Midterm exam

Mathematics for Machine Learning

3 November 2025

The use of any external aids (lecture notes, internet connection, etc.) is prohibited.

Questions on the course. (8 pts)

1. Define the estimation error and the approximation error of a model $\mathcal{F}$. How do they change when the complexity of the model increases? (1 pt)

2. Does an empirical risk minimizer always have good generalization properties? Give an example. (1 pt)

3. What is the F1-score of a classifier on a test set? (0.5 pt)

4. Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}$. Compute $\operatorname{argmin}_{a \in \mathbb{R}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - a x_i)^2$. (1 pt)

5. Write the function $f\colon \mathbb{R}^p \to \mathbb{R}$ encoded by a 1-layer perceptron with parameters $w_1, \ldots, w_p, b$ and activation function $\sigma\colon \mathbb{R} \to \mathbb{R}$. When σ is ReLU and $p = 1$, what does the graph of this function look like? (1 pt)

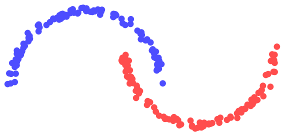
6. When using the k-NN algorithm, k should not be too small, nor too large. Why (for each case)? What does the algorithm predict when $k = n$? (1 pt)

7. We want to solve a regression problem with the k-NN algorithm. To chose the value of k, we perform 3-fold cross-validation on k. Explain the procedure. (1.5 pts)

8. What is the Hamming distance between the sets

$$E_1 = \{1, 2, 5\} \quad \text{and} \quad E_2 = \{1, 2, 3, 4\}? \quad (0.5 \text{ pt})$$

9. Reproducing this figure on your answer sheet, draw the decision frontier of the nearest neighbor algorithm (i.e., $k = 1$) on the following classification problem: (0.5 pt)



Exercise 1. (true risk minimizer vs empirical risk minimizer, 4 pts) Consider a regression problem with training data $(x_1, y_1), \dots, (x_n, y_n) \in [0,1] \times \mathbb{R}$. Assume that for all $i \in \{1, \dots, n\}$, we have $y_i = T(x_i)$ where $T(X) = 4X^2 + 3X$. 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. Assuming that the x_i are uniformly distributed on $[0,1]$, the true risk of a predictor $a \in \mathcal{F}$ for the quadratic cost reads

   $$R(a) = \frac{1}{2} \int_0^1 (ax - (4x^2 + 3x))^2 \mathrm{d}x.$$

   Determine the minimizer $a^* \in \mathbb{R}$ of the true risk $R(a)$. (1 pt)

2. Prove (for example by recurrence) that

   $$\sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6 \quad \text{and} \quad \sum_{i=1}^{n} i^3 = n^2(n+1)^2/4. \quad \text{(1 pt)}$$

3. Let us now determine the empirical risk minimizer inside $\mathcal{F}$. Assume that for all $i \in \{1, \dots, n\}$, we have $x_i = i/n$. We have seen in the lecture that the empirical risk minimizer in $\mathcal{F}$ 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 $\lim_{n \to +\infty} \hat{a}$? (2 pts)

Exercise 2. (quadratic cost vs absolute value cost, 7 pts) We consider a regression task $\mathbb{R} \to \mathbb{R}$, with data $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}$. Let $\mathcal{F}$ be the model containing all constant functions $f : \mathbb{R} \to \mathbb{R}$, i.e.

$$\mathcal{F} = \{f : x \mapsto a \mid a \in \mathbb{R}\}.$$

1. Let $f \in \mathcal{F}$. Denote $a \in \mathbb{R}$ such that $f(x) = a$ for all $x \in \mathbb{R}$. Write the empirical risk of f for the **quadratic** cost, using a and $y_1, \dots, y_n$. What is the empirical risk minimizer in $\mathcal{F}$ for the quadratic cost? (1 pt)

2. Let $f \in \mathcal{F}$. Denote $a \in \mathbb{R}$ such that $f(x) = a$ for all $x \in \mathbb{R}$. Write the empirical risk of f for the **absolute value** cost, using a and $y_1, \dots, y_n$. We denote it $\varphi(a)$. (0.5 pt)

3. Assume $y_1 < \dots < y_n$. We want to determine the empirical risk minimizers of $\mathcal{F}$ for the absolute value cost. Let $f \in \mathcal{F}$ equal to $a \in \mathbb{R}$.

   (a) Assume that $a \in [y_i, y_{i+1}]$. Prove that $\varphi(a)$ is equal to $\frac{1}{n}\sum_{j=1}^{i}(a - y_j) + \frac{1}{n}\sum_{j=i+1}^{n}(y_j - a)$. (0.5 pt)

   (b) Compute the derivative of φ with respect to a, on each interval $(-\infty, y_1]$, $[y_1, y_2]$, $\dots$, $[y_{n-1}, y_n]$, $[y_n, +\infty)$. *Hint: on* $[y_i, y_{i+1}]$, *you should obtain* $\varphi'(a) = \frac{2i}{n} - 1$. (1 pt)

(c) **Example 1:** Assume $n = 3$ with $y_1 = 1$, $y_2 = 2$ and $y_3 = 4$. Draw the graph of φ. What is (or are) the minimizer(s) of φ? (1 pt)

(d) **Example 2:** Assume $n = 4$ with $y_1 = 1$, $y_2 = 2$, $y_3 = 3$, $y_4 = 4$. Draw the graph of φ. What is (or are) the minimizer(s) of φ? (1 pt)

(e) **Back to the general case:** What is the sign of $\varphi'(a)$ on each interval $(-\infty, y_1]$, $[y_1, y_2]$, $\dots$, $[y_{n-1}, y_n]$, $[y_n, +\infty)$? What are the minimizers of the empirical risk φ? (1 pt)

4. According to what you obtained in the previous questions, which choice of cost, quadratic or absolute value, is more robust to outliers when doing empirical risk minimization? (1 pt)

Exercise 3. (infinite norm, 2.5 pts) The Minkowski distance of parameter $q \geq 1$ between two vectors $u, v \in \mathbb{R}^p$ reads $\|u - v\|_q = \left(\sum_{j=1}^{p} |u_j - v_j|^q\right)^{1/q}$. The aim of this exercise is to show that $\|u - v\|_q \to_{q \to +\infty} \max_{j=1,\dots,p} |u_j - v_j|$.

1. Prove the result when $u = v$. (0.5 pt)

2. Let us now assume $u - v \neq 0$. Let $\ell \in \{1, \dots, p\}$ be such that $|u_\ell - v_\ell| = \max_{j=1,\dots,p} |u_j - v_j|$.

   (a) Prove that $\|u - v\|_q \geq |u_\ell - v_\ell|$. (0.5 pt)

   (b) Prove that

$$\|u - v\|_q = |u_\ell - v_\ell| \left(\sum_{j=1}^{p} \left(\frac{|u_j - v_j|}{|u_\ell - v_\ell|} \right)^q \right)^{1/q} \leq |u_\ell - v_\ell| \, p^{1/q}. \quad \text{(1 pt)}$$

   (c) What is $\lim_{q \to +\infty} p^{1/q}$? Conclude the proof. (0.5 pt)

Bonus exercise. (decision frontier of nearest neighbors, 3 pts) Consider the following data for binary classification: $((1,2),+)$, $((2,1),+)$, $((2,2),-)$, $((2,3),+)$, $((3,1),-)$, $((3,2),+)$, where the features have dimension 2 and the labels of the classes are $+$ and $-$.

1. Draw the decision frontier of the nearest neighbor algorithm (i.e., $k = 1$) on this data. (1 pt)

2. Draw the decision frontier of the 3-NNs algorithm on this data. (2 pts)

Solutions

Questions on the course.

1. The estimation error of $\mathcal{F}$ is the difference between the true risk of the ERM $\hat{f}$ and the minimal true risk over $\mathcal{F}$:

$$\varepsilon_{\text{estimation}} = R(\hat{f}) - \min_{f \in \mathcal{F}} R(f).$$

The approximation error of $\mathcal{F}$ is the difference between the minimal true risk over $\mathcal{F}$ and the minimal true risk over the set of all predictors $\mathcal{X} \to \mathcal{Y}$:

$$\varepsilon_{\text{approximation}} = \min_{f \in \mathcal{F}} R(f) - \min_{h\colon \mathcal{X} \to \mathcal{Y}} R(h).$$

The estimation error (resp. the approximation error) increases (resp. decreases) when $\mathcal{F}$ becomes richer.

2. No. For example, in classification, when the training set contains only positive samples, the constant classifier that always output the positive class is an ERM. However, it performs badly on negative test samples.

3. F1 score $= \frac{2}{1/\text{precision}+1/\text{recall}} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$.

4. See the course.

5. $f\colon x \mapsto \sigma(b + \sum_{j=1}^{p} w_j x_j)$. See the course for the graph.

6. When k is too small, the algorithm is very sensitive to noise (large variance). When k is too large, the algorithm is robust to noise but not expressive enough (large bias). In particular, when $k = n$, the k-NN predictor is a constant function, that predicts:

   - the mean of training labels in regression,
   - the most common label in classification.

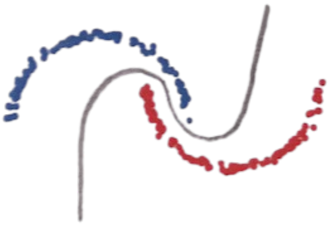
7. We divide our training set D in three subsets D_1, D_2, D_3. We select $\mathcal{K} = \{k_1, \ldots, k_r\}$ a set of values of k among which we want to pick the best one. For $\ell = 1, 2, 3$:

   - we train our k-NN algorithm on $D \setminus D_\ell$,
   - we compute $\hat{R}_{D_\ell}(k)$, its empirical risk on D_ℓ.

   We then pick $k \in \mathcal{K}$ that minimizes the average

$$\frac{1}{3}(\hat{R}_{D_1}(k) + \hat{R}_{D_2}(k) + \hat{R}_{D_3}(k)).$$

8.



Exercise 1.

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

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

where cst does not depend on a. The true risk is minimized at a^* such that $R'(a^*) = 0 \Leftrightarrow \frac{1}{3}a^* = 2 \Leftrightarrow a^* = 6$.

2. Denote P_n: $\sum_{i=1}^n i^2 = n(n+1)(2n+1)/6$. It is easy to check that P_1 is true. If P_n is true for some $n \geq 1$:

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

which proves P_{n+1}. By recurrence, P_n is true for all $n \geq 1$.

Denote Q_n: $\sum_{i=1}^n i^3 = n^2(n+1)^2/4$. It is easy to check that Q_1 is true. If Q_n is true for some $n \geq 1$:

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

which proves Q_{n+1}. By recurrence, Q_n is true for all $n \geq 1$.

3. Let us first compute the denominator.

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

Now the numerator:

$$\begin{aligned}
\sum_{i=1}^{n} x_i T(x_i) &= \sum_{i=1}^{n} \frac{i}{n}(4\frac{i^2}{n^2} + 3\frac{i}{n}) \\
&= \frac{4}{n^3}\sum_{i=1}^{n} i^3 + \frac{3}{n^2}\sum_{i=1}^{n} i^2 \\
&= \frac{4n^2(n+1)^2}{4n^3} + \frac{3n(n+1)(2n+1)}{6n^2} \\
&= \frac{(n+1)^2}{n} + \frac{(n+1)(2n+1)}{2n} \\
&= (n+1)\left(\frac{n+1}{n} + \frac{2n+1}{2n}\right) \\
&= \frac{(n+1)(4n+3)}{2n}.
\end{aligned}$$

Finally,

$$\begin{aligned}
\hat{a} &= \frac{(n+1)(4n+3)}{2n} \times \frac{6n}{(n+1)(2n+1)} \\
&= \frac{12n+6}{2n+1} \\
&\rightarrow_{n\rightarrow+\infty} 6.
\end{aligned}$$

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

Exercise 2.

1. The empirical risk is $\hat{R}_n(f) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - a)^2$. The ERM is equal to $\hat{a} = \frac{1}{n}\sum_{i=1}^{n} y_i$, i.e. the mean of training labels.

2. The empirical risk is $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}|y_i - a|$.

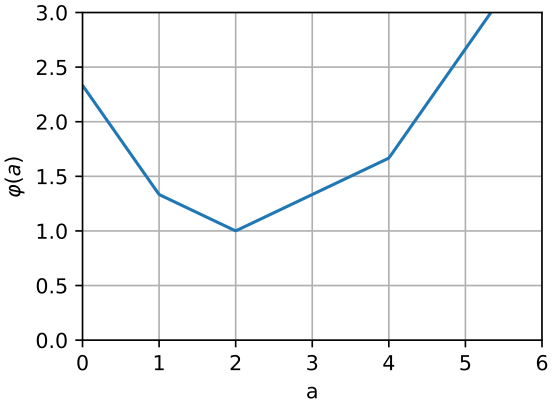
3. (a) For $j \in \{1, \dots, i\}$, we have $|y_j - a| = a - y_j$, as $a \geq y_i \geq y_j$. For $j \in \{i+1, \dots, n\}$, we have $|y_j - a| = y_j - a$, as $a \leq y_{i+1} \leq y_j$. This proves the result.

   (b) Denote $y_0 = -\infty$ and $y_{n+1} = +\infty$. Let $i \in \{0, \dots, n\}$. We have for $a \in [y_i, y_{i+1}]$:

$$\begin{aligned}
\varphi'(a) &= \frac{1}{n}\sum_{j=1}^{i} 1 - \frac{1}{n}\sum_{j=i+1}^{n} 1 \\
&= \frac{1}{n}(i - (n-i)) = \frac{2i}{n} - 1,
\end{aligned}$$

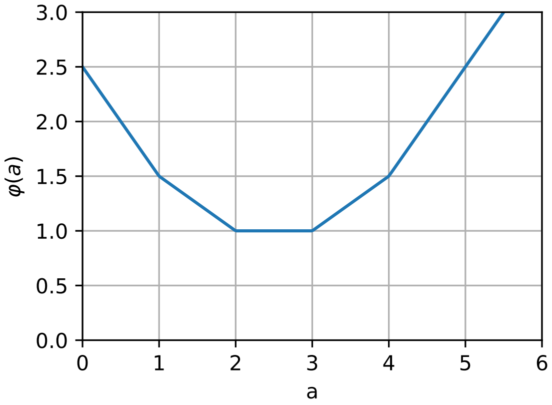
   as there are i terms in the first sum and $n - i$ in the second sum.

(c) We obtain the following plot:



2 is therefore the unique minimizer of φ. Notice that this is also the unique median of (y_1, y_2, y_3).

(d) We obtain the following plot:



The minimizers of φ are therefore all the values in the interval $[2, 3]$. Notice that this is also the set of all medians of (y_1, y_2, y_3, y_4).

(e) If n is odd, $\frac{2i}{n} - 1$ is negative (*i.e. strictement négative*) for $i \leq \frac{n-1}{2}$, and positive (*i.e. strictement positive*) for $i \geq \frac{n+1}{2}$. Therefore, φ is decreasing (*i.e. strictement décroissante*) on $(-\infty, y_{\frac{n+1}{2}}]$ and increasing (*i.e. strictement croissante*) on $[y_{\frac{n+1}{2}}, +\infty)$. Hence, φ reaches its minimum at $y_{\frac{n+1}{2}}$, the median of $(y_1, \ldots, y_n)$.

If n is even, $\frac{2i}{n} - 1$ is negative for $i \leq \frac{n}{2} - 1$, is zero for $i = \frac{n}{2}$, and is positive for $i \geq \frac{n}{2} + 1$. Therefore, φ is decreasing on $(-\infty, y_{\frac{n}{2}}]$, constant on $[y_{\frac{n}{2}}, y_{\frac{n}{2}+1}]$ and increasing on $[y_{\frac{n}{2}+1}, +\infty)$. The minimizers of ϕ are all the values in the interval $[y_{\frac{n}{2}}, y_{\frac{n}{2}+1}]$, namely all the medians of $(y_1, \ldots, y_n)$.

4. The median of $(y_1, \ldots, y_n)$ is insensitive to changing extreme values, which correspond to outliers, whereas the mean of $(y_1, \ldots, y_n)$ changes a lot when y_n becomes very large for example. Therefore, the absolute cost is more robust to outliers than the quadratic cost.

Exercise 3.

1. When $u = v$, we have $\|u - v\|_q = 0 = \max_{j=1,\dots,p}|u_j - v_j|$. We conclude by taking the limit $q \to +\infty$.

2. (a) We have

$$\begin{aligned}\|u - v\|_q^q &= \sum_{j=1}^{p}|u_j - v_j|^q \\ &\geq |u_\ell - v_\ell|^q,\end{aligned}$$

as the other terms in the sum are non-negative. We conclude by applying the increasing function $x \mapsto x^{\frac{1}{q}}$ to each side.

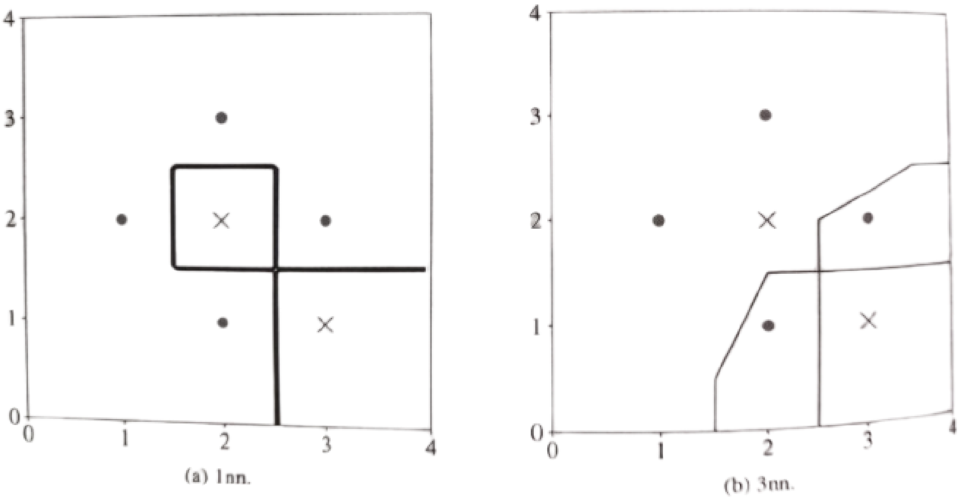
(b) We have

$$\begin{aligned}\|u - v\|_q &= \left(\sum_{j=1}^{p}|u_j - v_j|^q\right)^{1/q} \\ &= \left(\sum_{j=1}^{p}|u_\ell - v_\ell|^q \frac{|u_j - v_j|^q}{|u_\ell - v_\ell|^q}\right)^{1/q} \\ &= |u_\ell - v_\ell| \left(\sum_{j=1}^{p} \frac{|u_j - v_j|^q}{|u_\ell - v_\ell|^q}\right)^{1/q} \\ &\leq |u_\ell - v_\ell| p^{1/q},\end{aligned}$$

as $\frac{|u_j - v_j|^q}{|u_\ell - v_\ell|^q} \leq 1$ for each $j = 1, \dots, p$.

(c) $\lim_{q\to+\infty} p^{1/q} = \lim_{q\to+\infty} e^{\log(p)/q} = 1$ as $\log(p)/q \to 0$ when $q \to +\infty$. We conclude with the squeeze theorem (*théorème des gendarmes*).

Bonus exercise. Solution from *Introduction au Machine Learning*, by C-A Azencott, Chapter 8, exercise 1:



(a) 1nn.

(b) 3nn.