Machine Learning for EDS

Tutorial week 5

2023/2024

Generically in the following problems, consider $\mathcal{X}$ to be a measurable space of features, $\mathcal{Y}$ a measurable space of outputs, $\mathcal{F}$ be the set of all predictors from $\mathcal{X}$ into $\mathcal{Y}$, P a distribution over $\mathcal{X} \times \mathcal{Y}$, $(X, Y) \sim P$, $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be an i.i.d. P-distributed sample, with D_n independent of (X, Y).

Problem 1 (n-th shatter coefficient and VC-dimension)

Consider the binary classification framework with $\mathcal{Y} = \{0, 1\}$, c the 0-1 cost, i.e., $c(y, y') = \mathbb{1}_{y \neq y'}$, for any $y, y' \in \{0, 1\}$.

1. Recall the definition of the n-th shatter coefficient of a model S, denoted by $\mathcal{C}(S, n)$. Also recall the definition of the VC-dimension of a model S.

2. In the lecture, we proved that the model S of interval classifiers of the form

$$f_{a,b} : x \in \mathbb{R} \mapsto \mathbb{1}_{x \in [a,b]}, \qquad \text{for any} \quad a, b \in \mathbb{R},\ a < b\,.$$

has VC-dimension 2. Compute its shatter coefficient $\mathcal{C}(S, n)$ for $n \geq 0$. Compare the bound you obtain with the general bound $\mathcal{C}(S, n) \leq (en/d)^d$ we derived in the lecture, where d denotes the VC-dimension of the model S.

Now we consider the following example. Let $\mathcal{X} = \mathbb{R}^2$ and consider a model S that contains predictors that are axis-aligned rectangles in $\mathbb{R}^2$. More specifically, we consider a model S that contains predictors $f_{\boldsymbol{a},\boldsymbol{b}}$ of the form $f_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x}) = \mathbb{1}_{a_1 < x_1 < a_2} \mathbb{1}_{b_1 < x_2 < b_2}$, where $\boldsymbol{a} = (a_1, a_2) \in \mathbb{R}^2$ and $\boldsymbol{b} = (b_1, b_2) \in \mathbb{R}^2$ are such that $a_1 < a_2$ and $b_1 < b_2$. So

$$S = \{f_{\boldsymbol{a},\boldsymbol{b}} :\ \boldsymbol{a} \in \mathbb{R}^2, \boldsymbol{b} \in \mathbb{R}^2, a_1 < a_2 \text{ and } b_1 < b_2\}\,.$$

3. Does the model S shatter the points $\boldsymbol{x}_1 = (0, 0)$, $\boldsymbol{x}_2 = (1, 0)$, $\boldsymbol{x}_3 = (0, 1)$ and $\boldsymbol{x}_4 = (1, 1)$ (make a drawing)? If not, can you think if an alternative set of four points that the model *does* shatter?

4. Argue that there exists no sample of 5 points that is shattered by the model S.

5. What is the VC-dimension of this model?
6. Discuss whether the answers to the three questions above will change if S only contains predictors based on axis-aligned squares, i.e. if we restrict $a_2 - a_1 = b_2 - b_1$ for each $f_{\boldsymbol{a},\boldsymbol{b}}$ in S.

Problem 2 (Learning guarantee for empirical convexified risk minimizer)

Consider the binary classification framework with $\boldsymbol{\mathcal{Y} = \{-1, 1\}}$, c the 0-1 cost, i.e., $c(y, y') = \mathbb{1}_{y \neq y'}$, for any $y, y' \in \{-1, 1\}$. Denote $\eta(x) := \mathbb{P}(Y = 1|X = x) = 1 - \mathbb{P}(Y = -1|X = x)$ for all $x \in \mathbb{R}$.

1. Within a certain model, we would like to find a predictor minimising the empirical risk over a given sample. In the framework of binary classification with 0-1 cost, what practical optimisation difficulty do we face?

Instead of directly optimising the empirical risk with 0-1 cost, we introduce the notion of pseudo-classifiers and convexified risk. A pseudo-classifier is any measurable function $h : \mathcal{X} \to \mathbb{R}$. We consider a convex pseudo-loss function Φ that upperbounds the 0-1 cost function: $\Phi : z \mapsto \exp(z)$. The convexified risks with respect to the exponential loss are defined by:

$$\mathcal{R}_P^{\Phi}(h) := \mathbb{E}\Big[\Phi\big(-Yh(X)\big)\Big|D_n\Big],$$
$$\widehat{\mathcal{R}}_n^{\Phi}(h) := \frac{1}{n}\sum_{i=1}^{n}\Phi\big(-Y_i h(X_i)\big),$$

To a pseudo-classifier h, one can associate a classifier as $f_h : x \mapsto \operatorname{sgn}\big(h(x)\big)$.

2. Show that for any fixed $x \in \mathcal{X}$, the function $r_x : a \mapsto \mathbb{E}\Big[\exp\big(-Ya\big)\Big|X = x\Big]$ is strictly convex on $\mathbb{R}$.
3. Recall in one to two sentences the interpretation of the case: $\eta(X) \in \{0, 1\}$ almost surely.

In the rest of the problem, assume that $0 < \eta(x) < 1$ for all $x \in \mathcal{X}$.

4. Show that for any $a \in \mathbb{R}$:

$$r'_x(a) = e^a(1 - \eta(x)) - e^{-a}\eta(x),$$

where $r'_x(a) = \partial r_x(a)/\partial a$.

[*Hint: Use the fact that* $\mathbb{E}[-Ye^{-Ya}|X = x] = \mathbb{E}[-Ye^{-Ya}(\mathbb{1}_{Y=-1} + \mathbb{1}_{Y=1})|X = x]$]

5. Show that r_x admits a unique global minimum at $a_x^* \in \mathbb{R}$ which reads:
$$a_x^* = \frac{1}{2} \ln \left(\frac{\eta(x)}{1-\eta(x)} \right).$$

6. Let $\mathcal{H}$ be the set of all pseudo-classifiers, $\mathcal{R}_P^{\Phi *} := \inf_{h \in \mathcal{H}} \mathcal{R}_P^{\Phi}(h)$ and define also the pseudo-classifier $h^* : x \mapsto a_x^*$. Show that $\mathcal{R}_P^{\Phi}(h^*) = \mathcal{R}_P^{\Phi *}$.

7. Using the expression h^*, show that:
$$\mathcal{R}_P^{\Phi *} = \mathbb{E}\left[\left(\frac{\eta(X)}{1-\eta(X)} \right)^{1/2} \mathbb{1}_{Y=-1} \right] + \mathbb{E}\left[\left(\frac{\eta(X)}{1-\eta(X)} \right)^{-1/2} \mathbb{1}_{Y=+1} \right].$$

8. Using the law of iterated expectations, show that:
$$\mathbb{E}\left[\left(\frac{\eta(X)}{1-\eta(X)} \right)^{1/2} \mathbb{1}_{Y=-1} \right] = \mathbb{E}\left[\sqrt{\eta(X)(1-\eta(X))} \right].$$

9. Show that:
$$\mathcal{R}_P^{\Phi *} = 2\mathbb{E}\left[\sqrt{\eta(X)(1-\eta(X))} \right].$$

We will now show that a pseudo-classifier h with low convexified risk $\mathcal{R}_P^{\Phi}(h)$ entails a classifier f_h with low generalisation risk $\mathcal{R}_P(f_h)$. To this end, let us introduce the function:
$$\forall u \in [0,1], \quad G(u) := \inf_{\alpha \in \mathbb{R}} \Big\{ u\Phi(-\alpha) + (1-u)\Phi(\alpha) \Big\},$$
and let us assume the following result:

Zhang's (2004) lemma *For Φ a convex, non-negative, increasing function such that:*

(a) $\Phi(z) \geq \mathbb{1}_{z>0}$ for all $z \in \mathbb{R}$, $\Phi(0) = 1$ and $\lim_{z \to -\infty} \Phi(z) = 0$.

(b) there exist constants $c > 0$ and $s \geq 1$ such that for any $u \in [0,1]$,
$$\left| \frac{1}{2} - u \right|^s \leq c^s \left(1 - G(u)\right),$$

then for any pseudo-classifier $h \in H$ and its associated plug-in classifier $f_h \in \mathcal{F}$,
$$\mathcal{R}_P(f_h) - \mathcal{R}_P^* \leq 2c(\mathcal{R}_P^{\Phi}(h) - \mathcal{R}_P^{\Phi *})^{1/s}.$$

where $\mathcal{R}_P^*$ denotes Bayes risk.

10. In our context, with $\Phi = \exp$, it holds that $G(u) = 2\sqrt{u(1-u)}$ (you are not required to show this). Verify that assumptions (a) and (b) of Zhang's lemma hold with $c = \frac{1}{\sqrt{2}}$ and $s = 2$, and apply the lemma.

    [*Hint: for point (b), first show that* $|1/2 - u|^2 = \left(\frac{1}{2} - \sqrt{u(1-u)}\right)\left(\frac{1}{2} + \sqrt{u(1-u)}\right)$ *and use that* $u(1-u)$ *is upper-bounded by* $1/4$].

11. Let $S \subset \mathcal{H}$ be a model of pseudo-classifiers and define an empirical convexified risk minimiser (ECRM) as any pseudo-classifier $\hat{h} \in S$ such that $\widehat{\mathcal{R}}_n^{\Phi}(\hat{h}) = \inf_{h \in S} \widehat{\mathcal{R}}_n^{\Phi}(h)$. Propose an error decomposition for the convexified risk similar to the decomposition of excess risk of a sample-based predictor into estimation and approximation error.

12. Show that:

$$\mathcal{R}_P^{\Phi}(\hat{h}) - \inf_{h \in S} \mathcal{R}_P^{\Phi}(h) \leq 2 \sup_{h \in S} \left| \mathcal{R}_P^{\Phi}(h) - \widehat{\mathcal{R}}_n^{\Phi}(h) \right|.$$

    [*Hint: first show that* $\mathcal{R}_P^{\Phi}(\hat{h}) - \mathcal{R}_P^{\Phi}(g) \leq 2 \sup_{h \in S} \left| \mathcal{R}_P^{\Phi}(h) - \widehat{\mathcal{R}}_n^{\Phi}(h) \right|$ *for any pseudo-classifier* $g \in S$.]

Assume that we have a model $S \subset \mathcal{H}$ such that $\inf_{h \in S} \mathcal{R}_P^{\Phi}(h) = \mathcal{R}_P^{\Phi *}$, and such that the following result holds:

Lemma 2 of Lugosi and Vayatis (2004) *For any* $n \geq 1$,

$$\mathbb{E}\Big[\sup_{h \in S} \big|\mathcal{R}_P^{\Phi}(h) - \widehat{\mathcal{R}}_n^{\Phi}(h)\big|\Big] \leq 4e\sqrt{\frac{2d\ln(4n+2)}{n}},$$

where $d = V(S_{sgn})$ *is the VC dimension of the set of classifiers* $f_h = sgn(h)$ *associated to the pseudo-classfiers* h *in the model* S*:* $S_{sgn} = \{f_h : h \in S\}$.

13. Interpret the assumption $\inf_{h \in S} \mathcal{R}_P^{\Phi}(h) = \mathcal{R}_P^{\Phi *}$.

14. With $f_{\hat{h}}$ the classifier associated with the ECRM pseudo-classifier $\hat{h}$, show that:

$$\mathbb{E}\Big[\mathcal{R}_P(f_{\hat{h}})\Big] - \mathcal{R}_P^* \leq 4e^{1/2}\left(\frac{2d\ln(4n+2)}{n}\right)^{1/4},$$

    where d is the VC dimension of the set of classifiers f_h associated to the pseudo-classfiers h in the model S.

    [*Hint: Start by applying the result from question 10 to* $\hat{h}$.]