Machine Learning
EDS
Tutorial: Week 4

Janneke van Brummelen

Vrije Universiteit Amsterdam

Period 3.4
2023/2024

During this week's lectures:

We have looked at the decomposition of excess risk into:

- Approximation error, discrepancy performance ideal predictor and theoretically best predictor in model,
- Estimation error, discrepancy between performance of sample-based predictor and theoretically best predictor in model.

We have established learning guarantees for:

- Zero-error classification for finite models,
- Non-deterministic setting with bounded cost and finite models.

Table of Contents

Quiz: Question 1

Let $\mathcal{X}$ and $\mathcal{Y}$ be some spaces of features and outputs, and let $S \subset \mathcal{F}$ be some model. Say we consider the empirical risk minimiser $\hat{f}_S$ over the model S for some sample D_n.

Is the following statement true or false?

"*The approximation error $\ell(f^*, S)$ will decrease as the sample size n increases*"

- True
- **False**

Quiz: Question 2

Is the following statement true or false?

"*If we consider a model $S \subset \mathcal{F}$ which contains a Bayes predictor, then the excess risk of the ERM predictor based on a sample D_n for this model S will be equal to zero.*"

- True
- **False**

This week's exercises

This week's problems focus on deriving learning guarantees:

- Problem 1: learning guarantee for expected estimation error for ERM with bounded cost
- Problem 2: learning guarantee for ERM with bounded cost variance

Table of Contents

Problem 1 (ERM with bounded cost)

Let all usual assumptions hold, e.g. let D_n be some sample of n independent pairs $(X_i, Y_i) \sim P$. Consider the regression framework $\mathcal{Y} = \mathbb{R}$ with a cost function c.

1.1 Recall the definition of the empirical risk $\widehat{\mathcal{R}}_n(f)$ of a predictor $f \in \mathcal{F}$.

1.2 Prove that $\mathbb{E}[\widehat{\mathcal{R}}_n(f)] = \mathcal{R}_P(f)$ for any predictor $f \in \mathcal{F}$.

1.3 Recall the definition of a model. When do we say that a model is finite? And when do we say a model is infinite? Give an example in each case.

1.4 Let S denote a model. Recall the definition of an ERM predictor over the model S.

1.5 Let $\hat{f}$ denote an ERM predictor over the model S and let $\mathcal{R}_P(\hat{f})$ denote its generalisation risk. Why do we say that $\hat{f}$ and $\mathcal{R}_P(\hat{f})$ are random variables?

Problem 1 (ERM with bounded cost)

1.6 Propose an interpretation (in one sentence) of the quantity $\inf_{f \in S} \mathcal{R}_P(f)$.
Using the fact that for a random variable Z, and a measurable function h, it holds that: $\inf_t \mathbb{E}\Big[h(Z,t)\Big] \geq \mathbb{E}\Big[\inf_t h(Z,t)\Big]$, show the following inequality:

$$\inf_{f \in S} \mathcal{R}_P(f) \geq \mathbb{E}\Big[\widehat{\mathcal{R}}_n(\hat{f})\Big].$$

1.7 Deduce the following upper-bound on the expectation of the estimation error:

$$\mathbb{E}\Big[\mathcal{R}_P(\hat{f})\Big] - \inf_{f \in S} \mathcal{R}_P(f) \leq \mathbb{E}\left[\sup_{f \in S} \Big\{\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)\Big\}\right].$$

Problem 1 (ERM with bounded cost)

Assume in the rest of the exercise that the cost function c is bounded: there exists a positive constant C such that for any $y, y' \in \mathbb{R}$, $0 \leq c(y, y') \leq C$. Assume in addition that the model S is finite and denote $K = \operatorname{Card} S$, $S = \{f_1, \ldots, f_K\}$.

1.8 Using results from Tutorial 1, show that $U_i := \mathbb{E}\left[\frac{1}{n}c\big(f(X), Y\big)\right] - \frac{1}{n}c\big(f(X_i), Y_i\big)$, $i = 1, \ldots, n$, are i.i.d. b-sub-gaussian random variable, for some explicit parameter b that you will express in terms of C and n.

Result in Problem 3 Week 1:

If Z is a random variable that is bounded within $[c, d]$, then $Z - \mathbb{E}[Z]$ and $\mathbb{E}[Z] - Z$ are sub-Gaussian with parameter $\frac{d-c}{2}$.

Problem 1 (learning guarantee ERM with bounded cost)

1.9 Deduce that for each $f \in \mathcal{F}$, $\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)$ is $\bar{b}$-sub-gaussian for some explicit parameter $\bar{b}$ that you will express in terms of C and n.

Result in Problem 2 Week 1:

If $Z_1, \dots, Z_n$ are n independent b_i-sub-Gaussian random variables, for $i \in \{1, \dots, n\}$, then $S_n = \sum_{i=1}^n Z_i$ is $\bar{b}$-sub-Gaussian with parameter $\bar{b} = \sqrt{\sum_{i=1}^n b_i^2}$.

Problem 1 (learning guarantee ERM with bounded cost)

1.10 Finally, using the lemma proved in Problem 4 of Week 1 Problem set, prove the following learning guarantee for the ERM predictor:

$$\forall n \geq 1, \quad \mathbb{E}\left[\mathcal{R}_P(\hat{f})\right] \leq \inf_{f \in S} \mathcal{R}_P(f) + C\sqrt{\frac{\ln(\text{Card } S)}{2n}}.$$

Lemma of Problem 4 Week 1:

Let $Z_1, \ldots, Z_K$ be ν-sub-Gaussian random variables for some parameter $\nu > 0$. Then,

$$\mathbb{E}\left[\max_{1 \leq k \leq K} Z_k\right] \leq \nu\sqrt{2\ln(K)}.$$

1.11 Interpret the learning guarantee of Question 10. Comment in particular on the influence of the sample size, model complexity, and maximal cost C.

Table of Contents

Problem 2 (ERM with bounded cost variance)

In the regression framework $\mathcal{Y} = \mathbb{R}$, let c be a cost function with bounded variance in the sense that $\mathbb{V}\Big(c\big(f(X), Y\big)\Big) \leq v$ for some $v > 0$, S be a finite model and $\hat{f}$ be an ERM learning over S.

2.1 Is the bounded variance cost function assumption more or less restrictive than assuming the cost function is bounded? Give an example of a context without a bounded cost function where the bounded cost variance assumption is reasonable, and another example where the bounded cost function is reasonable.

Problem 2 (ERM with bounded cost variance)

2.2 Prove the inequality:

$$\mathcal{R}_P(\hat{f}) - \inf_{f \in S} \mathcal{R}_P(f) \leq 2 \sup_{f \in S} \left| \widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f) \right|,$$

What does the left-hand side represent?

2.3 Show that for any $\varepsilon > 0$:

$$\mathbb{P}\left(\mathcal{R}_P(\hat{f}) - \inf_{f \in S} \mathcal{R}_P(f) \geq 2\varepsilon \right) \leq \sum_{f \in S} \mathbb{P}\left(\left| \widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f) \right| \geq \varepsilon \right).$$

2.4 Show that

$$\mathbb{E}\left[\left(\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f) \right)^2 \right] \leq \frac{v}{n}.$$

Problem 2 (ERM with bounded cost variance)

2.5 Deduce from question 4 that

$$\mathbb{P}\left(\left|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\right| \geq \varepsilon\right) \leq \frac{v}{n\varepsilon^2}.$$

2.6 Finally, using questions 3 and 5, prove the learning guarantee

$$\mathbb{P}\left(\mathcal{R}_P(\hat{f}) \leq \inf_{f \in S} \mathcal{R}_P(f) + 2\sqrt{\frac{v \,\text{Card}\, S}{n\delta}}\right) \geq 1 - \delta,$$

for any $n \geq 1$ and $\delta > 0$.

Problem 2 (ERM with bounded cost variance)

2.7 Interpret. How does it compare to the learning guarantee obtained the lecture in the bounded cost function case? Comment in particular on the influence of the sample size, model complexity, and confidence level.

- For bounded cost we derived in lecture:

$$\mathbb{P}\left(\mathcal{R}_P(\hat{f}_S) \leq \inf_{f \in S} \mathcal{R}_P(f) + C\sqrt{\frac{2\ln\frac{2}{\delta} + 2\ln\left(\text{Card } S\right)}{n}}\right) \geq 1 - \delta\,.$$

- For cost with bounded variance we just derived:

$$\mathbb{P}\left(\mathcal{R}_P(\hat{f}) \leq \inf_{f \in S} \mathcal{R}_P(f) + 2\sqrt{\frac{v \text{ Card } S}{n\delta}}\right) \geq 1 - \delta.$$