Lecture-10: Margin Theory

1 Margin theory

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

Definition 1.1 (Margin). The geometric margin $\rho(x,y)$ of a point x with label y with respect to a linear classifier $h : x \mapsto \langle w,x\rangle + b$ is its distance to the hyperplane $\langle w,x\rangle + b = 0$. That is,

$$\rho(x,y) = \frac{y(\langle w,x\rangle + b)}{\|w\|}.$$

The margin of a linear classifier h for a sample $x \in \mathcal{X}^m$ is the minimum margin over the points in the sample, i.e.

$$\rho \triangleq \min\left\{\rho(x_i,y_i) = \frac{y_i(\langle w,x_i\rangle + b)}{\|w\|} : i \in [m]\right\}.$$

Corollary 1.2. *For any $\delta > 0$, with probability at least $1-\delta$, for any $h \in H = \{x \mapsto \langle w,x\rangle + b : w \in \mathbb{R}^n, b \in \mathbb{R}\}$,*

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

Proof. Recall that the VC-dimension of the family of hyperplanes or linear hypotheses in $\mathbb{R}^n$ is $n+1$. Application of corollary to Sauner's lemma to generalization bound for this hypothesis set, gives us the result. □

Remark 1. When the dimension of the feature space n is large compared to the sample size m, this bound is uninformative.

Theorem 1.3. *Let $x \in \mathcal{X}^m$ be a sample such that $\sup_{i\in[m]} \|x_i\| \leqslant r$. Then, the VC-dimension d of the set of canonical hyperplanes*

$$H \triangleq \left\{x \mapsto \operatorname{sign}(\langle w,x\rangle) : \min_{i\in[m]} |\langle w,x_i\rangle| = 1, \|w\| \leqslant \Lambda\right\}$$

is upper bounded as $\operatorname{VCdim}(H) = d \leqslant r^2\Lambda^2$.

Proof. Let $x \in \mathcal{X}^d$ be a sample set that can be fully shattered. Then, for all $y \in \{-1,1\}^d$, there exists w such that $y_i(\langle w,x_i\rangle) \geqslant 1$ for all $i \in [d]$. Summing up these inequalities, from the linearity of inner product and $\|w\| \leqslant \Lambda$, we get

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

Since this inequality holds for all $y \in \{-1,1\}^d$, 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

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

□

Remark 2. 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 1.4. *Let $x \in \mathcal{X}^m$ be a sample such that $\sup_{i\in[m]} \|x_i\| \leqslant r$ and $H = \{x \mapsto \langle w,x\rangle : \|w\| \leqslant \Lambda\}$. Then, the empirical Rademacher complexity of H can be bounded as*

$$\hat{\mathfrak{R}}_H(m) = \sqrt{\frac{r^2\Lambda^2}{m}}.$$

Proof. From the definition of 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

$$\hat{\mathfrak{R}}_H(m) = \frac{1}{m}\mathbb{E}\left[sup_w \left\langle w, \sum_{i=1}^m \sigma_i x_i \right\rangle\right] \leqslant \frac{\Lambda}{m}\mathbb{E}\left\|\sum_{i=1}^m \sigma_i x_i\right\| \leqslant \frac{\Lambda}{m}\left(\mathbb{E}\left\|\sum_{i=1}^m \sigma_i x_i\right\|^2\right)^{\frac{1}{2}} = \frac{\Lambda}{m}\sqrt{\sum_{i=1}^m \|x_i\|^2} \leqslant \sqrt{\frac{r^2\Lambda^2}{m}}.$$

□

To present the main margin-based generalization bounds for non-separable training data, we introduce a margin loss function, for the target margin $\rho > 0$.

Definition 1.5 (Margin loss function). For any $\rho > 0$, the ρ-margin loss is the function $L_\rho : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$ defined for all $y, y' \in \mathbb{R}$ by $L_\rho(y,y') = \Phi_\rho(yy')$ with,

$$\Phi_\rho(x) = \begin{cases} 0, & \rho \leqslant x, \\ 1 - x/\rho, & 0 \leqslant x \leqslant \rho, \\ 1, & x \leqslant 0. \end{cases}$$

Definition 1.6 (Empirical margin loss). Given a sample $x \in \mathcal{X}^m$ and a hypothesis $h : \mathcal{X} \to \{-1,1\}$, the empirical margin loss is defined by

$$\hat{R}_\rho(h) = \frac{1}{m}\sum_{i=1}^m \Phi_\rho(y_i h(x_i)).$$

Remark 3. (i) For any $i \in [m]$, we can upper bound the margin loss function $\Phi_\rho(y_i h(x_i)) \leqslant \mathbb{1}_{\{y_i h(x_i) \leqslant \rho\}}$. Thus, the empirical margin loss can be upper-bounded as

$$\hat{R}_\rho(h) \leqslant \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{\{y_i h(x_i)\leqslant\rho\}}.$$

(ii) The upper bound on the empirical margin loss is the fraction of the points in the training sample S that have been misclassified or classified with confidence less than ρ.

(iii) When h is a linear function defined by a weight vector w with $\|w\| = 1$, $y_i h(x_i)$ is the margin of point x_i. Thus, the upper bound is then the fraction of the points in the training data with margin less than ρ.

(iv) The slope of the function Φ_ρ defining the margin loss is at most $1/\rho$, thus Φ_ρ is $1/\rho$-Lipschitz.

Lemma 1.7 (Talagrand). *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be an ℓ-Lipschitz function. Then, for any hypothesis set H of real-valued functions, we have*

$$\hat{\mathfrak{R}}_z(\Phi \circ H) \leqslant \ell\hat{\mathfrak{R}}_z(H).$$

Proof. For a sample $S = (x_1,\ldots,x_m)$, by the definition of empirical Rademacher complexity, we have

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

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

$$u_{m-1}(h_1) + (\Phi\circ h_1)(x_m) \geqslant (1-\varepsilon)\sup_{h\in H}[u_{m-1}(h) + (\Phi\circ h)(x_m)],$$
$$u_{m-1}(h_2) - (\Phi\circ h_2)(x_m) \geqslant (1-\varepsilon)\sup_{h\in H}[u_{m-1}(h) - (\Phi\circ h)(x_m)].$$

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

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

Let $s = \operatorname{sign}(h_1(x_m) - h_2(x_m))$. Then, from the ℓ-Lipschitz property of Φ, we get

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

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

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

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

Theorem 1.8 (Margin bound for binary classification). *Let H be a set of real-valued functions. Fix $\rho > 0$, then, for any $\delta > 0$, with probability at least $1-\delta$, each of the following holds for all $h \in H$*

$$R(h) \leqslant \hat{R}_\rho(h) + \frac{2}{\rho}\mathcal{R}_m(H) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}, \qquad R(h) \leqslant \hat{R}_\rho(h) + \frac{2}{\rho}\mathcal{R}_z(H) + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}.$$

Proof. Let $\tilde{H} = \{z \mapsto yh(x) : h \in H\}$ for $z \in \mathcal{X}\times\mathcal{Y}$, and the family of loss functions $\tilde{\mathcal{H}} = \left\{\Phi_\rho \circ f : f \in \tilde{H}\right\}$. From the generalization bound on binary classification using Rademacher complexity, we get that with probability at least $1-\delta$, for all $g \in \tilde{\mathcal{H}}$

$$\mathbb{E}g(z) \leqslant \frac{1}{m}\langle \mathbf{1}, g_T\rangle + 2\mathcal{R}_m(\tilde{\mathcal{H}}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

From the definition of g and $\tilde{\mathcal{H}}$, we get

$$\mathbb{E}\Phi_\rho(yh(x)) \leqslant \hat{R}_\rho(h) + 2\mathcal{R}_m(\Phi_\rho\circ\tilde{H}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

Since $\mathbb{1}_{\{u\leqslant 0\}} \leqslant \Phi_\rho(u)$ for all $u \in \mathbb{R}$, we have $R(h) = \mathbb{E}\mathbb{1}_{\{yh(x)\leqslant 0\}} \leqslant \mathbb{E}\Phi_\rho(yh(x))$. Further, $\mathcal{R}_m$ is invariant to a constant shift. Therefore,

$$R(h) \leqslant \hat{R}_\rho(h) + 2\mathcal{R}_m((\Phi_\rho - 1)\circ\tilde{H}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

From the previous Lemma, Φ_ρ is $1/\rho$-Lipschitz, and hence so is $(\Phi_\rho - 1)$, further $(\Phi_\rho - 1)(0) = 0$. This implies that $\mathcal{R}_m(\Phi_\rho\circ\tilde{H}) \leqslant \frac{1}{\rho}\mathcal{R}_m(\tilde{H})$, and

$$\mathcal{R}_m(\tilde{H}) = \frac{1}{m}\mathbb{E}\left[\sup_{h\in H}\sum_{i=1}^m \sigma_i y_i h(x_i)\right] = \frac{1}{m}\mathbb{E}\left[\sup_{h\in H}\sum_{i=1}^m \sigma_i h(x_i)\right] = \mathcal{R}_m(H).$$

□

Remark 4. Target margin ρ is the trade-off parameter in the generalization bound above. Empirical margin loss $\hat{R}_\rho$ increases as a function of target margin ρ, and the complexity term decreases with the ρ.

Theorem 1.9. *Let H be a set of real-valued functions. Then, for any $\delta > 0$, with probability at least $1-\delta$, each of the following holds for all hypothesis $h \in H$ and target margin $\rho \in (0,1)$*

$$R(h) \leqslant \hat{R}_\rho(h) + \frac{4}{\rho}\mathcal{R}_m(H) + \sqrt{\frac{\ln\log_2\frac{2}{\delta}}{m}} + \sqrt{\frac{\ln\frac{2}{\delta}}{2m}}, \qquad R(h) \leqslant \hat{R}_\rho(h) + \frac{4}{\rho}\mathcal{R}_z(H) + \sqrt{\frac{\ln\log_2\frac{2}{\delta}}{m}} + 3\sqrt{\frac{\ln\frac{4}{\delta}}{2m}}.$$

Proof. Consider sequences $\rho = (\rho_k > 0 : k \in \mathbb{N})$ and $\varepsilon = (\varepsilon_k \in (0,1) : k \in \mathbb{N})$. By the previous theorem, we have

$$P\left\{R(h) - \hat{R}_{\rho_k}(h) > \frac{2}{\rho_k}\mathcal{R}_m(H) + \varepsilon_k\right\} \leqslant exp(-2m\varepsilon_k^2).$$

□

Choosing $\varepsilon_k = \varepsilon + \sqrt{\frac{\ln k}{m}}$ for each $k \in \mathbb{N}$, and using the union bound, we get

$$P\left(\cup_{k\in\mathbb{N}}\left\{R(h) - \hat{R}_{\rho_k}(h) > \frac{2}{\rho_k}\mathcal{R}_m(H) + \varepsilon_k\right\}\right) \leqslant \sum_{k\in\mathbb{N}} \exp(-2m(\varepsilon + \sqrt{\ln k/m})^2) \leqslant e^{-2m\varepsilon^2}\frac{\pi^2}{6} \leqslant 2e^{-2m\varepsilon^2}.$$

We can choose $\rho_k = 2^{-k}$ for all $k \in \mathbb{N}$. Then, for any $\rho \in (0,1)$, there exists $k \in \mathbb{N}$ such that $\rho \in (\rho_k, \rho_{k-1}]$ with $\rho_0 = 1$. For that k, we have $\rho \leqslant \rho_{k-1} = 2\rho_k$, and thus $1/\rho_k \leqslant 2/\rho$ and

$$\ln k = \sqrt{\ln \log_2(1/\rho_k)} \leqslant \sqrt{\ln \log_2(2/\rho)}.$$

Furthermore, for any $h \in H$, we have $\hat{R}_{\rho_k}(h) \leqslant \hat{R}_\rho(h)$ from the monotonicity of empirical marginal loss in the target margin ρ. Thus,

$$P\left(\cup_{k\in\mathbb{N}}\left\{R(h) - \hat{R}_\rho(h) > 4\mathcal{R}_m(H) + \sqrt{\frac{\ln \log_2(2/\rho)}{m}} + \varepsilon\right\}\right) \leqslant 2\exp(-2m\varepsilon^2).$$

Corollary 1.10. *Let $H = \{x \mapsto \langle w,x\rangle : \|w\| \leqslant \Lambda\}$ and assume that $X \subseteq x : \|x\| \leqslant r$. Fix the target margin $\rho > 0$, then, for any $\delta > 0$, with probability at least $1-\delta$, for any $h \in H$,*

$$R(h) \leqslant \hat{R}_\rho(h) + 2\sqrt{\frac{\frac{r^2}{\rho^2}\Lambda^2}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

Remark 5. This bound can be generalized to hold uniformly for all $\rho > 0$ at the cost of an additional $\sqrt{\frac{\ln \log_2 \frac{2}{\rho}}{m}}$.

Remark 6. This generalization bound for linear hypotheses does not depend directly on the dimension of the feature space, but only on the margin. It suggests that a small generalization error can be achieved when ρ/r is large (small second term) while the empirical margin loss is relatively small (first term). The latter occurs when few points are either classified incorrectly, or correctly with margin less than ρ.

Remark 7. Lack of dependence of the guarantee on the dimension of the feature space appears to contradict the VC-dimension lower bounds, which show that for any learning algorithm $\mathcal{A}$ there exists a *bad* distribution for which the error of the hypothesis returned by the algorithm is $\Omega(d/m)$ with a non-zero probability. However, the bound of the corollary does not rule out such bad cases, since for such bad distributions, the empirical margin loss would be large even for a relatively small margin ρ, and thus the bound of the corollary would be loose in that case.

Remark 8. Thus, in some sense, the learning guarantee of the corollary hinges upon the hope of a good margin value ρ. If there exists a relatively large margin value $\rho > 0$ for which the empirical margin loss is small, then a small generalization error is guaranteed by the corollary. This favorable margin situation depends on the distribution. While the learning bound is distribution-independent, the existence of a good margin is in fact distribution-dependent. A favorable margin seems to appear relatively often in applications.

Remark 9. The bound of the corollary gives a strong justification for margin-maximization algorithms such as SVMs. For $\rho = 1$, the margin loss can be upper bounded by the hinge loss. For all $x \in R$, we have $\Phi_1(x) \leqslant \max\{1-x, 0\}$. Using this fact, the bound of the corollary implies that with probability at least $1-\delta$, for all $h \in H = \{x \mapsto \langle w,x\rangle : \|w\| \leqslant \Lambda\}$, we have

$$R(h) \leqslant \sum_{i=1}^m \xi_i + 2\sqrt{\frac{r^2\Lambda^2}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}},$$

where $\xi_i = \max\{1 - y_i\langle w, x_i\rangle, 0\}$. The objective function minimized by the SVM algorithm has precisely the form of this upper bound: the first term corresponds to the slack penalty over the training set and the second to the minimization of the $\|w\|$ which is equivalent to that of $\|w\|^2$. Note that an alternative objective function would be based on the empirical margin loss instead of the hinge loss. However, the advantage of the hinge loss is that it is convex, while the margin loss is not.

Remark 10. These generalization bounds do not directly depend on the dimension of the feature space and guarantee good generalization with a favorable margin. Thus, we can seek large-margin separating hyperplanes in a very high-dimensional space. However, finding solution to SVM in higher dimensions require computing many inner products in that space, which could be very costly. However, if the inner products are represented by PDS Kernels, SVMs in higher dimensions work well.

2 Learning Guarantees: Kernel Methods

Consider a hypothesis set of the form $H = \{h \in \mathbb{H} : \|h\|_{\mathbb{H}} \leqslant \Lambda\}$, for some $\Lambda \geqslant 0$, where $\mathbb{H}$ is the RKHS associated to a kernel K. By the reproducing property, any $h \in H$ is of the form $x?? \mapsto \langle h, K(x,\cdot)\rangle = \langle h, \Phi(x)\rangle$ with $\|h\|_{\mathbb{H}} \leqslant \Lambda$, where Φ is a feature mapping associated to K, that is of the form $x \mapsto \langle w, \Phi(x)\rangle$ with $\|w\|_{\mathbb{H}} \leqslant \Lambda$.

Theorem 2.1 (Rademcaher complexity of kernel-based hypotheses). *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel and let $\Phi : \mathcal{X} \to \mathbb{H}$ be a feature mapping associated to K. Let $S \subseteq \{x : K(x,x) \leqslant r^2\}$ be a sample of size m, and let $H = \{x \mapsto \langle w, \Phi(x)\rangle : \|w\|_{\mathbb{H}} \leqslant \Lambda\}$ for some $\Lambda \geqslant 0$. Then ????*

$$\hat{R}_z(H) \leqslant \frac{\Lambda\sqrt{\mathrm{tr}[\mathbf{K}]}}{m} \leqslant \sqrt{\frac{r^2\Lambda^2}{m}}.$$

Proof. The proof steps are identical to the Theorem 1.4, where x_i is replaced by $\Phi(x_i)$ and hence $\|x_i\|^2 = K(x_i,x_i)$. It follows that $\sum_{i=1}^m K(x_i,x_i) = \mathrm{tr}[\mathbf{K}]$. □

Remark 11. Trace of the kernel matrix is an important quantity for controlling the complexity of hypothesis sets based on kernels.

Corollary 2.2 (Margin bounds for kernel-based hypotheses). *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel with $r = \sup_{x\in\mathcal{X}} K(x,x)$. Let $\Phi : \mathcal{X} \to \mathbb{H}$ be a feature mapping associated to K, and let $H = \{x \mapsto \langle w, \Phi(x)\rangle : \|w\|_{\mathbb{H}} \leqslant \Lambda\}$ for some $\Lambda \geqslant 0$. Fix $\rho > 0$, then for any $\delta > 0$, the following hold with probability at least $1-\delta$ for any hypothesis $h \in H$*

$$R(h) \leqslant \hat{R}_\rho(h) + 2\sqrt{\frac{\frac{r^2\Lambda^2}{\rho^2}}{m}} + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}, \qquad R(h) \leqslant \hat{R}_\rho(h) + 2\sqrt{\frac{\frac{\mathrm{tr}[\mathbf{K}]\Lambda^2}{\rho^2}}{m}} + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}.$$