Machine Learning for EDS

Tutorial week 3

2023/2024

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

Problem 1 (Empirical Risk Minimization with quadratic cost) Consider the regression framework $\mathcal{Y} = \mathbb{R}$ with quadratic cost $c(y, y') = (y - y')^2$ for $y, y' \in \mathcal{Y}$, and assume $\mathbb{E}[Y^2] < +\infty$. Consider the case that $\mathcal{X} = \mathbb{R}^2$ with $X = (X_1, X_2) \in \mathcal{X}$, $\mathbb{E}[X_i^2] < +\infty$ for $i =, 1, 2$, and where Y is such that:

$$Y = aX_1 + bX_1X_2 + \varepsilon,$$

where $a, b \in \mathbb{R}$, and ε is an error term independent of both X and Y, such that $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] = \sigma^2$.

1. What are Bayes risk and Bayes predictors in this framework? You can use that in the regression framework with quadratic cost, $\eta(X) = \mathbb{E}[Y|X]$ is a Bayes predictor and Bayes risk is equal to $\mathbb{E}[(Y - \eta(X))^2]$.

The goal of the following questions is to study the approximation of the above relationship by a linear function of X_1. Let us introduce predictors f_d of the form

$$f_d : (x_1, x_2) \in \mathbb{R}^2 \mapsto dx_1\,, \qquad \text{for any } d \in \mathbb{R}\,,$$

and define the model $S \subset \mathcal{F}$, containing all predictors of this type:

$$S := \Big\{ f_d : d \in \mathbb{R} \Big\}.$$

Let $D_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ with $\boldsymbol{x}_i = (x_{i1}, x_{i2}) \sim (X_1, X_2)$ for all $i = 1, \ldots, n$, be an independent identically distributed random sample following the studied relationship:

$$y_i = ax_{i1} + bx_{i1}x_{i2} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

2. Write down the empirical risk $\widehat{\mathcal{R}}_n(f)$ of a predictor $f \in S$ over the sample D_n.

Let $\hat{f} \in S$ be an empirical risk minimiser (ERM) of $\widehat{\mathcal{R}}_n(f)$ over the model S on the sample D_n, i.e., $\widehat{\mathcal{R}}_n(\hat{f}) = \inf_{f \in S} \widehat{\mathcal{R}}_n(f)$.

3. By minimising the function $d \mapsto \widehat{\mathcal{R}}_n(f_d)$, show that $\hat{f} = f_{\hat{d}}$ with

$$\hat{d} = \frac{\sum_{i=1}^n x_{i1} y_i}{\sum_{j=1}^n x_{j1}^2}.$$

4. Assume X_1 and X_2 are independent. Show that the expected risk of $f_{\hat{d}}$ for the sample D_n is given by

$$\mathcal{R}_P(f_{\hat{d}}) = \mathbb{E}[X_1^2]\ \mathbb{E}[(a - \hat{d} + bX_2)^2 | D_n] + \mathbb{E}[\varepsilon^2]\,.$$

   Also give the excess risk of $f_{\hat{d}}$. Say, $\mathbb{E}[X_2] = 0$ and $\mathbb{E}[X_2^2] > 0$, then under which circumstances can the excess risk be equal to zero?

5. Explain in one sentence why $\hat{d}$ and $\mathcal{R}_P(f_{\hat{d}})$ are random variables.

6. Give two reasons why in practice we can typically not write down an explicit expession of the expected risk of an empirical risk minimizer $\hat{f}$.

Problem 2 (properties of plug-in estimators) Consider the binary classification framework with $\mathcal{Y} = \{0, 1\}$ and the 0-1 cost.

1. Recall the expressions of Bayes risk, Bayes classifiers and of the excess risk of a classifier $f \in \mathcal{F}$ in this framework.

2. We will assume in the rest of question 2 that the joint distribution P between features and output is a *zero-error* distribution: $\eta(X) \in \{0, 1\}$ almost surely.

   (a) Denoting by f^* a Bayes classifier, show that the latter assumption implies that $f^*(X) = Y$ almost surely. What is Bayes risk equal to? Interpret the *zero-error* assumption; do you think it is a restrictive assumption? (two to three sentences)

   (b) Let $\hat{\eta}$ be a regression learning rule. Recall the definition of the plug-in classifier associated to $\hat{\eta}$.

   (c) Letting D_n be a sample, denote $\hat{f}_{\hat{\eta}}(D_n)$ the plug-in classifier associated to $\hat{\eta}$. Show the following implication:

$$\hat{f}_{\hat{\eta}}(D_n; X) \neq f^*(X) \qquad \Longrightarrow \qquad \hat{\eta}(D_n; X) \leq \frac{1}{2} < \eta(X) \quad \text{or} \quad \eta(X) \leq \frac{1}{2} < \hat{\eta}(D_n; X).$$

   (d) Deduce that

$$2\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X) \neq f^*(X)} \leq 2|\hat{\eta}(D_n; X) - \eta(X)| \mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n;X) \neq f^*(X)}.$$

(e) Denoting $\ell(f^*, \hat{f}_{\hat{\eta}}(D_n))$ the excess risk of the plug-in classifier $\hat{f}_{\hat{\eta}}(D_n)$, obtain that

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq 2\sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2 \Big| D_n\Big] \mathbb{P}\Big(\hat{f}_{\hat{\eta}}(D_n; X) \neq f^*(X) \Big| D_n\Big)}.$$

Hint: use Cauchy-Schwarz inequality: $|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$

(f) Show that

$$\mathbb{P}\Big(\hat{f}_{\hat{\eta}}(D_n; X) \neq f^*(X) \Big| D_n\Big) = \mathcal{R}_P(\hat{f}_{\hat{\eta}}(D_n)) - \mathcal{R}_P^*.$$

Hint: use question 2.a.

(g) Deduce that the excess risk is upper-bounded as follows:

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq 4\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2 \Big| D_n\Big].$$

(h) Compare this bound with the one obtained in the lecture (*A good regression rule gives a good classification rule*): does it suggest lower or higher excess risk for the plug-in classifier?

3. Instead of the *zero-error* assumption, assume now that P satisfies the *margin condition*:

$$\mathbb{P}\Big(\Big|\eta(X) - \frac{1}{2}\Big| \geq h\Big) = 1, \qquad \text{for some } h \in [0, 1/2].$$

(a) What does the case $h = 1/2$ correspond to? Does the case $h = 0$ impose any restrictions on the joint distribution P of features and outputs? Is the margin condition more or less general than the zero-error assumption?

(b) Assume in the rest of the problem that the margin condition holds for some $h \in (0, 1/2)$. Using the margin condition, first prove that:

$$\mathbb{E}\Big[|\eta(X) - 1/2| \mathbb{1}_{\hat{f}_{\hat{\eta}}(D_n; X) \neq f^*(X)} \mathbb{1}_{|\eta(X) - 1/2| < h} \Big| D_n\Big] = 0.$$

(c) Then show that

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq 2\mathbb{E}\Big[|\hat{\eta}(D_n; X) - \eta(X)| \mathbb{1}_{\big|\hat{\eta}(D_n; X) - \eta(X)\big| \geq h} \Big| D_n\Big]$$

(d) Deduce that:

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq 2\sqrt{\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2 \Big| D_n\Big] \mathbb{P}\Big(|\hat{\eta}(D_n; X) - \eta(X)| \geq h \Big| D_n\Big)}.$$

(e) Finally, obtain the following upper-bound of the excess risk of the plug-in classifier under the margin condition:

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq \frac{2}{h}\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2 \Big| D_n\Big].$$

(f) Compare the previous inequality with the one obtained under the zero-error assumption and the one obtained in the lecture (*A good regression rule gives a good classification rule*). Comment in particular on the cases $h = 1/2$ and $h \to 0$.

(g) It is said that the learning rule $\hat{f}_{\hat{\eta}}$ is weakly consistent if its average excess risk over samples of size n tends to zero as n tends to infinity:

$$\mathbb{E}\Big[\ell(f^*, \hat{f}_{\hat{\eta}}(D_n))\Big] \underset{n\to+\infty}{\longrightarrow} 0.$$

Assume that the regression rule $\hat{\eta}$ is such that $\mathbb{E}\Big[\big(\hat{\eta}(D_n; X) - \eta(X)\big)^2\Big] \underset{n\to+\infty}{\sim} \frac{c}{n}$, for some positive constant $c > 0$. Show that the plug-in learning rule is weakly consistent:

i. under the margin condition.

ii. without the margin condition (i.e., under the assumption of the lecture, slide *A good regression rule gives a good classification rule*)

Hint: For ii, use Jensen's inequality.

Compare the *rate of convergence*, that is, the speed at which $\mathbb{E}\Big[\ell(f^*, \hat{f}_{\hat{\eta}}(D_n))\Big]$ tends to 0 with the sample size n in both cases. Do we have the same learning guarantee with and without assumptions on the data?