Machine Learning
EDS
Tutorial: Week 3

Janneke van Brummelen

Vrije Universiteit Amsterdam

Period 3.4
2023/2024

This tutorial

1. Short recap of this week's material
2. A small quiz
3. Discuss Problem 2 of Problem set 3

This week:

We have introduced:

- Empirical risk, a computable performance measure based on observations
- Models, a subset of predictors (flexibility vs generalisability trade-off)
- Learning rules: the theoretical counterpart of algorithms
- Empirical Risk Minimisation: a natural learning rule, which includes OLS and histogram regression
- Empirical Convexified Risk Minimisation: 'Approximate' ERM for classification using convex surrogate loss function

Table of Contents

Quiz: Question 1

Let $\mathcal{X}$ and $\mathcal{Y}$ be a feature space and an output space, respectively.
Let $\mathcal{F}$ be the set of all predictors.
Is the following statement true or false?

"*A learning rule is a function from* $(\mathcal{X} \times \mathcal{Y})^n$ *to* $\mathcal{F}$"

- True
- **False**

Quiz: Question 2

Is the following statement true or false?

"If for a particular sample D_n and cost function c, we have that two predictors f and g are such that

$$\widehat{\mathcal{R}}_n^c(f; D_n) < \widehat{\mathcal{R}}_n^c(g; D_n)\,,$$

then it must be the case that

$$\mathcal{R}_P^c(f) < \mathcal{R}_P^c(g)\text{"}$$

- True
- **False**

Quiz: Question 3

Let D_n be some sample of examples $(X_i, Y_i) \sim P$ for some joint distribution P. Is the following statement true or false?

"*A sample based predictor $\hat{f}(D_n)$ and its risk $\mathcal{R}_P(\hat{f}(D_n))$ are random*"

- **True**
- False

Table of Contents

Introduction Problem 2

In this problem we will look at properties of **plug-in classifiers**

Let P be a feature/label distribution on $\mathcal{X} \times \{0, 1\}$, let $\eta(X) = \mathbb{P}(Y = 1|X)$ and consider the 0-1 cost function.

We derived in the lecture slides that for any regression rule $\hat{\eta}$ and corresponding **plug-in classifier** $\hat{f}_{\hat{\eta}} = \mathbb{1}_{\hat{\eta}>1/2}$, we have:

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq 2\sqrt{\mathbb{E}\left[(\hat{\eta}(D_n; X) - \eta(X))^2 | D_n\right]}$$

where D_n denotes a sample $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ with iid $(X_i, Y_i) \sim P$.

During this tutorial, we will investigate if we can find tighter bounds if we impose restrictions on P

Problem 2.1

Preliminaries:

- Let $\mathcal{X}$ be a measurable space of features and let $\mathcal{Y} = \{0, 1\}$
- Let P a distribution over $\mathcal{X} \times \mathcal{Y}$, $(X, Y) \sim P$
- Let $\mathcal{F}$ be the set of all predictors from $\mathcal{X}$ to $\mathcal{Y}$
- Let $\eta(X) = \mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X)$
- Let $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be an i.i.d. sample with $(X_i, Y_i) \sim P$
- Consider the 0-1 cost

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

Table of Contents

Problem 2.2

In this part of the problem, assume that the joint distribution P between features and output is a *zero-error* distribution: so that $\eta(X) \in \{0, 1\}$ 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?

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

Problem 2.2

(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:

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

(d) Deduce that

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

Problem 2.2

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

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

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

(f) Show that

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

Hint: use question 2.a.

Problem 2.2

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

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

(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? Recall, that the inequality derived in the lecture was:

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

Table of Contents

Problem 2.3

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

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

(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?

Problem 2.3

The *margin condition*:

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

(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:

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

(c) Then show that

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

Problem 2.3

(d) Deduce that:

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

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

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

Problem 2.3

(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 $h \to 0$.

The inequalities are restated here for convenience:

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq 2\sqrt{\mathbb{E}\left[(\hat{\eta}(D_n;X) - \eta(X))^2 | D_n\right]}, \quad \text{no ass. (lecture): } h = 0$$

$$\ell\left(f^*, \hat{f}_{\hat{\eta}}(D_n)\right) \leq 4\mathbb{E}\left[\left(\hat{\eta}(D_n;X) - \eta(X)\right)^2 \middle| D_n\right], \quad \text{zero-error ass: } h = 1/2$$

$$\ell(f^*, \hat{f}_{\hat{\eta}}(D_n)) \leq \frac{2}{h}\mathbb{E}\left[\left(\hat{\eta}(D_n;X) - \eta(X)\right)^2 \middle| D_n\right], \quad \text{margin cond: } h \in (0, 1/2)$$

Problem 2.3

(g) It is said that the learning rule $\hat{f}_{\hat{\eta}}$ is weakly consistent if

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

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

Hint: For (ii), use Jensen's inequality.

Compare the *rate of convergence*, i.e. the speed at which $\mathbb{E}\Big[\ell\big(f^*, \hat{f}_{\hat{\eta}}(D_n)\big)\Big]$ tends to 0 with the sample size n in both cases.