Math of ML : Exercises 5 *

October 13, 2025

Exercise 1 (Kernel ridge regression)**.** *Let $x_1, \dots, x_n \in \mathcal{X}$ be the observed design vectors and let $y_1, \dots, y_n \in \mathbb{R}$ be the observed response variables. Let $\phi : \mathcal{X} \to \mathbb{R}^p$ be a feature mapping, where p is possibly equal to infinity, and let $k(x, y) = \langle \phi(x), \phi(y) \rangle$ be the associated kernel. In this exercise, we are interested in obtaining the predictor $f_{\widehat{\theta}}(x) = \langle \widehat{\theta}, \phi(x) \rangle$, where $\widehat{\theta}$ is the solution to the optimization problem*

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (\theta^\top \phi(x_i) - y_i))^2 + \lambda \|\theta\|_2^2, \tag{1}$$

where $\lambda > 0$ is a regularization parameter. We will consider two separate approaches for solving the above optimization problem.

1. *(Explicit feature vector manipulation approach)*

   (a) *Write an explicit formula for the unique $\widehat{\theta}$ that minimizes the objective (1). What is the computational complexity of computing $\widehat{\theta}$?*

   (b) *Given a new point $x \in \mathcal{X}$, what is the computational complexity of computing $f_{\widehat{\theta}}(x)$?*

2. *(Kernel approach)*

   (a) *Suggest a way to solve the problem (1) without ever explicitly manipulating the feature vectors $\phi(x_i)$ (think of how to represent the output function $f_{\widehat{\theta}}$). Assuming that a computation of $k(x, y)$ takes $O(\kappa)$ number of arithmetic operations, what is the computational complexity of computing $f_{\widehat{\theta}}$ using your method? How does the computational complexity compare with the "explicit feature vector manipulation method" as $p \to \infty$?*

   Hint: Let $\Phi \in \mathbb{R}^{n \times p}$ be an arbitrary matrix. Check that for any $\lambda > 0$ it holds that

   $$\left(\Phi^\top \Phi + \lambda I_p\right)^{-1} \Phi^\top = \Phi^\top \left(\Phi \Phi^\top + \lambda I_n\right)^{-1}.$$

   (b) *Given a new point $x \in \mathcal{X}$, what is the computational complexity of computing $f_{\widehat{\theta}}(x)$?*

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

Exercise 2 (Covering number bounds for the supremum of a stochastic process)**.** *Let $(X_t)_{t\in T}$ be a collection of zero-mean σ^2-sub-Gaussian random variables, indexed by some set T equipped with a metric d. Suppose we wish to upper-bound $\sup_{t\in T} X_t$.*

Remark: *In week 3 of the course, we showed how to upper bound the excess risk of the Empirical Risk Minimization estimator over a hypothesis class $\mathcal{F}$, by controlling the supremum of an empirical process: $\sup_{f\in\mathcal{F}} \hat{\mathcal{R}}(f) - \mathcal{R}(f) = \sup_{f\in\mathcal{F}} \left\{ \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i) - \mathbf{E}_Z \ell(f, Z) \right\}$. This setting corresponds to taking $T = \mathcal{F}$ and $\forall f \in \mathcal{F}, X_f = \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i) - \mathbf{E}_Z \ell(f, Z)$.*

We already know how to handle the case where the index set T is finite (recall exercise 1.5 in exercise sheet 3). We have also seen how to deal with some well-behaved infinite index sets T and stochastic processes $(X_t)_{t\in T}$ via the use of Rademacher complexities (e.g., when bounding the Rademacher complexity of norm-constrained linear predictors, as in section 4.2 of lecture 3).

The purpose of this exercise is to introduce a general technique for bounding $\mathbf{E}[\sup_{t\in T} X_t]$ for infinite index sets T. The key idea is to approximate the infinite index set with a finite set, use exercise 1.5 of exercise sheet 3 to bound the maximum of the finite set, and pay an additional approximation error term. Observe that there is a trade-off between the size of the approximating finite set and the incurred approximation error. Before proceeding with the exercise, we need one additional definition.

Definition 1 (ε-net and ε-covering number)**.** *A set $\{t_1, \dots, t_N\} \subseteq T$ is said to be an ε-net for the metric space (T, d) if for any $t \in T$ there exists $j \in \{1, \dots, N\}$ such that $d(t, t_j) \leq \varepsilon$. The cardinality of the smallest ε-net, denoted $N(\varepsilon, T, d)$, is an ε-covering number of (T, d).*

1. *Suppose that there exists a random variable L such that for any $t, s \in T$ we have $|X_t - X_s| \leq L d(t, s)$. Prove that*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \leq \inf_{\varepsilon > 0} \left\{ \mathbf{E}[L]\varepsilon + \sqrt{2\sigma^2 \log N(\varepsilon, T, d)} \right\}. \tag{2}$$

2. *We will now consider a simple application of the bound (2). Let $M \in \mathbb{R}^{n\times m}$ be a rectangular matrix such that each entry M_{ij} is an independent zero-mean τ^2-sub-Gaussian random variable. The operator norm of M is defined by $\|M\| = \sup_{x\in B_n, y\in B_m} \langle x, My\rangle$, where $B_d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ is the unit ball in the d-dimensional Euclidean space. Prove that $\mathbf{E}\|M\| \leq c\tau\sqrt{n+m}$, where $c > 0$ is a universal constant.*

   Hint: *Let $x, y \in B_d$ and consider the metric induced by the Euclidean norm $d_2(x, y) = \|x - y\|_2$. Then, for any $\varepsilon \in (0, 1)$ we have*

$$\left(\frac{1}{\varepsilon}\right)^d \leq N(\varepsilon, B_d, d_2) \leq \left(1 + \frac{2}{\varepsilon}\right)^d. \tag{3}$$

   The above inequality is classical; see, e.g., [Wainwright, 2019, Example 5.8] for a proof.

Exercise 3 (Random Fourier features)**.** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous real-valued translation-invariant kernel scaled so that for any $x \in \mathbb{R}^d$ we have $k(x, x) = 1$. Then, by Bochner's theorem there exists a probability measure P such that*

$$k(x, x') = q(x - x') = \int_{\mathbb{R}^d} \exp\left(i\omega^\top (x - x')\right) P(d\omega) = \mathbf{E}_{\omega\sim P}\left[\exp\left(i\omega^\top (x - x')\right)\right]. \tag{4}$$

The aim of this exercise is to exploit the representation (4) to build an approximate (and random) feature mapping $\widehat{\phi} : \mathbb{R}^d \to \mathbb{R}^{2p}$ such that $k(x, x') \approx \widehat{\phi}(x)^\top \widehat{\phi}(x')$.

1. *Using the fact that the kernel k is real-valued, show using (4) that*

$$k(x, x') = \mathbf{E}_{\omega \sim P}\left[\cos(\omega^\top x)\cos(\omega^\top x') + \sin(\omega^\top x)\sin(\omega^\top x')\right]. \tag{5}$$

2. *The identity (5) suggests drawing p i.i.d. samples $\omega_1, \ldots, \omega_p$ from the distribution P to build the approximate (random) feature map*

$$\widehat{\phi}(x) = \frac{1}{\sqrt{p}}\left(\cos(\omega_1^\top x), \sin(\omega_1^\top x), \cos(\omega_2^\top x), \sin(\omega_2^\top x), \ldots, \cos(\omega_p^\top x), \sin(\omega_p^\top x)\right)^\top \in \mathbb{R}^{2p}.$$

Let $\widehat{k}(x, x') = \widehat{\phi}(x)^\top \widehat{\phi}(x')$ and let $r > 0$ be a positive constant and let $B_r = \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$. Let $\sigma_k = \mathbf{E}_{\omega \sim P}[\|\omega\|_2]$. Prove that

$$\mathbf{E}_{\omega_1, \ldots, \omega_p}\left[\sup_{x, x' \in B_r}\left\{k(x, x') - \widehat{k}(x, x')\right\}\right] \leq c\sqrt{\frac{d}{p}\log\left(1 + \frac{r\sigma_k\sqrt{p}}{\sqrt{d}}\right)},$$

where $c > 0$ is some universal constant. In particular, the above result shows that to get an ε-approximation of the kernel k, it suffices to take $p = \tilde{O}(d/\varepsilon^2)$, where the notation $\tilde{O}(\cdot)$ hides logarithmic factors.

Hint: *use the bounds (2) and (3) stated in Exercise 2.*

References

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Ramon Van Handel. Probability in high dimension. Technical report, Princeton University, 2014. URL https://web.math.princeton.edu/~rvan/APC550.pdf.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.