Computational Foundations of Data Sciences — Exercises

Théo Lacombe

December 6, 2024

Preamble. This exercise sheet gathers two types of exercises: some that have already been addressed during the lectures and which are mostly "prove what is written in the slides", and some that can be seen as "complements" (they consider new problems that where not mentioned during the lecture). The latter are indicated by a (*).

Some exercises also contains the tag "**(explain)**". For such exercises, you need to explain in your own words some concept studied in the course. Your speech should be accessible to, say, a M1 student (i.e. someone with some solid background in maths and computer science, but no advanced knowledge in machine learning / data sciences).

Some questions are marked as "Bonus". It means that they are not necessary to catch the main message carried by the exercise. They can be seen as complementary material on the topic covered by the exercise.

1 Chapter 0 & 1: Generalities

Exercise 1 Let $x_1, \dots, x_n \in \mathbb{R}^d$ be observations. As we seek for a *representative* of our observation, a natural goal is to minimize

$$\hat{x} \mapsto \sum_{i=1}^{n} \|x - x_i\|^p, \tag{1}$$

for some $p \geq 1$.

1. Find the optimal $\hat{x}$ when $p = 2$ (justify).
2. What happen if $p = 1$? if $p > 1$ and $p \neq 2$?

Exercise 2 (explain) What are the differences between supervised and unsupervised learning? What are the pro and cons of both problems?

2 Chapter 2: Regression

Exercise 3 (explain) What does the "overfitting" phenomenon refers to? How can it be mitigated?

Exercise 4 (linear regression) Let $(x_i, y_i)_{i=1}^n$ be couples of observations and labels in $\mathbb{R}^d \times \mathbb{R}^k$. We consider the linear model with parameter $M \in \mathbb{R}^{d\times k}$ defined as $F_M(x) = xM$ for $x \in \mathbb{R}^d$. We want to minimize the following objective function:

$$M \mapsto \sum_{i=1}^{n} \|F_M(x_i) - y_i\|^2, \tag{2}$$

for some loss function $\ell : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_+$.

1. Recall why this model encompasses (i) affine regression $x \mapsto ax + b$, (ii) polynomial regression $x \mapsto a_0 + a_1x_1 + \cdots + a_dx^d$.
2. Let $\|A\|_2^2 = \mathrm{Tr}(AA^T)$ denote the (squared) Froebenius norm of a matrix A. Show that in that case, (2) boils down to minimize the function $M \mapsto \|XM - Y\|$, where $X \in \mathbb{R}^{n\times d}$ is the matrix of observations, and $Y \in \mathbb{R}^{n\times k}$ the matrix of corresponding labels.
3. Assuming that $(X^TX)^{-1}$ is invertible, determine the expression of the optimal M.
4. Propose an interpretation of this assumption.
5. Using the notion of pseudo-inverse (click on the link or search on Wikipedia), show that one can still find an expression of optimal M even if $(X^TX)^{-1}$ is singular.

Exercise 5 (Ridge regression) With the same notation as Exercise 4, we consider the model $F_M : x \mapsto xM$ and the objective function

$$M \mapsto \|XM - Y\|^2 + \lambda\|M\|^2, \tag{3}$$

where $\lambda \geq 0$ is an hyperparameter.

1. Recall what this model is useful for.
2. Show that, under some suitable assumption, one has access to a closed form solution for the minimizer of (3)

3 Chapter 3: An optimization detour

Exercise 6 Let $L : \mathbb{R}^d \to \mathbb{R}$ denote a real-valued function, and assume that it is of class $\mathcal{C}^1$.

Explain the claim "$-\nabla L(\theta)$ indicates the steepest descent direction of L at θ."

Exercise 7 (Basic properties of convex functions) Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, assumed to be of class $\mathcal{C}^2$ for the sake of simplicity.

1. Prove that for all $x, y \in D(f)$ (the domain of f), one has
$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle . \tag{4}$$
2. Deduce that the gradient of f should be monotone, that is for all $x, y \in D(f)$,
$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0. \tag{5}$$
3. Deduce also that if $\nabla f(x) = 0$, then x should be a global minimizer of f.
4. Prove that for all $x \in D(f)$, the Hessian matrix of f at x should be positive semi-definite (i.e. all its eigenvalues should be ≥ 0).

Exercise 8 (strong convexity) Let $f : \mathbb{R}^d \to \mathbb{R}$ be an α-strongly convex function, for $\alpha > 0$, assumed to be of class $\mathcal{C}^2$ for the sake of simplicity.

1. Show that for all $x, y \in D(f)$, one has
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2. \tag{6}$$
2. Show that the function $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.
3. Show that the eigenvalues of the Hessian matrix of f at any $x \in D(f)$ should be larger than α.
4. Show that f admits a unique minimizer x^*.
5. Prove that f satisfies the PL-inequality
$$0 \leq f(x) - f(x^*) \leq \frac{1}{2\alpha}\|\nabla f(x)\|^2. \tag{7}$$

Exercise 9 (Convergence rate of the gradient descent) Let $f : \mathbb{R}^d \to \mathbb{R}$ be an α-strongly convex and β-smooth function. Let x^* denote the unique (global) minimizer of f. Let $x_0 \in \mathbb{R}^d$ and consider the sequence defined by $x_{t+1} = x_t - \lambda \nabla f(x_t)$ for some (fixed) $\lambda > 0$.

1. Using the β-smoothness assumption, show that

$$f(x_{t+1}) - f(x_t) \leq -\lambda \left(1 - \frac{\lambda\beta}{2}\right) \|\nabla f(x_t)\|^2.$$

2. Using the PL-inequality (see (7)), deduce that if λ is small enough (to be specified), one has

$$f(x_{t+1}) - f(x^*) \leq \left(1 - 2\alpha\lambda \left(1 - \frac{\beta\lambda}{2}\right)\right) (f(x_t) - f(x^*)).$$

3. Deduce that if λ belongs to some interval (to be specified), one obtains exponential convergence rate of the current value $(f(x_t))_t$ toward the minimal value $f(x^*)$.

4 Chapter 4: Classification

Exercise 10 (Convexity of the Logistic Regression) The goal of this exercise is to show the convexity of the cross-entropy for the logistic regression. More precisely, given $(x_i, y_i)_{i=1}^n$, with $x_i \in \mathbb{R}^d$ and $y_i = (0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{R}^K$ (K being the number of classes), we want to prove that

$$L : \mathbb{R}^{K\times d} \ni \theta \mapsto -\sum_{i=1}^{n} y_i \cdot \log(\text{smax}(\theta x_i)) \tag{8}$$

is convex, where $\text{smax}(a) = \left(\frac{e^{a[j]}}{\sum_{j'=1}^K e^{a[j']}}\right)_{j=1}^K$ and the log is applied term-wise.

1. Show that if a function $f : \mathbb{R}^K \to \mathbb{R}$ is convex and $x \in \mathbb{R}^d$, then $\mathbb{R}^{K\times d} \ni \theta \mapsto f(\theta x)$ is convex.

2. Show that the problem boils down to showing that the *log-sum-exp* function $\varphi : a \mapsto \log\left(\sum_{j=1}^K e^{a[j]}\right)$ is convex.

3. Using Hölder inequality, prove that φ is convex.

4. Propose another proof based using the Hessian matrix of φ.

5 Chapter 5: Unsupervised learning

Exercise 11 (Convergence of the Lloyd algorithm) TBA.

Exercise 12 (Quantization problem) (*) Let P be a probability distribution on $\mathbb{R}^d$ with finite second moment (i.e. $\int |x|^2 \mathrm{d}P(x) < \infty$; denoted as $P \in \mathcal{P}_2(\mathbb{R}^d)$), and $n \in \mathbb{N}$. We also assume that P is supported on at least n points.

A *n-quantizer* is a measurable map $f : \mathbb{R}^d \to \mathbb{R}^d$ whose image contains (at most) n distinct points, i.e. $|f(\mathbb{R}^d)| \leq n$. Let $\mathcal{F}_n$ denote the set of all n-quantizers. The energy of $f \in \mathcal{F}_n$ over P is defined as

$$\mathcal{E}_P(f) \coloneqq \mathbb{E}_P[|X - f(X)|^2] = \int_{\mathbb{R}^d} |x - f(x)|^2 \mathrm{d}P(x).$$

A n-quantizer f is said to be optimal for P if it minimizes $\mathcal{E}_P$. Let $V_n(P) \coloneqq \inf_{f \in \mathcal{F}_n} \mathcal{E}_P(f)$ denote the corresponding minimal value.

1. Show that for all $f \in \mathcal{F}_n$, $\mathcal{E}_P(f) < \infty$.
2. Show that
$$V_n(P) = \inf_{\alpha \subset \mathbb{R}^d, |\alpha| \leq n} \mathbb{E}_P[\min_{a \in \alpha} |X - a|^2]. \tag{9}$$
3. Let $\psi_P : (\mathbb{R}^d)^n \to \mathbb{R}$ be defined as $\psi_P(a_1, \dots, a_n) \coloneqq \mathbb{E}_P[\min_{i=1,\dots,n} |X - a_i|^2]$. Minimizers of ψ are called *n-optimal set of centers*. Show that minimizing ψ is equivalent to minimizing $f \mapsto \mathcal{E}_P(f)$ and that there is a natural correspondence between the minimizers of $\mathcal{E}_P$ and ψ_P.
4. Is ψ_P convex in general?
5. Show that ψ_P is continuous.
6. Show that $V_n(P) < V_{n-1}(P)$.
7. (hard) Prove that the sublevel set $\{(a_1, \dots, a_n),\ \psi_P(a_1, \dots, a_n) \leq c\}$ is compact for all $c \in (0, V_{n-1}(P))$.
8. Deduce that the set of n-optimal quantizers of P is not empty.
9. Assume that we are observing a finite set of N observations $x_1, \dots, x_N \in \mathbb{R}^d$, in which case it makes sense to set $P = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}$ (where δ_x denote the Dirac mass located at $x \in \mathbb{R}^d$). In that case, the quantization can be interpreted as a (i) supervised learning problem, (ii) unsupervised learning problem? (Pick one.)
10. Observe that the k-mean problem is a quantization problem.
11. (Bonus) Define the *Wasserstein distance*[1] between two probability measures μ and ν as
$$W_2(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu, \nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \mathrm{d}\pi(x, y) \right)^{\frac{1}{2}}, \tag{10}$$

[1] You may consider proving that this is indeed a distance on $\mathcal{P}_2(\mathbb{R}^d)$. This is an interesting yet pretty hard exercise from scratch; you may look at some references online.

where $\Pi(\mu, \nu)$ denote the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ that admit μ and ν as first and second marginal, respectively. For a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$, let also $T\#\mu$ denote the *pushforward* measure of μ by T, that is the measure defined for any borel $A \subset \mathbb{R}^d$ as $T\#\mu(A) = \mu(T^{-1}(A))$. Prove that

$$V_n(P) = \inf_{f \in \mathcal{F}_n} W_2(\mu, f\#\mu)^2. \tag{11}$$

Exercise 13 (Gaussian mixture models) (*) TBA

Exercise 14 (Spectral clustering) (*) TBA

Exercise 15 (Linear AE) (*) We consider a *linear* autoencoder, that is the encoder is given by $x \mapsto Ax$ and the decoder by $z \mapsto Bz$, where $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Z} = \mathbb{R}^k$, and $A \in \mathbb{R}^{k \times d}$ and $B \in \mathbb{R}^{d \times k}$. Training this autoencoder on a dataset $X \in \mathbb{R}^{d \times n}$ consists of minimizing

$$(A, B) \mapsto \|BAX - X\|^2.$$

1. Assuming that (XX^T) is non-singular (invertible), show that at fixed B such that B^TB is non-singular as well, one has $A = (B^TB)^{-1}B^T$. Note: this matrix is called the Moore-Penrose pseudo-inverse of B.
2. Deduce that training a linear AE is substantially equivalent to perform a PCA.
3. What happen if we consider an affine AE, i.e. the encoder is $x \mapsto Ax + a$ and the decoder is $z \mapsto Bz + b$ for some vectors a, b (that are also optimized at training time, along with the matrices A and B)?

6 Chapter 6: Kernel methods

Exercise 16 (Examples of kernels) (*)

The linear kernel. Let $\mathcal{X} = \mathbb{R}^d$ and define $K(x, y) = \langle x, y \rangle$ (the usual Euclidean inner-product on $\mathbb{R}^d$).

1. Show that K defines a PSD kernel on $\mathbb{R}^d$.
2. What is the corresponding RKHS $\mathcal{H}$? (specify the inner product in $\mathcal{H}$)

The quadratic kernel. Still with $\mathcal{X} = \mathbb{R}^d$, let $K(x, y) = \langle x, y \rangle^2$.

1. Expressing $\langle x, y \rangle^2$ as an inner-product, prove that K is a PSD kernel on $\mathbb{R}^d$.
2. Determine the corresponding RKHS.

Exercise 17 (Basic properties of Kernels) Let K and K' be two (PSD) kernels on a space $\mathcal{X}$. Show that

1. $K + K'$ is a kernel.
2. $K \cdot K'$ is a kernel.
3. (*) Show that if K is a kernel, so is e^K.

Exercise 18 (Kernel and Fourier transform) Let $K(x, y) = h(x - y)$ for some $h : \mathbb{R}^d \to \mathbb{R}$ (that satisfies $h(u) = h(-u)$, making K is symmetric). Assume that the Fourier transform of h—defined by $\hat{h}(\omega) \coloneqq \int_{\mathbb{R}^d} e^{-i\langle \omega, u \rangle} \mathrm{d}u$—is non-negative.

- Show that K is a PSD kernel on $\mathbb{R}^d$.
- Deduce that $(x, y) \mapsto \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ defines a kernel on $\mathbb{R}^d$ (for any $\sigma > 0$).

Exercise 19 (Kernelization of k-means) (*) Let $\mathcal{X}$ be a set and consider a PSD kernel K on $\mathcal{X}$ with corresponding RKHS $\mathcal{H}$ and feature map φ. The goal of this exercise is to show that, given a set of observation $x_1, \ldots, x_n$ in $\mathcal{X}$, we can consider solving the k-means problem on $\mathcal{H}$ with observations $\varphi(x_1), \ldots, \varphi(x_n)$ only through the manipulation of the Gram matrix $G \coloneqq (K(x_i, x_j))_{ij}$.

For this, we recall that the loss function of the k-means problem is given by

$$L(c_1, \ldots, c_k) \coloneqq \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|\varphi(x_i) - c_j\|_{\mathcal{H}}^2, \tag{12}$$

where $c_1, \ldots, c_k \in \mathcal{H}$.

1. Show that
$$L(c_1, \ldots, c_k) = \sum_{i=1}^{n} \|\varphi(x_i) - c_{s_i}\|_{\mathcal{H}}^2,$$
where $s_i = \arg\min_j \|\varphi(x_i) - c_j\|_{\mathcal{H}}^2$.
2. Show that the map $c \mapsto \sum_{i=1}^n \|\varphi(x_i) - c\|_{\mathcal{H}}^2$ is minimized when $c = \frac{1}{n}\sum_{i=1}^n \varphi(x_i)$. Beware that $\mathcal{H}$ may be an infinite dimensional Hilbert space, so you cannot faithfully rely on gradient computation as we usually did on $\mathbb{R}^d$.

3. Introducing $C_j \coloneqq \{i,\ s_i = j\}$ for $j = 1, \dots, k$, deduce that minimizing (12) boils down to minimizing

$$(s_1, \dots, s_n) \mapsto \sum_{i=1}^{n} \left\| \varphi(x_i) - \frac{1}{|C_{s_i}|} \sum_{j \in C_{s_i}} \varphi(x_j) \right\|_{\mathcal{H}}^2,$$

where $s_1, \dots, s_n$ are now variables in $\{1, \dots, k\}$.

4. Deduce that performing k-means in $\mathcal{H}$ boils down to an optimization problem on $(s_1, \dots, s_n)$ that only requires the knowledge of K. Note: you can reach a simple formula after careful simplification. Nonetheless, the optimization problem remains highly non-trivial.

Exercise 20 (Heat kernel on graphs) Let V denote a finite set $(v_1, \dots, v_n)$ and $G = (V, E)$ denote a non-oriented connected graph, i.e. $E \subset V \times V$ and $(v, v') \in E \Rightarrow (v', v) \in E$ and for any v, v' there always exists a *path* $(v_0, \dots, v_N)$ with $v_0 = v, v_N = v'$ and $(v_i, v_{i+1}) \in E$ for all $i = 0, \dots, N-1$. We introduce the following notations:

- The adjacency matrix $A \in \mathbb{R}^{n \times n}$ is defined by $A_{ij} = 1$ if $(v_i, v_j) \in E$, 0 otherwise.
- The degree matrix $D \in \mathbb{R}^{n \times n}$ is the diagonal matrix defined by $D_{ii} = \sum_{j=1}^{n} A_{ij}$.
- The Laplacian matrix of G is the matrix $L \coloneqq D - A \in \mathbb{R}^{n \times n}$.
- Given a square matrix $M \in \mathbb{R}^{n \times n}$, we let $\exp(M) \coloneqq \sum_{k \geq 0} \frac{M^k}{k!}$ denote the usual exponential of a matrix.

1. Let $f_0 \in \mathbb{R}^n$ be a n-dimensional vector (that can be understood as assigning values on the nodes of G). Consider the *discrete diffusion equation* on $\mathbb{R}^n$ defined as

$$\partial_t f(t) = -L f(t),$$

with $f(0) = f_0$.

Show that $f(t) = f_0 e^{-tL}$ is a solution of this equation.

2. Show that L is a symmetric, positive definite matrix.

3. Deduce that for any $t > 0$, $K_t \coloneqq e^{-tL}$ defines a PSD kernel on G (formally, the kernel is $(v_i, v_j) \mapsto (K_t)_{ij}$).

This kernel is called the *heat kernel* on G with temperature t.

4. (bonus) Compute the heat kernel (for any temperature $t > 0$) for the complete graph ($A_{ij} = 1$ for all $i \neq j$).

5. (bonus) It may be tempting to consider the natural *shortest path* distance on G defined by $d(v, v') = \min_N \{(v_0, \dots, v_N) \subset V,\ v_0 = v, v_N = v',\ (v_i, v_{i+1}) \in E \text{ for all } i = 0, \dots, N-1\}$ and build a kernel on G through $K_{ij} = e^{-d(v_i, v_j)^2}$. Show that this approach does not work (i.e. such a K may not be PSD).