Machine Learning for EDS

TUTORIAL WEEK 4 : SOLUTIONS

2023/2024

Important: this file is meant for students of the course Machine Learning for EDS (2023/2024) and is not allowed to be distributed to others.

Generically in the following exercises, consider $\mathcal{X}$ to be a measurable space of features, $\mathcal{Y}$ a measurable space of outputs, P a distribution over $\mathcal{X} \times \mathcal{Y}$, $(X, Y) \sim P$, $\eta(X) = \mathbb{E}[Y|X]$ and $\mathcal{F}$ is the set of all predictors (that is, measurable functions) from $\mathcal{X}$ into $\mathcal{Y}$.

Problem 1

1. Empirical risk is defined as follows:

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} c(f(X_i), Y_i)\,.$$

2. Using the definition above, for any $f \in \mathcal{F}$:

$$\begin{aligned}\mathbb{E}[\widehat{\mathcal{R}}_n(f)] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} c(f(X_i), Y_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[c(f(X_i), Y_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[c(f(X), Y)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathcal{R}_P(f) = \mathcal{R}_P(f)\,,\end{aligned}$$

   where the second equality uses the linearity of the expectation, and the third equality follows from the fact that (X, Y) has the same distribution as (X_i, Y_i). So the empirical risk is an unbiased estimator of the true risk.

3. A model S is a set of predictors, $S \subset \mathcal{F}$. A model is finite if it contains a finite number of predictors, so if $\mathrm{Card}(S) < \infty$. A model is infinite if it contains infinitely many predictors. For examples see the lecture slides of week 4.

4. The Empirical Risk Minimization (ERM) predictor $\hat{f}_S$ minimizes the empirical risk over S:

$$\hat{f}_S \in S \quad \text{and} \quad \widehat{\mathcal{R}}_n(\hat{f}_S; D_n) = \inf_{g \in S}\widehat{\mathcal{R}}_n(g; D_n)\,.$$

5. Recall that the generalisation error of $\hat{f}$ is equal to:

$$\mathcal{R}_P(\hat{f}) = \mathbb{E}[c(\hat{f}(X), Y)|D_n]\,.$$

   So clearly $\hat{f}$ and $\mathcal{R}_P(\hat{f})$ are random variables, because they both depend on the sample D_n which is random.

6. $\inf_{f \in S} \mathcal{R}_P(f)$ is the minimal generalisation error that can be achieved by the predictors in the model S (or strictly speaking, it is the highest lower bound of the generalisation error that can be achieved by the predictors in the model S).

$$\inf_{f \in S} \mathcal{R}_P(f) \underbrace{=}_{\text{by 2.}} \inf_{f \in S} \mathbb{E}[\widehat{\mathcal{R}}_n(f)] \qquad \underbrace{\geq}_{\text{given (see remark)}} \qquad \mathbb{E}[\inf_{f \in S} \widehat{\mathcal{R}}_n(f)] \underbrace{=}_{\text{by 4.}} \mathbb{E}[\widehat{\mathcal{R}}_n(\hat{f})]$$

Remark: [Super-additivity of infimum and expectation] For a random variable X, and an appropriately defined measurable function h, it holds that:

$$\inf_t \mathbb{E}\Big[h(X,t)\Big] \geq \mathbb{E}\Big[\inf_t h(X,t)\Big].$$

To see this, notice first that for any s: $h(X,s) \geq \inf_t h(X,t)$. Taking expectations, we get that for any s:

$$\mathbb{E}\Big[h(X,s)\Big] \geq \mathbb{E}\Big[\inf_t h(X,t)\Big],$$

and taking the infimum over s:

$$\inf_s \mathbb{E}\Big[h(X,s)\Big] \geq \mathbb{E}\Big[\inf_t h(X,t)\Big].$$

7. From point 6 we know that $-\inf_{f \in S} \mathcal{R}_P(f) \leq -\mathbb{E}[\widehat{\mathcal{R}}_n(\hat{f})]$, thus:

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

8. To show this, see that $\frac{1}{n}c(f(X_i), Y_i)$ is a random variable bounded within $[0, C/n]$. In the tutorial of week 1 (problem 3), we have seen that if Z is a random variable bounded within $[c, d]$, then both $Z - \mathbb{E}[Z]$ and $\mathbb{E}[Z] - Z$ are sub-Gaussian with parameter $\frac{d-c}{2}$. Hence, $U_i = \mathbb{E}[\frac{1}{n}c(f(X), Y)] - \frac{1}{n}c(f(X_i), Y_i)$ is sub-Gaussian with parameter $C/(2n)$ for all i.

9. In problem 2 of week 1, we have shown that the sum of n independent sub-Gaussian random variables $Z_1, \ldots, Z_n$, each with respective parameter b_i, $i \in \{1, \ldots, n\}$, is itself sub-Gaussian with parameter $b = \sqrt{\sum_{i=1}^n b_i^2}$. Hence for each $f \in \mathcal{F}$:

$$\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) = \sum_{i=1}^n \left(\mathbb{E}\Big[\frac{1}{n}c(f(X_i), Y_i)\Big] - \frac{1}{n}c(f(X_i), Y_i)\right),$$

is sub-Gaussian with parameter $\sqrt{\sum_{i=1}^n \frac{C^2}{4n^2}} = \frac{C}{2\sqrt{n}}$

10. From point 7, we know that

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

Let's upper-bound the expectation on the right-hand side. Recall from the previous point that for a fixed f, $\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)$ is b-sub-Gaussian for $b = \dfrac{C}{2\sqrt{n}}$. Since the model S is finite, we can denote it as $S = \{f_1, \dots, f_K\}$, with $K = \mathrm{Card} S$. So we can apply the lemma from problem 4 of Tutorial 1 with $Z_k = \mathcal{R}_P(f_k) - \widehat{\mathcal{R}}_n(f_k)$, $\nu = \dfrac{C}{2\sqrt{n}}$:

$$\begin{aligned}\mathbb{E}\Big[\sup_{f\in S}\{\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)\}\Big] &= \mathbb{E}\Big[\max_{k\in\{1,\dots,K\}}\{\mathcal{R}_P(f_k) - \widehat{\mathcal{R}}_n(f_k)\}\Big]\\ &= \mathbb{E}\Big[\max_{k\in\{1,\dots,K\}} Z_k\Big]\\ &\leq \frac{C}{2\sqrt{n}}\sqrt{2\ln(K)}\\ &= \frac{C}{2\sqrt{n}}\sqrt{2\ln(\mathrm{Card}S)}.\end{aligned}$$

Therefore:

$$\boxed{\begin{array}{l}\mathbb{E}\Big[\mathcal{R}_P(\hat{f})\Big] \leq \inf_{f\in S} \mathcal{R}_P(f) + C\sqrt{\dfrac{\ln(\mathrm{Card}S)}{2n}},\\ \text{or equivalently}\\ \mathbb{E}\Big[\ell(f^*, \hat{f})\Big] \leq \ell(f^*, S) + C\sqrt{\dfrac{\ln(\mathrm{Card}S)}{2n}}.\end{array}}$$

11. The learning guarantee gives an upperbound to the difference between the expectation of the generalization error of the ERM and the minimal generalization error that can be attained by the predictors in S. As the sample size n increases, the learning guarantees becomes stronger, i.e. the bound decreases, and moves towards the lower bound $\inf_{f\in S}\mathcal{R}_P(f)$ at a $n^{-1/2}$ rate. This is intuitive, because if there is more data available for constructing the ERM, then the empirical risk is likely to be more similar to the true risk. As the model size increases, i.e. if $K = \mathrm{Card}S$ increases, the learning guarantee becomes weaker, but only at a $(\ln(K))^{1/2}$ rate. This is intuitive, because if $\mathrm{Card}S$ increases, there will be more predictors to choose from, which makes it less likely that the ERM is close to $\inf_{f\in S}\mathcal{R}_P(f)$. Finally, if the maximal cost C increases, the learning guarantee becomes weaker, because then the cost function has a different scale and for example $\mathcal{R}_P(f)$ is then also likely to be higher.

Problem 2

1. It is less restrictive, because if the cost function is bounded, it follows immediately that the variance of the cost function is also bounded. On the other hand, the cost function having a bounded variance does not imply a bounded cost function. Think of examples yourself. In general, the bounded cost assumption requires $\mathcal{Y}$ to be bounded or, if $\mathcal{Y}$ is unbounded (e.g. if $\mathcal{Y} = \mathbb{R}$), the cost function should have some imposed upperbound (e.g. a truncated cost function). In case the bounded cost function assumption is violated, the bounded cost variance assumption can still hold, but it requires restrictions on P and possibly on the model S. For instance, for quadratic cost, the assumptions $\mathbb{E}[Y^2] < \infty$ and $\mathbb{E}[f(X)^2] < \infty$ for any $f \in S$ would be sufficient for the bounded cost variance assumption to hold.

2. This was shown in the lectures. For completeness we will repeat the derivation here. For $\hat{f}$ the ERM over S and for any predictor $f \in S$, we have:

$$\begin{aligned}\mathcal{R}_P(\hat{f}) - \mathcal{R}_P(f) &= \underbrace{\mathcal{R}_P(\hat{f}) - \widehat{\mathcal{R}}_n(\hat{f})}_{\leq \sup_f |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|} + \underbrace{\widehat{\mathcal{R}}_n(\hat{f}) - \widehat{\mathcal{R}}_n(f)}_{\leq 0} + \underbrace{\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)}_{\leq \sup_f |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|} \\ &\leq 2 \sup_f |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|\end{aligned}$$

By taking the supremum over $f \in S$ on both sides of the inequality, we get the result:

$$\sup_{f \in S} [\mathcal{R}_P(\hat{f}) - \mathcal{R}_P(f)] = \mathcal{R}_P(\hat{f}) - \inf_{f \in S} \mathcal{R}_P(f) \leq 2 \sup_f |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|$$

The left-hand side represents the estimation error.

3. Using the global upper-bound from point 2, we have that for any $\varepsilon > 0$:

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

Using a union bound:

$$\begin{aligned}\mathbb{P}\Big(\sup_{f \in S} \Big|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\Big| \geq \varepsilon\Big) &= \mathbb{P}\Big(\bigcup_{f \in S} \Big\{\Big|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\Big| \geq \varepsilon\Big\}\Big) \\ &\leq \sum_{f \in S} \mathbb{P}\Big(\Big|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\Big| \geq \varepsilon\Big).\end{aligned}$$

4. Since the empirical risk is an unbiased estimator of the true risk (i.e., $\mathbb{E}\Big[\widehat{\mathcal{R}}_n(f)\Big] = \mathcal{R}_P(f)$), we notice that

$$\mathbb{E}\Big[\Big(\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\Big)^2\Big] = \mathbb{E}\Big[\Big(\widehat{\mathcal{R}}_n(f) - \mathbb{E}[\widehat{\mathcal{R}}_n(f)]\Big)^2\Big]$$

$$
\begin{aligned}
&= \mathbb{V}\Big(\widehat{\mathcal{R}}_n(f)\Big) = \mathbb{V}\Big(\frac{1}{n}\sum_{i=1}^{n} c(f(X_i), Y_i)\Big) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}\Big(c(f(X_i), Y_i)\Big) \le \frac{v}{n},
\end{aligned}
$$

where we used the independence between the examples $(X_1, Y_1), \dots, (X_n, Y_n)$ and the upper bound of the variance of the cost function v.

5. Apply the Markov inequality to the random variable $\Big|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\Big|$ for some fixed $f \in S$:

$$
\begin{aligned}
\mathbb{P}\Big(\Big|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\Big| \ge \varepsilon\Big) &= \mathbb{P}\Big(\Big(\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\Big)^2 \ge \varepsilon^2\Big) \\
&\le \frac{\mathbb{E}\Big[\Big(\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\Big)^2\Big]}{\varepsilon^2} \le \frac{v}{n\varepsilon^2}.
\end{aligned}
$$

6. Using the results from question 3 and question 5, we obtain:

$$
\begin{aligned}
\mathbb{P}\Big(\mathcal{R}_P(\hat{f}) - \inf_{f\in S}\mathcal{R}_P(f) \ge 2\varepsilon\Big) &\le \sum_{f\in S}\mathbb{P}\Big(\Big|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\Big| \ge \varepsilon\Big) \\
&\le \sum_{f\in S}\frac{v}{n\varepsilon^2} = \frac{v\mathrm{Card}S}{n\varepsilon^2}.
\end{aligned}
$$

So to obtain the final expression, rewrite by multiplying the inequality by -1 and then adding 1:

$$
\mathbb{P}\Big(\mathcal{R}_P(\hat{f}) \le \inf_{f\in S}\mathcal{R}_P(f) + 2\varepsilon\Big) \ge 1 - \frac{v\mathrm{Card}S}{n\varepsilon^2}.
$$

Now rewrite the result with $\delta := \dfrac{v\mathrm{Card}S}{n\varepsilon^2}$ (implying $\varepsilon = \sqrt{v\mathrm{Card}S/(n\delta)}$), to get the final expression:

$$
\boxed{\begin{array}{l}\text{For any } \delta > 0, n \ge 1, \\ \mathbb{P}\Big(\mathcal{R}_P(\hat{f}) \le \inf_{f\in S}\mathcal{R}_P(f) + 2\sqrt{\dfrac{v\mathrm{Card}S}{n\delta}}\Big) \ge 1 - \delta.\end{array}}
$$

7. The interpretation of the learning guarantee is that when drawing a sample D_n at random, the estimation error of the ERM predictor based on D_n is bounded by $2\sqrt{v\mathrm{Card}S/n\delta}$ with probability of at least $1-\delta$.

In the case the cost function c was bounded, $c(y, y') \le C$ for all $y, y' \in \mathbb{R}$, for some constant $C > 0$, we obtained the bound:

$$
\mathbb{P}\left(\mathcal{R}_P(\hat{f}) \le \inf_{f\in S}\mathcal{R}_P(f) + C\sqrt{\frac{2\ln(\frac{1}{\delta}) + 2\ln\left(2\mathrm{Card}S\right)}{n}}\right) \ge 1 - \delta.
$$

The bounded cost case offers a much stronger guarantee: the complexity of the model intervenes through $\sqrt{\ln(\mathrm{Card} S)}$ instead of $\sqrt{\mathrm{Card} S}$, while the confidence δ intervenes through $\sqrt{\ln(1/\delta)}$ instead of $\sqrt{1/\delta}$. The influence of the sample size is comparable (in both cases sample size intervenes through $\sqrt{1/n}$)