Lecture-13: Margin based generalization bounds

1 Margin based generalization bounds

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 1.1 (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 $y, y' \in \mathbb{R}$ as $L_\rho(y, y') \triangleq \Phi_\rho(yy')$ where

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

Remark 1. The following statements hold for margin loss function.

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

(ii) Margin loss function is monotonic in ρ, i.e. for any $\rho < \rho'$, we have $\Phi_\rho(x) \leqslant \Phi_{\rho'}(x)$ for all $x \in \mathbb{R}$.

Definition 1.2 (Empirical margin loss). Consider binary label set $\mathcal{Y} \triangleq \{-1, 1\}$ and hypothesis set $H \subseteq \mathcal{Y}^{\mathcal{X}}$. Given a sample $x \in \mathcal{X}^m$ and a hypothesis $h \in H$, the *empirical margin loss* is defined as

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

Remark 2. The following statements hold for empirical margin loss.

(i) For any $i \in [m]$, we can bound the margin loss function $\mathbb{1}_{\{y_i h(x_i) \leqslant 0\}} \leqslant \Phi_\rho(y_i h(x_i)) \leqslant \mathbb{1}_{\{y_i h(x_i) \leqslant \rho\}}$. Thus, the empirical margin loss can be bounded as

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{y_i h(x_i) \leqslant 0\}} \leqslant \hat{R}_\rho(h) \leqslant \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{y_i h(x_i) \leqslant \rho\}}.$$

(ii) 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.

Definition 1.3. For each $h \in H \subseteq \mathbb{R}^{\mathcal{X}}$, we can define another map $\bar{h} \in \mathbb{R}^{\mathcal{Z}}$ defined for each $z \in \mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ as $\bar{h}(z) \triangleq yh(x)$. We define the set of function $\bar{H} \triangleq \{z \mapsto yh(x) : h \in H\} \subseteq \mathbb{R}^{\mathcal{Z}}$, and the family of loss functions $\tilde{H} \triangleq \{\Phi_\rho \circ \bar{h} : \bar{h} \in \bar{H}\} \subseteq [0,1]^{\mathcal{Z}}$ where $\Phi_\rho \in [0,1]^{\mathbb{R}}$ is the ρ-margin loss function.

Remark 3. From the definition of loss functions set $\tilde{H}$ and empirical margin loss $\hat{R}_\rho$, we get for any $\tilde{h} \in \tilde{H}$

$$\tilde{h}(z) = \Phi_\rho(yh(x)), \qquad \frac{1}{m} \langle \mathbf{1}, \tilde{h}_z \rangle = \hat{R}_\rho(h), \qquad \mathcal{R}_m(\tilde{H}) = \mathcal{R}_m(\Phi_\rho \circ \bar{H}).$$

1.1 Linear binary classification

Theorem 1.4 (Margin bound for binary classification). *Consider hypothesis set $H \subseteq \mathbb{R}^{\mathcal{X}}$ and $\rho, \delta > 0$, then*

$$P\Big(\bigcap_{h \in H} \Big\{ R(h) \leqslant \hat{R}_\rho(h) + \frac{2}{\rho} \mathcal{R}_m(H) + \sqrt{\frac{1}{2m} \ln \frac{1}{\delta}} \Big\} \Big) \geqslant 1 - \delta,$$

$$P\Big(\bigcap_{h \in H} \Big\{ R(h) \leqslant \hat{R}_\rho(h) + \frac{2}{\rho} \hat{\mathcal{R}}_z(H) + 3\sqrt{\frac{1}{2m} \ln \frac{2}{\delta}} \Big\} \Big) \geqslant 1 - \delta.$$

Proof. From the generalization bound on binary classification using Rademacher complexity, we get that

$$P\Big(\cap_{\tilde{h}\in\tilde{H}}\left\{\mathbb{E}\tilde{h}(z)\leqslant\frac{1}{m}\langle\mathbf{1},\tilde{h}_z\rangle+2\mathcal{R}_m(\tilde{H})+\sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}\right\}\Big)\geqslant 1-\delta.$$

Since $\mathbb{1}_{\{u\leqslant 0\}}\leqslant\Phi_\rho(u)$ for all $u\in\mathbb{R}$, we have $R(h)=\mathbb{E}\mathbb{1}_{\left\{\tilde{h}(z)\leqslant 0\right\}}=\mathbb{E}\mathbb{1}_{\{yh(x)\leqslant 0\}}\leqslant\mathbb{E}\Phi_\rho(yh(x))=\mathbb{E}\tilde{h}(z)$. Further, $\frac{1}{m}\langle\mathbf{1},\tilde{h}_z\rangle=\hat{R}_\rho(h)$ and $\tilde{H}=\Phi_\rho\circ\bar{H}$. It follows that

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

Since $y=c(x)$, there is a one-to-one relationship between $\bar{H}$ and H, and therefore

$$P\Big(\cap_{h\in H}\left\{R(h)\leqslant\hat{R}_\rho(h)+2\mathcal{R}_m(\Phi_\rho\circ\bar{H})+\sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}\right\}\Big)\geqslant 1-\delta.$$

Since Φ_ρ is $1/\rho$-Lipschitz, Talagrand's inequality Lemma implies that $\mathcal{R}_m(\Phi_\rho\circ\bar{H})\leqslant\frac{1}{\rho}\mathcal{R}_m(\bar{H})$, and

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

□

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. If for a relatively large value of margin ρ the empirical margin loss of h remains relatively small, then h benefits from a very favorable guarantee on its generalization error.

Remark 5. For Theorem 1.4, the margin parameter ρ must be selected beforehand. We next show that the bounds of the theorem can be generalized to hold uniformly for all margins $\rho\in(0,r]$ at the cost of a modest additional term $\sqrt{\frac{1}{m}\ln\log_2\frac{2r}{\rho}}$.

Theorem 1.5. *Consider hypothesis set $H\subseteq\mathbb{R}^{\mathcal{X}}$, $\delta>0$, and margin $\rho\in(0,r]$. Then,*

$$P\Big(\bigcap_{\rho<r,h\in H}\left\{R(h)\leqslant\hat{R}_\rho(h)+\frac{4}{\rho}\mathcal{R}_m(H)+\sqrt{\frac{1}{m}\ln\log_2\frac{2r}{\rho}}+\sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}\right\}\Big)\geqslant 1-\delta,$$

$$P\Big(\bigcap_{\rho<r,h\in H}\left\{R(h)\leqslant\hat{R}_\rho(h)+\frac{4}{\rho}\hat{\mathcal{R}}_z(H)+\sqrt{\frac{1}{m}\ln\log_2\frac{2r}{\rho}}+3\sqrt{\frac{1}{2m}\ln\frac{4}{\delta}}\right\}\Big)\geqslant 1-\delta.$$

Proof. Consider sequences $\rho\in\mathbb{R}_+^{\mathbb{N}}$ and $\epsilon\in(0,1)^{\mathbb{N}}$. By the previous theorem, we have

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

Choosing $\epsilon_k=\epsilon+\sqrt{\frac{1}{m}\ln k}$ for each $k\in\mathbb{N}$, using the union bound, the fact that $(a+b)^2\geqslant a^2+b^2$ for $a,b>0$, and the upper bound on sum $\sum_{k\in\mathbb{N}}\frac{1}{k^2}\leqslant\frac{\pi^2}{6}\leqslant 2$, we get

$$P\left(\cup_{k\in\mathbb{N}}\cup_{h\in H}\left\{R(h)-\hat{R}_{\rho_k}(h)>\frac{2}{\rho_k}\mathcal{R}_m(H)+\epsilon_k\right\}\right)\leqslant\sum_{k\in\mathbb{N}}\exp(-2m(\epsilon+\sqrt{\frac{1}{m}\ln k})^2)\leqslant e^{-2m\epsilon^2}\frac{\pi^2}{6}\leqslant 2e^{-2m\epsilon^2}.$$

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

$$\sqrt{\ln k}=\sqrt{\ln\log_2(\rho_0/\rho_k)}\leqslant\sqrt{\ln\log_2(2\rho_0/\rho)}.$$

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. This implies that

$$\left\{R(h) - \hat{R}_\rho(h) > \frac{4}{\rho}\mathcal{R}_m(H) + \sqrt{\frac{1}{m}\ln\log_2\frac{2r}{\rho}} + \epsilon\right\} \subseteq \left\{R(h) - \hat{R}_{\rho_k}(h) > \frac{4}{\rho}\mathcal{R}_m(H) + \sqrt{\frac{1}{m}\ln\log_2\frac{2r}{\rho}} + \epsilon\right\}$$
$$\subseteq \left\{R(h) - \hat{R}_{\rho_k}(h) > \frac{2}{\rho_k}\mathcal{R}_m(H) + \epsilon_k\right\}.$$

Taking union over all $h \in H$ and $\rho \in (0, r)$, we observe that

$$\cup_{\rho\in(0,r)} \cup_{h\in H} \left\{R(h) - \hat{R}_\rho(h) > \frac{4}{\rho}\mathcal{R}_m(H) + \sqrt{\frac{1}{m}\ln\log_2\frac{2r}{\rho}} + \epsilon\right\} \subseteq \cup_{k\in\mathbb{N}} \cup_{h\in H} \left\{R(h) - \hat{R}_{\rho_k}(h) > \frac{2}{\rho_k}\mathcal{R}_m(H) + \epsilon_k\right\}.$$

Taking probability on both sides, we obtain

$$P\left(\cup_{\rho\in(0,r)} \cup_{h\in H} \left\{R(h) - \hat{R}_\rho(h) > \frac{4}{\rho}\mathcal{R}_m(H) + \sqrt{\frac{1}{m}\ln\log_2\frac{2r}{\rho}} + \epsilon\right\}\right) \leqslant 2\exp(-2m\epsilon^2).$$

□

Corollary 1.6. *Consider hypothesis set $H = \{x \mapsto \langle w, x\rangle : \|w\| \leqslant \Lambda\}$ pf separating hyperplanes and an unlabeled sample $x \in \mathcal{X}^m$ such that $\sup_{i\in[m]} \|x_i\| \leqslant r$. Fix the target margin $\rho > 0$ and target accuracy $\delta > 0$, then*

$$P\Big(\cap_{h\in H}\left\{R(h) \leqslant \hat{R}_\rho(h) + \frac{2r\Lambda}{\rho\sqrt{m}} + \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}\right\}\Big) \geqslant 1 - \delta.$$

Remark 6. This bound can be generalized to hold uniformly for all $\rho \in (0, r)$ at the cost of an additional $\sqrt{\frac{1}{m}\ln\log_2\frac{2r}{\rho}}$ in the generalization upper bound.

Remark 7. 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 8. 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 9. 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 10. 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. i.e. we have $\Phi_1(x) \leqslant \max\{1 - x, 0\}$for all $x \in \mathbb{R}$. This implies that $\Phi_1(y_i h(x_i)) \leqslant \max\{1 - y_i h(x_i), 0\}$ where $y_i h(x_i) \geqslant 1 - \xi_i$ and slack variables $\xi_i \geqslant 0$ for all $i \in [m]$. That is, $\xi_i = \max\{1 - y_i h(x_i), 0\}$ and we have $\Phi_1(y_i h(x_i)) \leqslant \xi_i$ for all $i \in [m]$. Thus, the empirical 1-margin loss for any hypothesis $h \in H \triangleq \{x \mapsto \langle w, x\rangle : \|w\| \leqslant \Lambda\}$ is upper bounded by $\hat{R}_1(h) \leqslant \frac{1}{m}\sum_{i=1}^m \xi_i$. Using this fact, the bound of the corollary implies that

$$P\Big(\cap_{h\in H}\left\{R(h) \leqslant \frac{1}{m}\sum_{i=1}^m \xi_i + 2\frac{r\Lambda}{\sqrt{m}} + \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}\right\}\Big) \geqslant 1 - \delta.$$

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 11. 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.

1.2 Kernel based binary classification

Corollary 1.7 (Margin bounds for kernel-based hypotheses). *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel, denote its associated RKHS with $\mathbb{H}$, associated feature mapping with $\Phi : \mathcal{X} \to \mathbb{H}$, and the hypothesis set of separating hyperplanes with a bounded RKHS norm with $H \triangleq \{x \mapsto \langle w, \Phi(x)\rangle : \|w\|_{\mathbb{H}} \leqslant \Lambda\}$ for some $\Lambda \geqslant 0$. For an unlabeled sample $x \in \mathcal{X}^m$ with $r = \max_{i\in[m]} k(x_i, x_i)$, target margin $\rho > 0$ and target accuracy $\delta > 0$, we have*

$$P\Big(_{h\in H}\left\{R(h) \leqslant \hat{R}_\rho(h) + \frac{2r\Lambda}{\rho\sqrt{m}} + \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}\right\}\Big) \geqslant 1-\delta,$$

$$P\Big(\cap_{h\in H}\left\{R(h) \leqslant \hat{R}_\rho(h) + \frac{2\Lambda}{\rho}\sqrt{\frac{\operatorname{tr} K}{m}} + 3\sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}\right\}\Big) \geqslant 1-\delta.$$