dim$(H) = d$, *then $\Pi_H(m) \leqslant \left(\frac{em}{d}\right)^d = O(m^d)$, for all $m \geqslant d$.*

Remark 8. 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 2.8 (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 9. 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

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

3 Margin theory

We present generalization bounds for SVM algorithms based on the notion of margin.

Definition 3.1 (Affine hypothesis set). Consider binary label set $\mathcal{Y} \triangleq \{-1, 1\}$, input space $\mathcal{X} \subseteq \mathbb{R}^N$, a labeled sample $z \in (\mathcal{X} \times \mathcal{Y})^m$, and define an affine hypothesis set

$$H \triangleq \left\{x \mapsto \langle w, x\rangle + b : w \in \mathbb{R}^N, b \in \mathbb{R}\right\} \subseteq \mathbb{R}^{\mathcal{X}}.$$

Definition 3.2 (Margin). The geometric margin $\rho(z_i)$ of example $i \in [m]$ with respect to an affine hypothesis $h^{w,b} \in H$ is its distance to the hyperplane $E_{w,b} \triangleq \left\{x \in \mathbb{R}^N : \langle w, x\rangle + b = 0\right\}$. That is,

$$\rho(z_i) \triangleq \frac{y_i h^{w,b}(x_i)}{\|w\|} = \frac{y_i(\langle w, x_i\rangle + b)}{\|w\|}.$$

The margin of an affine classifier $h^{w,b} \in H$ for a labeled sample $z \in (\mathcal{X} \times \mathcal{Y})^m$ is the minimum margin over the points in the sample, i.e. $\rho \triangleq \min\{\rho(z_i) : i \in [m]\}$.

Corollary 3.3. *For any $\delta > 0$ and $H \triangleq \left\{x \mapsto \langle w, x\rangle + b : w \in \mathbb{R}^N, b \in \mathbb{R}\right\}$, we have*

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

Proof. Recall that the VC-dimension of the family of hyperplanes or linear hypotheses in $\mathbb{R}^N$ is $N + 1$. The result follows from the application of corollary to Sauer's lemma to generalization bound for this hypothesis set. □

Remark 10. When the dimension of the feature space N is large compared to the sample size m, this bound is uninformative.