Lecture-12: Multi-class classification: Generalization bounds

1 Generalization bounds: mono-label case

In the binary setting, classifiers are often defined based on the sign of a scoring function. In the multi-class setting, a hypothesis is defined based on a scoring function $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The label associated to point x is the one resulting in the largest score $h(x,y)$, which defines the following mapping from $\mathcal{X}$ to $\mathcal{Y}$

$$x \mapsto \arg\max \{h(x,y) : y \in \mathcal{Y}\}.$$

Definition 1.1. The margin $\rho_h(x,y)$ for a scoring function $h : \mathcal{X} \to \mathcal{Y}$ at a labeled example (x,y) is defined as

$$\rho_h(x,y) \triangleq h(x,y) - \max_{y \neq y'} h(x,y').$$

A scoring function h misclassifies an example x iff $\rho_h(x,y) \leqslant 0$. For any $\rho > 0$, we can define the **empirical margin loss** of a hypothesis h for multi-class classification as

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

where Φ_ρ is the margin loss function defined as $\Phi_\rho(x) = \mathbb{1}_{\{x \leqslant 0\}} + (1 - \frac{x}{\rho})\mathbb{1}_{\{0 \leqslant x \leqslant \rho\}}$.

Remark 1. Since $\Phi_\rho(x) \leqslant \mathbb{1}_{\{x \leqslant \rho\}}$, we obtain $\hat{R}_{S,\rho}(h) \leqslant \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{\rho_h(x_i,y_i) \leqslant \rho\}}$.

Lemma 1.2. *Let $\mathcal{F}_1, \dots, \mathcal{F}_L$ be L hypothesis sets in $\mathbb{R}^{\mathcal{X}}$, and let $\mathcal{G} \triangleq \{\max\{h_1, \dots, h_L\} : h_i \in \mathcal{F}_i, i \in [L]\}$. Then, for any sample S of size m, the empirical Rademacher complexity of $\mathcal{G}$ can be upper bounded as*

$$\hat{\mathfrak{R}}_S(\mathcal{G}) \leqslant \sum_{\ell=1}^{L} \hat{\mathfrak{R}}_S(\mathcal{F}_\ell).$$

Proof. Let $S = (x_1, \dots, x_m) \in \mathcal{X}^m$ be a sample of size m. We show this for $L = 2$, and then it follows inductively. We observe that for $h_1 \in \mathcal{F}_1, h_2 \in \mathcal{F}_2$, we have $h_1 \vee h_2 = \frac{1}{2}(h_1 + h_2 + |h_1 - h_2|)$. Therefore, we can write from the definition of Rademacher complexity, that

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{G}) &= \frac{1}{m}\mathbb{E}_\sigma[\sup_{h_1 \in \mathcal{F}_1, h_2 \in \mathcal{F}_2} \sum_{i=1}^{m} \sigma_i \max\{h_1(x_i), h_2(x_i)\}] \\
&= \frac{1}{2m}\mathbb{E}_\sigma[\sup_{h_1 \in \mathcal{F}_1, h_2 \in \mathcal{F}_2} \sum_{i=1}^{m} \sigma_i (h_1 + h_2 + |h_1 - h_2|)(x_i)] \\
&\leqslant \frac{1}{2}(\hat{\mathfrak{R}}_S(\mathcal{F}_1) + \hat{\mathfrak{R}}_S(\mathcal{F}_2)) + \frac{1}{2m}\mathbb{E}_\sigma[\sup_{h_1 \in \mathcal{F}_1, h_2 \in \mathcal{F}_2} \sum_{i=1}^{m} \sigma_i \left|h_1 - h_2\right|(x_i)].
\end{aligned}$$

The result follows from Talagrand's Lemma since $x \mapsto |x|$ is 1-Lipschitz. □

Definition 1.3. For any family $\mathcal{H}$ of hypotheses mapping $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we define

$$\Pi_1(\mathcal{H}) \triangleq \{x \mapsto h(x,y) : y \in \mathcal{Y}, h \in \mathcal{H}\}.$$

Remark 2. Recall that for a family of functions $\mathcal{G} \subseteq [0,1]^{\mathcal{Z}}$ and any *i.i.d.* sample $S \in \mathcal{Z}^m$, we have with probability greater than or equal to $1 - \delta$, for any function $g \in \mathcal{G}$

$$\mathbb{E}g(z) \leqslant \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(\mathcal{G}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

This follows from the application of McDiarmid's inequality. In addition, recall that the empirical Rademacher complexity of family $\mathcal{G}$ for $\sigma : \Omega \to \{-1,1\}^m$ *i.i.d.* Rademacher random sequence, is given by

$$\mathcal{R}_m(\mathcal{G}) \triangleq \frac{1}{m}\mathbb{E}_\sigma\Big[\sup_{g\in\mathcal{G}}\sum_{i=1}^m \sigma_i g(z_i)\Big].$$

Theorem 1.4 (Margin bound for multi-class classification). *Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a hypothesis set with $\mathcal{Y} = [k]$. Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following multi-class classification generalization bound holds for all $h \in \mathcal{H}$*

$$R(h) \leqslant \hat{R}_{S,\rho}(h) + \frac{4k}{\rho}\mathcal{R}_m(\Pi_1(\mathcal{H})) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

Proof. Let us define the margin $\rho_{\theta,h}(x,y) \triangleq \min_{y'\in\mathcal{Y}}[h(x,y) - h(x,y') + \theta\mathbb{1}_{\{y=y'\}}]$ for some constant $\theta > 0$. Then, we observe that

$$\rho_{\theta,h}(x,y) \leqslant \min_{y'\neq y}[h(x,y) - h(x,y') + \theta\mathbb{1}_{\{y=y'\}}] = \rho_h(x,y).$$

Therefore, it follows that $\mathbb{1}_{\{\rho_h(x,y)\leqslant 0\}} \leqslant \mathbb{1}_{\{\rho_{\theta,h}(x,y)\leqslant 0\}}$. Since $\mathbb{1}_{\{u\leqslant 0\}} \leqslant \Phi_\rho(u)$ for all $u \in \mathbb{R}$, we have $R(h) = \mathbb{E}\mathbb{1}_{\{\rho_h(x,y)\leqslant 0\}} \leqslant \mathbb{E}\mathbb{1}_{\{\rho_{\theta,h}(x,y)\leqslant 0\}} \leqslant \mathbb{E}\Phi_\rho(\rho_{\theta,h}(x,y))$. Defining the family of functions $\tilde{\mathcal{H}} \triangleq \{(x,y) \mapsto \rho_{\theta,h}(x,y) : h \in \mathcal{H}\}$ and family of composition of functions $\tilde{\mathscr{H}} \triangleq \{\Phi_\rho \circ \tilde{h} : \tilde{h} \in \tilde{\mathcal{H}}\}$. Applying the remark to family $\tilde{\mathscr{H}}$, we get

$$R(h) \leqslant \mathbb{E}\Phi_\rho(\rho_{\theta,h}(x,y)) \leqslant \frac{1}{m}\sum_{i=1}^m \Phi_\rho(\rho_{\theta,h}(x_i,y_i)) + 2\mathcal{R}_m(\tilde{\mathscr{H}}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

Fixing $\theta = 2\rho$, we observe that $\rho_{\theta,h}(x_i,y_i) = \rho_h(x_i,y_i)$ if $\rho_h(x_i,y_i) < 0$, and $\rho_{\theta,h}(x_i,y_i) = 2\rho \leqslant \rho_h(x_i,y_i)$ otherwise. This implies that

$$\Phi_\rho(\rho_{\theta,h}(x_i,y_i)) = \mathbb{1}_{\{\rho_{\theta,h}(x_i,y_i)\leqslant 0\}} + (1 - \frac{\rho_{\theta,h}(x_i,y_i)}{\rho})\mathbb{1}_{\{0\leqslant\rho_{\theta,h}(x_i,y_i)\leqslant\rho\}} = \mathbb{1}_{\{\rho_{\theta,h}(x_i,y_i)\leqslant 0\}} = \mathbb{1}_{\{\rho_h(x_i,y_i)\leqslant 0\}} = \Phi_\rho(\rho_h(x_i,y_i)).$$

From Talagrand's Lemma, we have $\mathcal{R}_m(\tilde{\mathscr{H}}) \leqslant \frac{1}{\rho}\mathcal{R}_m(\tilde{\mathcal{H}})$ since Φ_ρ is $\frac{1}{\rho}$-Lipschitz function. Therefore, with probability at least $1-\delta$, we have for all $h \in \mathcal{H}$

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

It suffices to show that $\mathcal{R}_m(\tilde{\mathcal{H}}) \leqslant 2k\mathcal{R}_m(\Pi_1(\mathcal{H}))$. To this end, we write

$$\begin{aligned}\mathcal{R}_m(\tilde{\mathcal{H}}) &= \frac{1}{m}\mathbb{E}_{S,\sigma}\Big[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i(h(x_i,y_i) - \max_y(h(x_i,y) - 2\rho\mathbb{1}_{\{y=y_i\}}))\Big]\\ &\leqslant \frac{1}{m}\mathbb{E}_{S,\sigma}\Big[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i h(x_i,y_i)\Big] + \frac{1}{m}\mathbb{E}_{S,\sigma}\Big[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i \max_y(h(x_i,y) - 2\rho\mathbb{1}_{\{y=y_i\}})\Big].\end{aligned}$$

We bound both the terms on the right hand side of the above equation individually. Defining $\epsilon_i \triangleq 2\mathbb{1}_{\{y_i=y\}} - 1 \in \{-1,1\}$, and observing that $\sigma_i\epsilon_i$ and σ_i have identical distribution, we can write the first term as

$$\frac{1}{2m}\mathbb{E}_{S,\sigma}\Big[\sup_{h\in\mathcal{H}}\sum_{i=1}^m\sum_{y\in\mathcal{Y}} \sigma_i h(x_i,y)(\epsilon_i+1)\Big] \leqslant \sum_{y\in\mathcal{Y}}\frac{1}{2m}\mathbb{E}_{S,\sigma}\Big[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i h(x_i,y)(\epsilon_i+1)\Big] \leqslant k\mathcal{R}_m(\Pi_1(\mathcal{H})).$$

We apply Lemma **??** to the second term, to obtain

$$\frac{1}{m}\mathbb{E}_{S,\sigma}\Big[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i\max_y(h(x_i,y) - 2\rho\mathbb{1}_{\{y=y_i\}})\Big] \leqslant \sum_{y\in\mathcal{Y}}\frac{1}{m}\mathbb{E}_{S,\sigma}\Big[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i(h(x_i,y) - 2\rho\mathbb{1}_{\{y=y_i\}})\Big]$$

$$= \sum_{y\in\mathcal{Y}}\frac{1}{m}\mathbb{E}_{S,\sigma}\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i(h(x_i,y) \leqslant k\mathcal{R}_m(\Pi_1(\mathcal{H})).$$

□

Remark 3. Larger margin means smaller second term and larger first term. That is, there is a trade-off between empirical error and complexity.

1.1 Rademacher complexity of family $\Pi_1(\mathcal{H})$

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel and let $\Phi : \mathcal{X} \to \mathbb{H}$ be the associated feature map. In multi-class classification, a kernel-based hypothesis is based on k weight vectors $\mathbf{w}_1, \ldots, \mathbf{w}_k \in \mathbb{H}$, where each weight vector $\mathbf{w}_i$ defines a scoring function $x \mapsto \langle \mathbf{w}_i, \Phi(x) \rangle$ for each $i \in [k]$ and the class associated to point $x \in \mathcal{X}$ is given by $\arg\max_{y \in \mathcal{Y}} \langle \mathbf{w}_y, \Phi(x) \rangle$. Let $\mathbf{W} \triangleq \begin{bmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_k \end{bmatrix}^T$ and for $p \geqslant 1$ we define the $L_{\mathbb{H},p}$ group norm of $\mathbf{W}$ as

$$\|\mathbf{W}\|_{\mathbb{H},p} \triangleq \Big(\sum_{i=1}^{k} \|\mathbf{w}_i\|_{\mathbb{H}}^{p} \Big)^{\frac{1}{p}}.$$

For any $p \geqslant 1$, the family of kernel-bases hypotheses under consideration is

$$\mathcal{H}_{K,p} \triangleq \Big\{ (x, y) \mapsto \langle \mathbf{w}_y, \Phi(x) \rangle : \|\mathbf{W}\|_{\mathbb{H},p} \leqslant \Lambda \Big\}.$$

Proposition 1.5 (Rademacher complexity of multi-class 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 the associated feature mapping. Assume that there exists $r > 0$ such that $K(x, x) \leqslant r^2$ for all $x \in \mathcal{X}$. Then, for any $m \in \mathbb{N}$, we have*

$$\mathcal{R}_m(\Pi_1(\mathcal{H}_{K,p})) \leqslant \sqrt{\frac{r^2 \Lambda^2}{m}}.$$

Proof. Let $S \in \mathcal{X}^m$ be an *i.i.d.* sample. We observe that for each weight vector, we have $\|\mathbf{w}_i\|_{\mathbb{H}} \leqslant \|\mathbf{W}\|_{\mathbb{H},p}$ for all $i \in [k]$. Thus, for any $\mathbf{W} \in \mathcal{H}_{K,p}$, we have $\mathbf{w}_i \leqslant \Lambda$ for all $i \in [k]$. Therefore, form Cauchy-Schwarz and Jensen's inequality, we have

$$\mathcal{R}_m(\Pi_1(\mathcal{H}_{K,p})) = \frac{1}{m} \mathbb{E}_{S,\sigma} \Big[\sup_{y \in \mathcal{Y}, \|\mathbf{W}\| \leqslant \Lambda} \Big\langle \mathbf{w}_y, \sum_{i=1}^{m} \sigma_i \Phi(x_i) \Big\rangle \Big] \leqslant \frac{\Lambda}{m} \mathbb{E}_{S,\sigma} \Big\| \sum_{i=1}^{m} \sigma_i \Phi(x_i) \Big\|_{\mathbb{H}} \leqslant \frac{\Lambda}{m} \Big(\mathbb{E}_{S,\sigma} \Big\| \sum_{i=1}^{m} \sigma_i \Phi(x_i) \Big\|_{\mathbb{H}}^2 \Big)^{\frac{1}{2}}.$$

Since the Rademacher random sequence σ is *i.i.d.* zero mean, we get $\mathbb{E}_{S,\sigma} \|\sum_{i=1}^{m} \sigma_i \Phi(x_i)\|_{\mathbb{H}}^2 = \mathbb{E}_S \sum_{i=1}^{m} \|\Phi(x_i)\|_{\mathbb{H}}^2 = \mathbb{E}_S \sum_{i=1}^{m} K(x_i, x_i) \leqslant mr^2$, and the result follows. □

Corollary 1.6 (Margin bound for multi-class classification with 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 an associated feature map. Assume that there exists $r > 0$ such that $K(x, x) \leqslant r^2$ for all $x \in \mathcal{X}$. Fix $\rho > 0$. Then, with probability at least $1 - \delta$ for all $h \in \mathcal{H}_{K,p}$*

$$R(h) \leqslant \hat{R}_{S,\rho}(h) + 4k \sqrt{\frac{r^2 \Lambda^2}{\rho^2 m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$