Empirical Processes

MAA110 - EPFL

Nikitas Georgakis*, Myrto Limnios*

08/04/2025

Notations. Without additional notice, we consider the same notations as in the lecture notes.

Linear classifiers. In this session, we will prove fundamental generalization guarantees for linear classifiers.

Consider the r.v. X valued in $\mathbb{R}^d$, with $d \in \mathbb{N}^*$, endowed with its Borelian σ-algebra, and let its random label Y valued in $\{0, 1\}$. Define P the joint distribution of (X, Y). We consider the class $\mathcal{H}$ of all possible classifiers defined as measurable functions $h : \mathbb{R}^d \to \{0, 1\}$, i.e., it assigns to any point x a label 0 or 1.

The goal is to find the *best* classifier h minimizing the classification risk

$$\mathcal{L} : h \in \mathcal{H} \mapsto \mathbb{P}(Y \neq h(X)) \in [0, 1] \ , \tag{1}$$

where we consider that $\mathcal{L}^* = \inf_{h \in \mathcal{H}} \mathcal{L}(h)$ exists. Consider an i.i.d. sample $\{(X_i, Y_i)\}_{i \leq n}$ drawn from P, and denote by $\widehat{h}$ the empirical minimizer of the empirical risk function defined as follows:

$$\widehat{h} \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} 1\{h(X_i) \neq Y_i\} =: \arg\min_{h \in \mathcal{H}} \mathcal{L}_n(h) \ .$$

Exercise 1 (Binary Loss). *In this setup, we consider linear classifiers taking the form of*

$$h_{\theta_0, \theta} : x \mapsto 1\{\langle \theta, x \rangle + \theta_0 > 0\} \ , \tag{2}$$

where $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ and $\theta_0 \in \mathbb{R}$ are the parameters we want to optimize. We denote by $\mathcal{H}_0$ this class.

1. *Show that we can equivalently consider the problem with $\theta_0 = 0$. We set $\theta_0 = 0$ in the remaining exercise.*
2. *Prove that*
$$\mathcal{L}(\widehat{h}) - \inf_{h \in \mathcal{H}_0} \mathcal{L}(h) \leq 2 \sup_{h \in \mathcal{H}_0} |\mathcal{L}_n(h) - \mathcal{L}(h)| \ . \tag{3}$$
3. *Prove that the class linear classifiers can be reformulated as a* VC*-class of sets and provide an upperbound on its* VC*-dimension.*

*{**first.last**}**@epfl.ch**, Office MA1493, CM1618

4. *Prove that, for any $\delta \in (0,1)$, it holds true with probability at least $1-\delta$*

$$\mathcal{L}(\widehat{h}) - \inf_{h \in \mathcal{H}_0} \mathcal{L}(h) \leq 8\sqrt{\frac{2(d+1)\log(n+1)}{n}} + 8\sqrt{\frac{\log(1/\delta)}{n}} \ , \tag{4}$$

as soon as $n \geq 2$.

5. *Correct the previous statement and give an order of lower bound for the sample size n for the inequality to hold true.*

6. *Notice that the optimal output $\widehat{h}$ depends on the random dataset. Prove a generalization guarantee for $\widehat{h}$, i.e., derive the upperbound of $\mathbb{E}[\mathcal{L}(\widehat{h})] - \inf_{h \in \mathcal{H}_0} \mathcal{L}(h)$.*

7. *Conclude that, if we are able to generate an infinite amount of data from P, then it holds true that*

$$\mathcal{L}(\widehat{h}) - \inf_{h \in \mathcal{H}_0} \mathcal{L}(h) \to 0 \ , \tag{5}$$

in probability, and for $n \to \infty$.

Exercise 2 (Hinge Loss)**.** *We consider here $\theta_0 = 0$. We now study the generalization perfomance of Support Vectore Machine (SVM) algorithm. Consider the same framework, and defined the hinge loss composing the hypothesis class $\mathcal{H}_1$, by*

$$h_\theta : x \mapsto \max(0, 1 - y\langle \theta, x \rangle) = (1 - y\langle \theta, x \rangle)_+ \ , \tag{6}$$

with $\theta \in \Theta \subseteq \mathbb{R}^d$, and with associated risk function

$$\mathcal{L} : h \in \mathcal{H} \mapsto \mathbb{E}[h(X,Y)] \ . \tag{7}$$

Based on an i.i.d. random sample $\{(X_i, Y_i)\}_{i \leq n}$ drawn from P, then we can consider the optimal empirial discriminant function as being solution of

$$\widehat{h} \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - Y_i\langle \theta, X_i \rangle) =: \arg\min_{h \in \mathcal{H}} \mathcal{L}_n(h) \ .$$

1. *Show the graph of the hinge loss function to explain why the associate classifier is of the form $1\{sign(\langle \theta, x \rangle) \neq y\}$.*

2. *Prove that the hinge loss function is 1-Lipschitz.*

3. *Prove similar upperbound as in question 4, using the empirical Rademacher complexity associated with the hinge loss, ie., with probability at least $1-\delta$,*

$$\mathcal{L}(\widehat{h}) - \inf_{h \in \mathcal{H}_1} \mathcal{L}(h) \leq 2\mathcal{R}_n(\mathcal{H}_1) + C\sqrt{\frac{\log(1/\delta)}{n}} \ , \tag{8}$$

and determine the explicit constant $C > 0$.