Machine Learning for EDS

Tutorial week 4

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 (learning guarantee ERM with bounded cost) Consider the regression framework $\mathcal{Y} = \mathbb{R}$ with a cost function c.

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

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

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.

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

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?

   (**Note:** in this problem set we write $\hat{f}$ for notational convenience, but we actually mean $\hat{f}(D_n)$. So here $\hat{f}$ denotes the ERM predictor over the model S for the sample D_n.)

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:

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

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

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

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 $K = \operatorname{Card} S$, $S = \{f_1, \ldots, f_K\}$.

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

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.

10. Finally, using the lemma proved in Problem 4 of Week 1-2 Tutorial, prove the following learning guarantee for the ERM predictor:

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

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

Problem 2 (learning guarantee 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.

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.

2. Prove the inequality:

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

What does the left-hand side represent?

3. Show that for any $\varepsilon > 0$:

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

4. Show that

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

5. Deduce from question 4 that

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

6. Finally, using questions 3 and 5, prove the learning guarantee

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

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

7. Interpret. How does it compare to the learning guarantee obtained the lecture in the bounded cost function case (slide *Estimation error upper-bound for ERM with finite models and bounded cost function*)? Comment in particular on the influence of the sample size, model complexity, and confidence level.