Exercice Sheet 2 – Fundamentals of supervised learning

Mathematics for Machine Learning

6 October 2025

Exercise 1. (The approximation error decreases with the size of the model.) Let $\mathcal{F}, \mathcal{F}'$ be two models such that $\mathcal{F}' \subset \mathcal{F}$. Prove that for a given learning task, the approximation error of $\mathcal{F}'$ is larger than the approximation error of $\mathcal{F}$.

Exercise 2. (Lagrange interpolation polynomials.) Let $x_0, \dots, x_d \in \mathbb{R}$ be pairwise distinct. The aim of this exercise is to prove that, for any $y_0, \dots, y_d \in \mathbb{R}$, there exists a unique polynomial P of degree $\leq d$ such that

$$\forall i \in \{0, 1, \dots, d\}, \quad P(x_i) = y_i. \tag{1}$$

1. Assume first that $d = 2$. We are going to construct polynomials Q_0, Q_1, Q_2 of degree 2, such that
$$\forall i \in \{0, 1, 2\}, \quad Q_i(x_i) = y_i \quad \text{and} \quad Q_i(x_j) = 0 \quad \text{if } j \neq i$$
   (a) Assume that we have constructed such Q_0, Q_1, Q_2. Prove that $P = Q_0 + Q_1 + Q_2$ satisfies the condition (1).

   (b) Let $R_0(X) = (X - x_1)(X - x_2)$. We look for Q_0 in the form $a_0 R_0$ for $a_0 \in \mathbb{R}$. Which value should we choose for a_0? Same question for Q_1 and Q_2.

2. Let us go back to the general case. How to construct a polynomial P of degree d satisfying the constraint (1)?

3. Prove that such a polynomial is unique. *Hint: a polynomial of degree d that has at least $d + 1$ roots has to be zero.*

Exercise 3. (About polynomial regression.) Consider a regression problem with training data $(x_0, y_0), \dots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}$. Assume that the x_i are pairwise distinct. We do empirical risk minimization with the quadratic cost, choosing the model $\mathcal{F}_d$ of polynomial functions of degree at most d:

$$\mathcal{F}_d = \{x \mapsto a_d x^d + a_{d-1} x^{d-1} + \dots + a_1 x + a_0 : (a_0, \dots, a_d) \in \mathbb{R}^{d+1}\}.$$

1. (Noiseless case.) Assume that for all $i \in \{0, \dots, n\}$, we have $y_i = T(x_i)$, where T is a polynomial of degree $\delta \leq d$.

   (a) Assume $n < d$. Is there a unique empirical risk minimizer for this problem? What is the value of the empirical risk of an ERM?

   (b) Assume $n \geq d$. Is there a unique empirical risk minimizer for this problem? Does it minimize the true risk?

2. (Noisy case.) Assume that for all $i \in \{0, \dots, n\}$, we have $y_i = T(x_i) + \varepsilon_i$, where T is a polynomial of degree $\delta \leq d$, and $\varepsilon_i \in \mathbb{R}$. When $n = d$, is the ERM equal to T? Close to T?

Exercise 4. (True risk minimizer vs. empirical risk minimizer.) Consider a regression problem with training data $(x_1, y_1), \ldots, (x_n, y_n) \in [0,1] \times \mathbb{R}$. Assume that for all $i \in \{1, \ldots, n\}$, we have $y_i = T(x_i)$ where $T(X) = X^2 + 1$. We consider a linear model $\mathcal{F} = \{x \mapsto ax : a \in \mathbb{R}\}$. With a slight abuse of notation, we will denote as a the function $x \mapsto ax$.

1. Let us determine the minimizer of the true risk inside $\mathcal{F}$.

   (a) Assuming that the x_i are uniformly distributed on $[0,1]$, prove that the true risk of a predictor $a \in \mathcal{F}$ for the quadratic cost reads

   $$R(a) = \frac{1}{2}\int_0^1 (ax - x^2 - 1)^2 \mathrm{d}x.$$

   (b) Which $a^* \in \mathbb{R}$ minimizes the true risk $R(a)$?

2. Let us now determine the empirical risk minimizer inside $\mathcal{F}$. Assume that for all $i \in \{1, \ldots, n\}$, we have $x_i = i/n$. We have seen in the lecture that the empirical risk minimizer is

   $$\hat{a} = \frac{\sum_{i=1}^n x_i T(x_i)}{\sum_{i=1}^n x_i^2}.$$

   Compute its value as a function of n. What is $\lim_{n \to +\infty} \hat{a}$? *Hint: you can use that* $\sum_{i=1}^n i = n(n+1)/2$, $\sum_{i=1}^n i^2 = n(n+1)(2n+1)/6$ *and* $\sum_{i=1}^n i^3 = n^2(n+1)^2/4$.

Exercise 5. Consider a binary classifier that outputs randomly 0 or 1 with the same probability 0.5. Take a test set with 85% of labels 1, and 15% of labels 0. What are the recall, the precision and the accuracy of this classifier?

Solutions

Exercise 1. Let us prove the following more general result: if $\mathcal{F}$ and $\mathcal{F}'$ are two sets such that $\mathcal{F}' \subset \mathcal{F}$, then for any function $\varphi \colon \mathcal{F} \to \mathbb{R}$, we have

$$\min_{x \in \mathcal{F}'} \varphi(x) \geq \min_{x \in \mathcal{F}} \varphi(x).$$

Denote

$$m := \min_{x \in \mathcal{F}} \varphi(x), \qquad m' := \min_{x \in \mathcal{F}'} \varphi(x).$$

Since $\mathcal{F}' \subset \mathcal{F}$, every $x \in \mathcal{F}'$ is also in $\mathcal{F}$, hence for every $x \in \mathcal{F}'$ we have $\varphi(x) \geq m$. Taking the minimum over $x \in \mathcal{F}'$ yields $m' \geq m$. This proves that the minimum of φ over the smaller model $\mathcal{F}'$ cannot be smaller than the minimum over the larger model $\mathcal{F}$. Applied to approximation error (which is of the form $\min_{f \in \mathcal{F}} R(f) - c$ where c is a constant, equal to the risk of the best predictor for the considered task), this gives the desired result: if $\mathcal{F}' \subset \mathcal{F}$ then the approximation error of $\mathcal{F}'$ is greater than or equal to that of $\mathcal{F}$.

Exercise 2. (Lagrange interpolation polynomials.)

1. Case $d = 2$ (three points). Suppose we have constructed polynomials Q_0, Q_1, Q_2 of degree ≤ 2 such that

   $$Q_i(x_i) = y_i \quad \text{and} \quad Q_i(x_j) = 0 \quad (j \neq i).$$

(a) Then $P := Q_0 + Q_1 + Q_2$ is of degree at most 2, and we have for each i

$$P(x_i) = Q_0(x_i) + Q_1(x_i) + Q_2(x_i) = 0 + \cdots + y_i + \cdots + 0 = y_i,$$

so P satisfies the interpolation conditions.

(b) To build Q_0 take $R_0(X) = (X - x_1)(X - x_2)$ (degree 2). Then $R_0(x_1) = R_0(x_2) = 0$ and $R_0(x_0) \neq 0$ (because the x_i are distinct). Hence choose

$$Q_0(X) = a_0 R_0(X) \quad \text{with} \quad a_0 = \frac{y_0}{R_0(x_0)} = \frac{y_0}{(x_0 - x_1)(x_0 - x_2)}.$$

This gives

$$Q_0(X) = y_0 \frac{X - x_1}{x_0 - x_1} \frac{X - x_2}{x_0 - x_2}.$$

Analogously,

$$Q_1(X) = y_1 \frac{X - x_0}{x_1 - x_0} \frac{X - x_2}{x_1 - x_2}, \qquad Q_2(X) = y_2 \frac{X - x_0}{x_2 - x_0} \frac{X - x_1}{x_2 - x_1}.$$

2. General d. For general d define the following polynomials for $i = 0, \ldots, d$:

$$L_i(X) := \prod_{\substack{0 \leq j \leq d \\ j \neq i}} \frac{X - x_j}{x_i - x_j}.$$

Each L_i is a polynomial of degree d and satisfies $L_i(x_j) = \mathbf{1}_{i=j}$ (indicator function). Then the interpolating polynomial is

$$P(X) := \sum_{i=0}^{d} y_i \, L_i(X).$$

Clearly $P(x_k) = \sum_i y_i L_i(x_k) = y_k$ for every k.

3. Uniqueness. Suppose P and Q are two polynomials of degree $\leq d$ satisfying $P(x_i) = Q(x_i) = y_i$ for all $i = 0, \ldots, d$. Then the polynomial $H := P - Q$ has degree $\leq d$ and has zeros at each $x_0, \ldots, x_d$ (so it has at least $d + 1$ distinct roots). A nonzero polynomial of degree at most d can have at most d distinct roots, hence H must be the zero polynomial. Therefore $P = Q$ and the interpolant is unique.

Exercise 3. (About polynomial regression.)

We have training points $(x_0, y_0), \ldots, (x_n, y_n)$ with distinct x_i. The model $\mathcal{F}_d$ is polynomials of degree $\leq d$.

1. Noiseless case: $y_i = T(x_i)$ for some polynomial T of degree $\delta \leq d$.

   (a) If $n < d$ (fewer data points than parameters), then there are infinitely many polynomials in $\mathcal{F}_d$ that interpolate the $n+1$ points. Indeed, for any $(x_{n+1}, y_{n+1}), \ldots, (x_d, y_d)$, there exists a polynomial P such that $P(x_i) = y_i$ for all $i \in \{0, \ldots, d\}$, according to Exercise 2. Such a polynomial is an ERM with a zero empirical risk. As there are infinitely many choices for the additional constraints $(x_{n+1}, y_{n+1}), \ldots, (x_d, y_d)$, there are infinitely many ERMs, with zero empirical risk.

(b) If $n \geq d$, using the argument of Exercise 2, question 3, there is at most one polynomial P in $\mathcal{F}_d$ that interpolates the (x_i, y_i), i.e. such that $P(x_i) = y_i$ for all $i \in \{0, \dots, n\}$. Moreover, the polynomial T is in $\mathcal{F}_d$, because $\delta \leq d$ by assumption, and T interpolates the x_i by definition. Therefore, T is the unique ERM, and it has a zero empirical risk. It also has a zero (so minimal) true risk.

2. Noisy case: $y_i = T(x_i) + \varepsilon_i$ with nonzero noise. When $n = d$, the ERM is the unique polynomial of degree $\leq d$ interpolating the noisy values $(x_i, T(x_i) + \varepsilon_i)$. Unless all $\varepsilon_i = 0$, this interpolant is not equal to T. Even when the noise is small, this interpolant is not necessarily close to T.

Exercise 4. (True risk minimizer vs. empirical risk minimizer.)

1. (a) Let X be a uniform random variable on $[0, 1]$. By definition

$$\begin{aligned} R(a) &= \mathbb{E}[c(aX, T(X))] \\ &= \mathbb{E}[\frac{1}{2}(aX - T(X))^2] \\ &= \frac{1}{2}\int_0^1 (ax - x^2 - 1)^2 \mathrm{d}x. \end{aligned}$$

(b) $R(a)$ is a convex degree-2 polynomial in a. We have

$$\begin{aligned} R(a) &= \frac{1}{2}\int_0^1 (ax - x^2 - 1)^2 \mathrm{d}x \\ &= \frac{1}{2}(\int_0^1 x^2 \mathrm{d}x) a^2 - \left(\int_0^1 x(x^2+1)\mathrm{d}x\right) a + \text{cst} \\ &= \frac{1}{2}[x^3/3]_0^1 \; a^2 - [x^4/4 + x^2/2]_0^1 \; a + \text{cst} \\ &= \frac{1}{6}a^2 - \frac{3}{4}a + \text{cst} \end{aligned}$$

where cst does not depend on a. The true risk is minimized at a^* such that $R'(a^*) = 0 \Leftrightarrow \frac{1}{3}a^* = \frac{3}{4} \Leftrightarrow a^* = \frac{9}{4}$.

2. Let us first compute the denominator.

$$\begin{aligned} \sum_{i=1}^n x_i^2 &= \frac{1}{n^2}\sum_{i=1}^n i^2 \\ &= \frac{n(n+1)(2n+1)}{6n^2} \\ &= \frac{(n+1)(2n+1)}{6n} \\ &= \frac{2n^2 + 3n + 1}{6n}. \end{aligned}$$

Now the numerator:

$$\begin{aligned}
\sum_{i=1}^{n} x_i T(x_i) &= \sum_{i=1}^{n} \frac{i}{n}\left(\frac{i^2}{n^2}+1\right) \\
&= \frac{1}{n^3}\sum_{i=1}^{n} i^3 + \frac{1}{n}\sum_{i=1}^{n} i \\
&= \frac{n^2(n+1)^2}{4n^3} + \frac{n(n+1)}{2n} \\
&= \frac{(n+1)^2}{4n} + \frac{n+1}{2} \\
&= \frac{n^2+2n+1+2n(n+1)}{4n} \\
&= \frac{3n^2+4n+1}{4n}.
\end{aligned}$$

Finally,

$$\begin{aligned}
\hat{a} &= \frac{3n^2+4n+1}{4n} \times \frac{6n}{2n^2+3n+1} \\
&= \frac{9n^2+12n+3}{4n^2+6n+2} \\
&\to_{n\to+\infty} \frac{9}{4}.
\end{aligned}$$

Therefore, when $n \to +\infty$, the empirical risk minimizer of linear regression for the quadratic cost converges to the true risk minimizer.

Exercise 5. We compute

$$\mathrm{TP} = 0.425, \quad \mathrm{FP} = 0.075, \quad \mathrm{FN} = 0.425, \quad \mathrm{TN} = 0.075.$$

Hence

$$\text{Precision} = \frac{0.425}{0.5} = 0.85, \quad \text{Recall} = \frac{0.425}{0.85} = 0.5, \quad \text{Accuracy} = 0.425 + 0.075 = 0.5.$$