Theory of Machine Learning

Exercise sheet 9 — Session 9

Exercise I (elementary properties) ✎. Show the following elementary properties of the Rademacher complexity:

1. if $\mathcal{G} \subset \mathcal{G}'$, then $R_n(\mathcal{G}) \leq R_n(\mathcal{G}')$;
2. $R_n(\mathcal{G} + \mathcal{G}') = R_n(\mathcal{G}) + R_n(\mathcal{G}')$;
3. $R_n(\alpha\mathcal{G}) = |\alpha| \, R_n(\mathcal{G})$;
4. if g_0 is a function, $R_n(\mathcal{G} + \{g_0\}) = R_n(\mathcal{G})$;
5. $R_n(\mathcal{G}) = R_n(\mathrm{conv}(\mathcal{G}))$.

Exercise II (Massart's lemma) ✎. Let us assume that $\mathcal{G}$ is *finite*, that is, $\mathcal{G} = \{g_1, \dots, g_m\}$. Let us assume further that $\frac{1}{n}\sum_{i=1}^{n} g_j(X_i)^2 \leq R^2$ for all $j \in [d]$. Show that the Rademacher complexity of the function class $\mathcal{G}$ satisfies

$$R_n(\mathcal{G}) \leq \sqrt{\frac{2\log m}{n}} R \, .$$

For simplicity's sake, we consider the X_is fixed.

1. Given $\lambda > 0$, show that

$$\exp\left(\lambda \mathbb{E}_\varepsilon\left[\sup_{g\in\mathcal{G}} \sum_{i=1}^{n} \varepsilon_i g(X_i)\right]\right) \leq \sum_{g\in\mathcal{G}} \mathbb{E}_\varepsilon\left[\exp\left(\lambda \sum_{i=1}^{n} \varepsilon_i g(X_i)\right)\right] .$$

   (*Hint: Jensen's inequality and property of* sup.)

2. Show that

$$\sum_{g\in\mathcal{G}} \mathbb{E}_\varepsilon\left[\exp\left(\lambda \sum_{i=1}^{n} \varepsilon_i g(X_i)\right)\right] = \sum_{g\in\mathcal{G}} \prod_{i=1}^{n} \frac{1}{2}\left(\exp\left(\lambda g(X_i)\right) + \exp\left(-\lambda g(X_i)\right)\right) .$$

   (*Hint*: independence of the ε_is and direct computation of the remaining expectation.)

3. Using $\frac{\mathrm{e}^{x} + \mathrm{e}^{-x}}{2} \leq \mathrm{e}^{x^2/2}$ (for all $x \in \mathbb{R}$), show that

$$\sum_{g\in\mathcal{G}} \prod_{i=1}^{n} \frac{1}{2}\left(\exp\left(\lambda g(X_i)\right) + \exp\left(-\lambda g(X_i)\right)\right) \leq \sum_{g\in\mathcal{G}} \exp\left(\frac{\lambda^2}{2} \sum_{i=1}^{n} g(X_i)^2\right) .$$

4. Show that

$$\sum_{g\in\mathcal{G}} \exp\left(\frac{\lambda^2}{2} \sum_{i=1}^{n} g(X_i)^2\right) \leq m \exp\left(\frac{n\lambda^2 R^2}{2}\right) .$$

   (*Hint*: refer to the assumptions in the exercise statement on g and $\mathcal{G}$.)

5. By putting everything together, show that

$$R_n(\mathcal{G}) \leq \frac{1}{n\lambda} \log m + \frac{\lambda R^2}{2} .$$

6. Show that $\lambda^\star = \frac{1}{R}\sqrt{\frac{2\log m}{n}}$ minimizes the previous bound.
7. Show the Massart's lemma bound.

Exercise III (On the generalization bound of linear regression in fixed design) ✎. The objective of this exercise is to prove a generalization bound for an OLS estimator in the fixed design setting.

Consider vector-valued inputs and real-valued outputs ($\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$) with $X := (X_1, \dots, X_n)^\top \in \mathbb{R}^{n\times d}$ the input vector and $Y := (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$ the response vector. Let $\phi(x) = (x_1, \dots, x_d)^\top$ and $\Phi \in \mathbb{R}^{n\times d}$ the matrix of inputs with row i defined as $\Phi_{i,:} := \phi(X_i)^\top$. We work in the fixed design setting where for a fixed input $X \in \mathbb{R}^{n\times d}$, the output is $Y = \Phi\theta + \varepsilon$ (ε i.i.d. $\mathcal{N}(0,1)$) and $\theta \in \mathbb{R}^d$. The true risk $\mathcal{R}(\theta)$ of θ and its empirical counterpart $\widehat{\mathcal{R}}(\theta)$ are defined as:

$$\mathcal{R}(\theta) := \mathbb{E}_\varepsilon\left[\frac{1}{n}\|Y - \Phi\theta\|_2^2\right],$$

$$\widehat{\mathcal{R}}(\theta) := \frac{1}{m}\sum_{i=1}^m \frac{1}{n}\left\|Y^{(i)} - \Phi\theta\right\|_2^2,$$

where $\{Y^{(i)}\}_{i=1}^m$ are i.i.d. random vectors sampled as $Y^{(i)} = \Phi\theta + \varepsilon^{(i)}$ ($\varepsilon^{(i)}$ i.i.d. $\mathcal{N}(0, \mathbf{I}_n)$).

Finally, a remainder on chi-squared distribution: if $Z_1, \dots, Z_n$ are independent standard normal $\mathcal{N}(0,1)$ random variables, then the sum of their squares, $K := \sum_{i=1}^n Z_i^2$, is distributed according to the chi-squared distribution with n degrees of freedom, which is denoted as $K \sim \chi^2(n)$. Also, $\mathbb{E}[K] = n$ and $\mathrm{Var}(K) = 2n$.

The objective is to compute a generalization bound for a fixed θ using Chebyshev's inequality:

1. Show that $Y_j^{(i)} \sim \mathcal{N}(\Phi_{j,:}\theta, 1)$ with $i \in [\![m]\!]$ and $j \in [\![n]\!]$.
2. Show that $\left\|Y^{(i)} - \Phi\theta\right\|_2^2 = \left(Y^{(i)} - \Phi\theta\right)^\top\left(Y^{(i)} - \Phi\theta\right)$ is distributed as $\chi^2(n)$.
   (Hint: rewrite the previous quantity as a sum.)
3. Compute the variance of $\left\|Y^{(i)} - \Phi\theta\right\|_2^2$.
4. Compute the variance of $\widehat{\mathcal{R}}(\theta)$.
   (Hint: the elements of the sum $\widehat{\mathcal{R}}(\theta)$ are independent.)
5. Using Chebyshev's inequality with $t > 0$, show that:
   $$\mathbb{P}\left(\left|\widehat{\mathcal{R}}(\theta) - \mathcal{R}(\theta)\right| > t\right) \le \frac{\mathrm{Var}\left(\widehat{\mathcal{R}}(\theta)\right)}{t^2} = \frac{2}{mnt^2}.$$
6. Solve in $t > 0$, $\delta = \frac{2}{mnt^2}$.
7. Rewrite the previous bound as follows:
   $$\mathbb{P}\left(\left|\widehat{\mathcal{R}}(\theta) - \mathcal{R}(\theta)\right| \le \sqrt{\frac{2}{mn\delta}}\right) \ge 1 - \delta.$$
8. (Numerical application) set $\delta = 0.01$ in the previous bound and rewrite it as:
   $$\mathbb{P}\left(\left|\widehat{\mathcal{R}}(\theta) - \mathcal{R}(\theta)\right| \le 10\sqrt{\frac{2}{mn}}\right) \ge 0.99.$$