

Lecture-12: Complexity bounds for separating hyperplanes

1 Complexity bounds for separating hyperplanes

We will find tighter upper bounds on the Rademacher complexity and VC-dimension on the hypothesis class of separating hyperplanes and its analog in higher dimensions using kernel methods.

1.1 Separating hyperplanes based hypotheses

Theorem 1.1 (VC-dimension for hypothesis set of canonical separating hyperplanes). Consider an unlabeled sample $x \in \mathcal{X}^m$ such that $\sup_{i \in [m]} \|x_i\| \leq r$ and the hypothesis set of canonical hyperplanes $H \triangleq \{x \mapsto \text{sign}(\langle w, x \rangle) : \min_{x \in \mathcal{X}^m} |\langle w, x \rangle| = 1, \|w\| \leq \Lambda\}$. Then, $\text{VC-dim}(H) \leq r^2 \Lambda^2$.

Proof. Let $\text{VC-dim}(H) = d$, and unlabeled sample $x \in \mathcal{X}^d$ that can be fully shattered, i.e. $|H_x| = 2^d$. Then, for any label sequence $y \in \{-1, 1\}^d$, there exists $h^w \in H$ such that $h_x^w = y$. That is, there exists $w \in \mathbb{R}^N$ such that $y_i(\langle w, x_i \rangle) \geq 1$ for all $i \in [d]$. Summing up these inequalities for each $i \in [d]$, from the linearity of inner product, Cauchy-Schwartz inequality, and hypothesis $\|w\| \leq \Lambda$, we get

$$d \leq \left\langle w, \sum_{i=1}^d y_i x_i \right\rangle \leq \|w\| \left\| \sum_{i=1}^d y_i x_i \right\| \leq \Lambda \left\| \sum_{i=1}^d y_i x_i \right\|.$$

Since this inequality holds for any label sequence $y \in \{-1, 1\}^d$, it also holds on expectation over $y \in \{-1, 1\}^d$ drawn *i.i.d.* according to a uniform distribution. From the independence assumption, we have $\mathbb{E}[y_i y_j] = \mathbb{E}[y_i] \mathbb{E}[y_j]$ for $i \neq j$. Thus, since the distribution is uniform, $\mathbb{E}[y_i y_j] = \mathbb{1}_{\{i=j\}}$. Taking expectation and applying Jensen's inequality to convex square function, we get

$$d^2 \leq \Lambda^2 \left(\mathbb{E}_y \left\| \sum_{i=1}^d y_i x_i \right\|^2 \right) \leq \Lambda^2 \mathbb{E}_y \left\| \sum_{i=1}^d y_i x_i \right\|^2 = \Lambda^2 \sum_{i=1}^d \|x_i\|^2 \leq d r^2 \Lambda^2.$$

□

Remark 1. When the training data is linearly separable, the maximum-margin canonical hyperplane with $\|w\| = 1/\rho$ can be plugged into above theorem. In this case, $\Lambda = 1/\rho$, and the upper bound can be rewritten as r^2/ρ^2 . Note that the choice of Λ must be made before receiving the sample $x \in \mathcal{X}^d$.

Theorem 1.2 (Rademacher complexity for separating hyperplanes). Consider unlabeled sample $x \in \mathcal{X}^m$ such that $\sup_{i \in [m]} \|x_i\| \leq r$ and the hypothesis set of hyperplanes $H \triangleq \{x \mapsto \langle w, x \rangle : \|w\| \leq \Lambda\}$. Then, $\hat{\mathcal{R}}_x(H) \leq \frac{1}{\sqrt{m}} r \Lambda$.

Proof. From the definition of empirical Rademacher complexity, the linearity of inner products, the application of Cauchy-Schwarz inequality to inner products, the application of Jensen's inequality to convex square function, and *i.i.d.* uniform nature of Rademacher vector σ , we get

$$\hat{\mathcal{R}}_x(H) = \frac{1}{m} \mathbb{E}_\sigma \left[\sup_w \left\langle w, \sum_{i=1}^m \sigma_i x_i \right\rangle \right] \leq \frac{\Lambda}{m} \mathbb{E}_\sigma \left\| \sum_{i=1}^m \sigma_i x_i \right\| \leq \frac{\Lambda}{m} \left(\mathbb{E}_\sigma \left\| \sum_{i=1}^m \sigma_i x_i \right\|^2 \right)^{\frac{1}{2}} = \frac{\Lambda}{m} \sqrt{\sum_{i=1}^m \|x_i\|^2} \leq \frac{1}{\sqrt{m}} r \Lambda.$$

□

Lemma 1.3. Consider a binary label set $\mathcal{Y} \triangleq \{-1, 1\}$, input space $\mathcal{X} \subseteq \mathbb{R}^d$, unlabeled sample $x \in \mathcal{X}^m$, hypothesis set $H \subseteq \mathbb{R}^{\mathcal{X}}$, and $b \in \mathbb{R}$. Define $b + H \triangleq \{b + h : h \in H\}$, then $\hat{\mathcal{R}}_x(H) = \hat{\mathcal{R}}_x(\bar{H})$.

Proof. From the definition of empirical Rademacher complexity and the fact that Rademacher vector σ is *i.i.d.* zero mean, we get

$$\hat{\mathcal{R}}_x(H) = \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{h \in H} \sum_{i=1}^m \langle \sigma_i, h(x_i) + b \rangle \right] = \frac{1}{m} \mathbb{E}_\sigma \left[b \sum_{i=1}^m \sigma_i + \sup_{h \in H} \sum_{i=1}^m \langle \sigma_i, h(x_i) \rangle \right] = \hat{\mathcal{R}}_x(H).$$

□

1.2 Kernel based hypotheses

Consider an input space $\mathcal{X} \subseteq \mathbb{R}^d$, binary label set $\mathcal{Y} \triangleq \{-1, 1\}$, a PDS kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, associated RKHS \mathbb{H} , and a hypothesis set of the form $H \triangleq \{h \in \mathbb{H} : \|h\|_{\mathbb{H}} \leq \Lambda\}$ for some $\Lambda \geq 0$. Recall that the feature map $\Phi : \mathcal{X} \rightarrow \mathbb{H}$ for the associated RKHS is defined as $\Phi(x) \triangleq e_x = k(x, \cdot)$ for all $x \in \mathcal{X}$. Further any $w \in \mathbb{H}$ has the form $x \mapsto \langle w, e_x \rangle = \langle w, \Phi(x) \rangle$ due to the reproducing property. For any $w \in H$, we have $\|w\|_{\mathbb{H}} \leq \Lambda$.

Theorem 1.4 (Rademacher complexity of kernel-based hypotheses). *Consider a PDS kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, associated RKHS \mathbb{H} and feature mapping $\Phi : \mathcal{X} \rightarrow \mathbb{H}$, an unlabeled sample $x \in \mathcal{X}^m$ such that $\sup_{i \in [m]} k(x_i, x_i) \leq r^2$, and a hypothesis set $H \triangleq \{x \mapsto \langle w, \Phi(x) \rangle : \|w\|_{\mathbb{H}} \leq \Lambda\} \subseteq \mathbb{H}$ for some $\Lambda \geq 0$. Denoting the kernel matrix $K \in \mathbb{R}^{m \times m}$ associated with kernel k and unlabeled sample x , defined as $K_{ij} \triangleq k(x_i, x_j)$, we observe that*

$$\hat{\mathcal{R}}_x(H) \leq \frac{\Lambda}{m} \sqrt{\text{tr } K} \leq \frac{r\Lambda}{\sqrt{m}}.$$

Proof. From the definition of empirical Rademacher complexity, the linearity of inner products, the application of Cauchy-Schwarz inequality to inner products, the application of Jensen's inequality to convex square function, and *i.i.d.* uniform nature of Rademacher vector σ , we get

$$\hat{\mathcal{R}}_x(H) = \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{w \in H} \left\langle w, \sum_{i=1}^m \sigma_i \Phi(x_i) \right\rangle_{\mathbb{H}} \right] \leq \frac{\Lambda}{m} \mathbb{E}_\sigma \left\| \sum_{i=1}^m \sigma_i \Phi(x_i) \right\|_{\mathbb{H}} \leq \frac{\Lambda}{m} \left(\mathbb{E}_\sigma \left\| \sum_{i=1}^m \sigma_i \Phi(x_i) \right\|_{\mathbb{H}}^2 \right)^{\frac{1}{2}} = \frac{\Lambda}{m} \sqrt{\sum_{i=1}^m \|\Phi(x_i)\|_{\mathbb{H}}^2}.$$

The result follows since $\|\Phi(x_i)\|_{\mathbb{H}}^2 = k(x_i, x_i)$ and $\sum_{i=1}^m k(x_i, x_i) = \text{tr } K$. □

Remark 2. Trace of the kernel matrix is an important quantity for controlling the complexity of hypothesis sets based on kernels.

A Talagrand's inequality

Lemma A.1 (Talagrand). *Let $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ be an ℓ -Lipschitz function. Then, for any hypothesis set H of real-valued functions, we have*

$$\hat{\mathcal{R}}_x(\Phi \circ H) \leq \ell \hat{\mathcal{R}}_x(H).$$

Proof. The empirical Rademacher complexity for an unlabeled sample $x \in \mathcal{X}^m$, is

$$\hat{\mathcal{R}}_x(\Phi \circ H) = \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i (\Phi \circ h)(x_i) \right] = \frac{1}{m} \mathbb{E} \left[\mathbb{E} \left[\sup_{h \in H} u_{m-1}(h) + \sigma_m (\Phi \circ h)(x_m) \mid \sigma^{m-1} \right] \right],$$

where $u_{m-1}(h) \triangleq \sum_{i=1}^{m-1} \sigma_i (\Phi \circ h)(x_i)$ for any hypothesis $h \in H$. Fix $\epsilon > 0$. By the definition of supremum, there exist $h_1, h_2 \in H$ such that

$$\begin{aligned} u_{m-1}(h_1) + (\Phi \circ h_1)(x_m) &\geq (1 - \epsilon) \sup_{h \in H} [u_{m-1}(h) + (\Phi \circ h)(x_m)], \\ u_{m-1}(h_2) - (\Phi \circ h_2)(x_m) &\geq (1 - \epsilon) \sup_{h \in H} [u_{m-1}(h) - (\Phi \circ h)(x_m)]. \end{aligned}$$

Thus for any $\epsilon > 0$, by the definition of \mathbb{E}_{σ_m} , we have

$$(1 - \epsilon) \mathbb{E} \left[\sup_{h \in H} u_{m-1}(h) + \sigma_m (\Phi \circ h)(x_m) \right] \leq \frac{1}{2} [u_{m-1}(h_1) + (\Phi \circ h_1)(x_m)] + \frac{1}{2} [u_{m-1}(h_2) - (\Phi \circ h_2)(x_m)].$$

From the ℓ -Lipschitz property of Φ , we get $(\Phi \circ h_1)(x_m) - (\Phi \circ h_2)(x_m) \leq \ell |h_1(x_m) - h_2(x_m)|$. Defining $s \triangleq \text{sign}(h_1(x_m) - h_2(x_m))$, we get

$$\begin{aligned} (1 - \epsilon) \mathbb{E} \left[\sup_{h \in H} u_{m-1}(h) + \sigma_m(\Phi \circ h)(x_m) \right] &\leq \frac{1}{2} [u_{m-1}(h_1) + s\ell h_1(x_m)] + \frac{1}{2} [u_{m-1}(h_2) - s\ell h_2(x_m)] \\ &\leq \frac{1}{2} \sup_{h \in H} [u_{m-1}(h) + s\ell h(x_m)] + \frac{1}{2} \sup_{h \in H} [u_{m-1}(h) - s\ell h(x_m)] = \mathbb{E}_{\sigma_m} \left[\sup_{h \in H} [u_{m-1}(h) + \sigma_m \ell h(x_m)] \right]. \end{aligned}$$

Since the inequality holds for all $\epsilon > 0$, we have

$$\mathbb{E}_{\sigma_m} \left[\sup_{h \in H} u_{m-1}(h) + \sigma_m(\Phi \circ h)(x_m) \right] \leq \mathbb{E}_{\sigma_m} \left[\sup_{h \in H} u_{m-1}(h) + \sigma_m \ell h(x_m) \right].$$

Proceeding in the same way for all other $\sigma_i, i \in [m-1]$ proves the lemma. \square