Introduction to Statistical Supervised Machine Learning

Master 1 MIASHS (2022-2023)

Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Abstract

This document contains a list of exercises related to the Supervised Machine Learning lessons in order to illustrate or apply the different notions discussed in the course.

It also contains some exercises in Mathematics/Statistics, and more precisely in Analysis, Optimization or Linear Algebra, fundamental tools to understand the intuitive approaches or the demonstrations presented in the course.

The exercises are arranged by theme but not necessarily by order of difficulty, so it will not be uncommon to see difficult exercises before simpler ones.

I take this opportunity to thank the students who allowed the realization of this booklet via their questions and their request, but also Ben Gao (M1 MIASHS) for the participation to the writing of the corrections.

Contents

1 Linear Algebra

1.1 Inner Product and Norms

Exercice 1.1. *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be two vectors. Which of the following applications define an inner product :*

1. $f(\mathbf{x}, \mathbf{y}) = x_1 y_1 + x_2 y_2$.
2. $f(\mathbf{x}, \mathbf{y}) = x_1 y_1 + x_2 y_2 - x_3 + y_3$.
3. $f(\mathbf{x}, \mathbf{y}) = x_1 y_1 + 2x_2 y_2 + 3x_3 y_3$.
4. $f(\mathbf{x}, \mathbf{y}) = x_1^2 y_1^2 + x_2^2 y_2^2 + x_3^3 y_3^3$.

Correction.

To prove that the following functions define an inner product, you have to check that they bilinear, symmetric, positive and definite form. In other words, you have to check that:

- $f(\mathbf{x}, \mathbf{x}) \geq 0$,
- $f(\mathbf{x}, \mathbf{x}) = 0 \iff \mathbf{x} = 0$,
- $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}, \mathbf{x})$,
- $f(\mathbf{x} + \lambda \mathbf{y}, \mathbf{z}) = f(\mathbf{x}, \mathbf{z}) + \lambda f(\mathbf{y}, \mathbf{z})$.

1. This first function define an inner product by definition
2. This second function do not satisfy the positive definite character. Indeed, if we take $\mathbf{x} = (0, 0, 1)$, we have $f(\mathbf{x}, \mathbf{x}) = 0$ but $\mathbf{x} \neq \mathbf{0}$.
3. You can easily check that this function is bilinear positive and definite.
4. This last function is not bilinear due to the presence of the quadratic and cubic terms.

Exercice 1.2. *The aim of this exercise is to work with norms and inner product with matrices.*

1. *Show that the application* $\langle \bullet, \bullet \rangle : \mathscr{M}_{m,d}(\mathbb{R}) \times \mathscr{M}_{m,d}(\mathbb{R}) \to \mathbb{R}$ *defined by :*

$$\langle \mathbf{A}, \mathbf{B} \rangle = trace(\mathbf{A}^T \mathbf{B}),$$

defined an inner product.

2. *Show that* $\|\mathbf{A}\|_F = \sqrt{trace(\mathbf{A}^T\mathbf{A})} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(a_{ij}^2)}$*, and show that it defines a norm.*

3. *Show that* $\|\mathbf{Ax}\|_2 \leq \|\mathbf{A}\|_F\|\mathbf{x}\|_2$ *where* $\mathbf{A} \in \mathcal{M}_{m,d}(\mathbb{R})$ *and* $\mathbf{x} \in \mathbb{R}^d$.

4. *Show that* $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F\|\mathbf{B}\|_F$ *where* $\mathbf{A} \in \mathscr{M}_{n,m}(\mathbb{R})$ *and* $\mathbf{B} \in \mathscr{M}_{m,p}(\mathbb{R})$

5. *Compute the Frobenius norm of the following matrices:*

$$\mathbf{A} = \begin{pmatrix} 1 & -3 \\ -3 & 1 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 3 & -2 & 3 \\ -2 & 1 & -2 \\ -3 & 2 & 3 \end{pmatrix}$$

Correction.

We consider $\mathbf{A} = (a_{ij})_{i,j=1}^{m,d}$ and $\mathbf{B} = (b_{ij})_{i,j=1}^{m,d}$

1. In this question, we need to check the four points as it was done in the previous exercise.

   Remember that $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i=1}^{m}\sum_{j=1}^{d} a_{ij}b_{ij}$.
   $\langle \mathbf{A}, \mathbf{A} \rangle = \sum_{i=1}^{m}\sum_{j=1}^{d} a_{ij}^2 \geq 0$ and this sum is equal to 0 if and only if $\mathbf{A} = \mathbf{0}$.
   It is clearly symmetric and for all $\lambda \in \mathbb{R}$ and $\mathbf{C} \in \mathscr{M}_{m,d}(\mathbb{R})$, we have:

$$\langle \mathbf{A}+\lambda\mathbf{C}, \mathbf{B} \rangle = \sum_{i=1}^{m}\sum_{j=1}^{d}(a_{ij}+\lambda c_{ij})(b_{ij}) = \lambda\sum_{i=1}^{m}\sum_{j=1}^{d} a_{ij}c_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{d} a_{ij}b_{ij} = \lambda\langle \mathbf{A}, \mathbf{C} \rangle + \langle \mathbf{A}, \mathbf{B} \rangle.$$

2. For the first part of the question, you just have to apply the definition on an inner product. We then check that this application defines a norm.
   It is obviously positive and definite for the same reason as before.
   For the scalability property, we consider $\lambda \in \mathbb{R}$ and evaluate $\|\lambda\mathbf{A}\|$:

$$\|\lambda\mathbf{A}\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{d}\lambda^2 a_{ij}^2\right)^{\frac{1}{2}} = |\lambda|\left(\sum_{i=1}^{m}\sum_{j=1}^{d} a_{ij}^2\right)^{\frac{1}{2}} = |\lambda|\|\mathbf{A}\|_F.$$

   Finally, the triangle inequality:

$$\|\mathbf{A}+\mathbf{B}\|_F^2 = Tr((\mathbf{A}+\mathbf{B})^T(\mathbf{A}+\mathbf{B})),$$

   ↓ develop the product

$$
\begin{aligned}
&= Tr(\mathbf{A}^T\mathbf{A} + \mathbf{A}^T\mathbf{B} + \mathbf{B}^T\mathbf{A} + \mathbf{B}^T\mathbf{B}),\\
&\quad \downarrow \text{ use of linearity}\\
&= Tr(\mathbf{A}^T\mathbf{A}) + Tr(\mathbf{A}^T\mathbf{B}) + Tr(\mathbf{B}^T\mathbf{A}) + Tr(\mathbf{B}^T\mathbf{B}),\\
&\quad \downarrow \text{ definition of Frobenius norm}\\
&= \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + 2Tr(\mathbf{A}^T\mathbf{B}),\\
&\leq \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + 2\|\mathbf{A}\|_F\|\mathbf{B}\|_F,\\
&\quad \downarrow \text{ Cauchy Scwharz}\\
&\leq \left(\|\mathbf{A}\|_F + \|\mathbf{B}\|_F\right)^2.
\end{aligned}
$$

We conclude by taking the square root.

3. We will simply apply the definition of norm to the vector $\mathbf{Ax}$ where $(\mathbf{Ax})_k = \sum_{j=1}^{d} x_j a_{kj}$. We then have:

$$
\begin{aligned}
\|\mathbf{Ax}\|_2^2 &= \sum_{k=1}^{m}\left(\sum_{j=1}^{d} x_j a_{kj}\right)^2,\\
&\quad \downarrow \text{ Use the Cauchy-Schwarz inequality}\\
&\leq \left(\sum_{j=1}^{d} x_j^2\right)\left(\sum_{j=1}^{d}\sum_{k=1}^{m} a_{kj}^2\right),\\
&= \|\mathbf{x}\|_2^2\|\mathbf{A}\|_F^2.
\end{aligned}
$$

We also conclude by tacking the square root on both sides.

- We will use the definition of norm and apply the Cauchy-Scwharz inequality:

$$
\begin{aligned}
\|\mathbf{AB}\|_F^2 &= \sum_{i=1}^{d}\sum_{j=1}^{d}\left(\sum_{k=1}^{m} a_{ik}b_{kj}\right)^2,\\
&\quad \downarrow \text{ Cauchy-Scwharz}\\
&\leq \sum_{i=1}^{d}\sum_{j=1}^{d}\left(\sum_{k=1}^{m} a_{ik}^2\sum_{k=1}^{m} b_{kj}^2\right),\\
&\leq \sum_{k=1}^{m}\left(\sum_{i=1}^{d} a_{ik}^2\sum_{j=1}^{d} b_{kj}^2\right),\\
&\leq \left(\sum_{i=1}^{m}\sum_{k=1}^{d} a_{ik}^2\right)\left(\sum_{j=1}^{p}\sum_{k=1}^{m} b_{kj}^2\right),\\
&\leq \|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2.
\end{aligned}
$$

- $$\|\mathbf{A}\|_F = 2\sqrt{5} \quad \textit{and} \quad \|\mathbf{B}\|_F = \sqrt{53}.$$

Exercice 1.3. *Show that, for all* $\mathbf{x} \in \mathbb{R}^d$*, the function* $\| \bullet \|_2 : \mathbf{x} \mapsto \sqrt{\sum_{j=1}^d x_j^2}$ *defines a norm.*

Correction.

We are going to show the same points as it was previously done

- For all $\mathbf{x} \in \mathbb{R}^d$

$$\|\mathbf{x}\|_2^2 = \sum_{j=1}^d x_j^2 > 0.$$

- The sum of positives numbers is equal to 0 if and if all these numbers are equal to 0, *i.e.*

$$\|\mathbf{x}\|_2^2 = 0 \iff \mathbf{x} = \mathbf{0}.$$

- Let us consider $\lambda \in \mathbb{R}$, then:

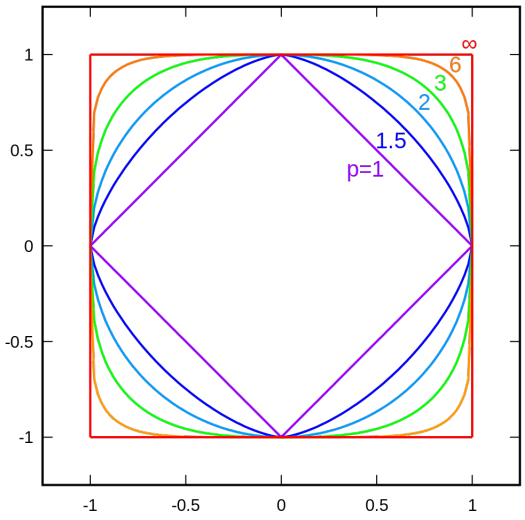
$$\begin{aligned}
\|\lambda\mathbf{x}\|_2^2 &= \sum_{j=1}^d \lambda^2 x_j^2, \\
&= \lambda^2 \sum_{j=1}^d x_j^2, \\
&= \lambda^2 \|\mathbf{x}\|_2^2.
\end{aligned}$$

  The result fall by taking the square root on both sides.

- It remains to prove the triangle inequality. Let us consider $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$\begin{aligned}
\|\mathbf{x}+\mathbf{y}\|_2^2 &= (\mathbf{x}+\mathbf{y})^T(\mathbf{x}+\mathbf{y}), \\
&= \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 + 2\mathbf{x}^T\mathbf{y}, \\
&\quad \downarrow \text{ Cauchy-Schwarz} \\
&= \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 + 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2, \\
&= \left(\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2\right)^2.
\end{aligned}$$

Exercice 1.4. *The aim of the exercise is to prove the inequality of Minkowski, i.e the triangular inequality for the L^p norm for $p \in [1, \infty[$.*



Let us consider $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ are considered as vectors here.

1. *Let $0 < p, q < \infty$ such that $\dfrac{1}{p} + \dfrac{1}{q} = 1$*

   (a) *Show that $\ln(ab) = \dfrac{\ln(a^p)}{p} + \dfrac{\ln(b^q)}{q}$ for all $a, b > 0$.*

   (b) *Use the convexity of the exponential to show Young's inequality:*

   $$|ab| \leq \frac{|a|^p}{p} + \frac{|b|^q}{q}.$$

2. *We want to prove now that : $\|\mathbf{xy}\|_1 \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q$ (Hölder's inequality)*
   We consider $0 < p, q < \infty$ such that $\dfrac{1}{r} = \dfrac{1}{p} + \dfrac{1}{q}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

   (a) *By a good choice of $p, q, \mathbf{x}$ and $\mathbf{y}$, show that, applying Young's inequality :*

   $$|x_i y_i|^r \leq \frac{1}{p'}|x_i|^p + \frac{1}{q'}|y_i|^q,$$

   where you should determine the value of p' and q'.

   (b) *Prove Hölder's inequality using the previous result (first you have to take the sum on all i and consider the special case where $r = 1$)*[1].

[1] *Hint:* set $x_i = \dfrac{x_i}{\|\mathbf{x}\|_p^p}$ and $y_i = \dfrac{y_i}{\|\mathbf{y}\|_q^q}$

3. *Let us now proove the triangle inequality for the L^p norm.*

   (a) *Use successively the triangle inequality and Hölder's inequality to show that :*

   $$\|\mathbf{x}+\mathbf{y}\|_p^p \leq (\|\mathbf{x}\|_p + \|\mathbf{y}\|_p)\frac{\|\mathbf{x}+\mathbf{y}\|_p^p}{\|\mathbf{x}+\mathbf{y}\|_p}.$$

   This last inequality is called the inequality of Minkowski.

   (b) *Show that the application $f(\mathbf{x}) = \sqrt[p]{\sum_{i=1}^n |x_i|^p}$ is a norm.*

Correction.

1. Let $0 < p, q < \infty$ such that $\dfrac{1}{p} + \dfrac{1}{q} = 1$.

   (a) This equality holds because $\ln(a^p) = p\ln(a)$ and $\ln(ab) = \ln(a) + \ln(b)$ for all $a, b > 0$.

   (b) We apply the exponential function to the previous equality and use the definition of convexity:

   $$\begin{aligned} |ab| &= \exp\left(\frac{\ln(a^p)}{p} + \frac{ln(b^q)}{q}\right), \\ &\quad \downarrow \text{ previous question} \\ &\leq \frac{\exp(\ln(a^p))}{p} + \frac{\exp(\ln(b^q))}{q}, \\ &\quad \downarrow \text{ convexity of the exponential} \\ &= \frac{a^p}{p} + \frac{b^q}{q}. \end{aligned}$$

2. We want to prove now that : $\|\mathbf{xy}\|_1 \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q$ (*Hölder's inequality*)
   We consider $0 < p, q < \infty$ such that $\dfrac{1}{r} = \dfrac{1}{p} + \dfrac{1}{q}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

   (a) We first apply *Young's Inequality* using $p = p'r$ and $q = q'r$ so that $1 = \dfrac{1}{p'} + \dfrac{1}{q'}$, we have:

   $$|ab| \leq \frac{1}{p'}|a|^{p'} + \frac{1}{q'}|b|^{q'},$$

   where a and b are real numbers. We then set $a = x_i^r$ and $b = y_i^r$ to find the inequality.

(b) In the previous inequality we begin by using the hint and consider the special case where $r = 1$. In particular, we consider unit vectors, *i.e.* we will consider

$$a_i = \frac{x_i}{\|\mathbf{x}\|_p} \quad \text{and} \quad b_i = \frac{y_i}{\|\mathbf{y}\|_q}.$$

Thus

$$\begin{aligned}\sum_{i=1}^{d} \frac{|x_i y_i|}{\|\mathbf{y}\|_q \|\mathbf{x}\|_p} &\leq \frac{1}{p} \sum_{i=1}^{d} \frac{|x_i|^p}{\|\mathbf{x}\|_p} + \frac{1}{q} \sum_{i=1}^{d} \frac{|y_i|^q}{\|\mathbf{y}\|_q}, \\ &= \frac{1}{p} + \frac{1}{q}, \\ &= 1.\end{aligned}$$

Multiplying on both sides by $\|\mathbf{y}\|_q \|\mathbf{x}\|_p$ we get the result.

3. Let us now prove the triangle inequality for the L^p norm.

(a) We begin by writing the left hand side, the triangle inequality and *Hölder's Inequality* with $\frac{1}{q} = \frac{p-1}{p}$

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|_p^p &= \sum_{i=1}^{n} |x_i + y_i|^p, \\ &\quad \downarrow \text{ triangle inequality} \\ &\leq \sum_{i=1}^{n} (|x_i| + |y_i|)|x_i + y_i|^{p-1}, \\ &\quad \downarrow \text{ we separate in two sums} \\ &\leq \sum_{i=1}^{n} |x_i||x_i + y_i|^{p-1} + \sum_{i=1}^{n} |y_i||x_i + y_i|^{p-1}, \\ &\quad \downarrow \text{ Apply Hölder's inequality} \\ &\leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |x_i + y_i|^{(p-1)\times q}\right)^{\frac{1}{q}} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |x_i + y_i|^{(p-1)\times q}\right)^{\frac{1}{q}}, \\ &\leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{\frac{p-1}{p}} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{\frac{p-1}{p}}, \\ &\quad \downarrow \text{ factorization} \\ &\leq \left[\left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n} |i_i|^p\right)^{\frac{1}{p}}\right] + \left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{\frac{p-1}{p}}\end{aligned}$$

$$\|\mathbf{x}+\mathbf{y}\|_p^p \leq (\|x\|_p + \|y\|_p)\,\|\mathbf{x}+\mathbf{y}\|_p^{p-1}.$$

We divide on both sides by $\|\mathbf{x}+\mathbf{y}\|_p^{p-1}$ to have the result.

(b) The hardest part of the job has been during the previous questions.

1.2 ...

2 Analysis

2.1 Derivatives

Exercice 2.1. *Compute the first order derivative of the following functions*

1. $f(\mathbf{x}) = \exp(x_1x_2x_3) + x_1^2 + x_2 + \ln(x_3)$ *for all* $\mathbf{x} \in \mathbb{R}^3$.
2. *Given* $\mathbf{X} \in \mathscr{M}_{m,d}$ *and* $\mathbf{y} \in \mathbb{R}^m$ *let* $f(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\theta}$
3. *Let* $\mathbf{b} \in \mathbb{R}^d$ *and* $f(\mathbf{x}) = \ln\left(\sum_{j=1}^d \exp(x_j + b_j)\right)$ *for all* $\mathbf{x} \in \mathbb{R}^d$

Correction

We will just provide the answers here as you just need to apply the definition of the derivative.

1. For all $\mathbf{x} \in \mathbb{R}^3$, we have:

$$\frac{\partial f}{\partial x_1} = x_2x_3\exp(x_1x_2x_3) + 2x_1,\ \frac{\partial f}{\partial x_2} = x_1x_3\exp(x_1x_2x_3) + 1$$
$$\frac{\partial f}{\partial x_3} = x_1x_2\exp(x_1x_2x_3) + \frac{1}{x_3}.$$

2. The derivative is given by

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}).$$

3. The gradient of the function is given by:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} \frac{\exp(x_1 + b_j)}{\sum_{j=1}^d \exp(x_j + b_j)} \\ \vdots \\ \frac{\exp(x_d + b_j)}{\sum_{j=1}^d \exp(x_j + b_j)} \end{pmatrix}.$$

Exercice 2.2. *Calculate the first and second order derivatives of the following functions where $x, y \in \mathbb{R}$:*

1. $f(x, y) = 4x^2 + \exp(xy)$.
2. $f(x, y) = 7xy + \cos(x) + x^2 + 4y^2$.
3. $f(x, y) = 4(x - y)^2 + 5(x^2 - y)^2$.
4. $f(x, y) = \exp(x^2 + y^2)$.

Correction.

1.
$$\nabla_{x,y} f(x, y) = \begin{pmatrix} 8x + y\exp(xy) \\ x\exp(xy) \end{pmatrix}.$$
$$\nabla^2_{x,y} f(x, y) = \begin{pmatrix} 8 + y^2\exp(xy) & \exp(xy)(1 + y) \\ \exp(xy)(1 + y) & x^2\exp(xy) \end{pmatrix}$$

2.
$$\nabla_{x,y} f(x, y) = \begin{pmatrix} 7y - \sin(x) + 2x \\ 7x + 8y \end{pmatrix}.$$
$$\nabla^2_{x,y} f(x, y) = \begin{pmatrix} 2 - \cos(x) & 7 \\ 7 & 8 \end{pmatrix}$$

3.
$$\nabla_{x,y} f(x, y) = \begin{pmatrix} 20x^3 - 20xy + 8x - 8y \\ -10x^2 + 18y - 8x \end{pmatrix}.$$
$$\nabla^2_{x,y} = f(x, y) \begin{pmatrix} 60x^2 - 20y + 8 & -20x - 8 \\ 20x - 8 & 18 \end{pmatrix}$$

4.
$$\nabla_{x,y} f(x, y) = \begin{pmatrix} 2x\exp(x^2 + y^2) \\ 2y\exp(x^2 + y^2) \end{pmatrix}.$$
$$\nabla^2_{x,y} f(x, y) = \begin{pmatrix} \exp(x^2 + y^2)(2 + 4x^2) & 4xy\exp(x^2 + y^2) \\ 4xy\exp(x^2 + y^2) & \exp(x^2 + y^2)(2 + 4y^2) \end{pmatrix}$$

Exercice 2.3. *We can show that given a function f twice continuously differentiable, we always have:*

$$\frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right) = \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right),$$

i.e. the order of the derivatives does not matter.

Let us now consider the function f defined by for all $x, y \in \mathbb{R}$ by

$$f(x,y) = \frac{xy(x^2-y^2)}{x^2+y^2}, if\ (x,y) \neq (0,0)\ and\ f(x,y) = 0\ if\ (x,y) = 0.$$

1. *Compute* $\frac{\partial f}{\partial x}(0,y)$ *and* $\frac{\partial f}{\partial y}(x,0)$.
2. *Compute* $\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)(0,0)$ *and* $\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right)(0,0)$.
3. *What can we say about f ?*

Correction.

Let us now consider the function f defined by for all $x, y \in \mathbb{R}$ by

$$f(x,y) = \frac{xy(x^2-y^2)}{x^2+y^2}, if\ (x,y) \neq (0,0)\ and\ f(x,y) = 0\ if\ (x,y) = 0.$$

The two differential are equal to:

$$\frac{\partial f}{\partial x}(x,y) = \frac{[x^2+y^2][y(x^2-y^2)+2x^2y] - 2x(xy)(x^2-y^2)}{(x^2+y^2)^2},$$

and

$$\frac{\partial f}{\partial y}(x,y) = \frac{[x^2+y^2][x(x^2-y^2)-2xy^2] - 2y(xy)(x^2-y^2)}{(x^2+y^2)^2}.$$

1. Evaluated at the given point, we have: $\frac{\partial f}{\partial x}(0,y) = -y$ and $\frac{\partial f}{\partial y}(x,0) = x$.
2. According to the previous question: $\frac{\partial}{\partial y}\frac{\partial f}{\partial x}(0,0) = -1$ and $\frac{\partial}{\partial x}\frac{\partial f}{\partial y}(0,0) = 1$
3. The function f is not twice continuously differentiable at at the origin, that is the reason why the theorem does not hold.

2.2 Convex Set

Exercice 2.4. *Show that the unit ball $\mathcal{B}_2$, i.e. the set $\mathbf{x}$ such that $\|\mathbf{x}\|_2 \leq 1$, is a convex set.*

Correction.

We will simply apply the definition of convexity. Let us consider two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{B}_2$ and for all $t \in [0,1]$ we have to show that the point $\mathbf{z}(t)$ belongs to this set, *i.e.* its norm is less than 1.

Exercice 2.5. *Based on the definition of Convex set, try to prove the following statements*

1. *Given two convex sets C_1 and C_2, the intersection $C = C_1 \cap C_2$ is also convex.*
2. *A set C is convex if and only if its intersection with every straight line is convex .*
3. *The definition of convexity holds for more than two points (do it by induction)*

Correction.

1. Let us consider $\mathbf{x}$ and $\mathbf{y}$ in $C_1 \cap C_2$. Because C_1 is convex, for all $t \in [0,1]$, $t\mathbf{x} + (1-t)\mathbf{y} \in C_1$. Similarly, because C_2 is convex, for all $t \in [0,1]$, $t\mathbf{x} + (1-t)\mathbf{y} \in C_2$. Thus for all for all $t \in [0,1]$, $t\mathbf{x} + (1-t)\mathbf{y} \in C_1 \cap C_2$.

2. Suppose that C is convex and denote by D any straight line. Then the intersection $I = C \cap D$ is convex using the previous question and the fact that a straight line is convex.
   Conversely suppose that the intersection $I = C \cap D$ is convex. So for all points $\mathbf{x}, \mathbf{y}$ and all $t \in [0,1]$ the point $\mathbf{z}(t)$ defined by $\mathbf{z}(t) = t\mathbf{x} + (1-t)\mathbf{y} \in I$. It means that $z(t)$ belongs to both C and D, so it belongs to C. We can conclude that C is convex.

3. We have seen case when $k = 2$ (i.e. taking the combination of two points).Let us now consider that the definition holds for a given $k = n-1$ and let us show the definition holds for $k = n$.

   We consider any n points $\mathbf{x}_1, ..., \mathbf{x}_n \in C$ and the set of weights $\{\theta_1, ..., \theta_n \mid \sum_{i=1}^n \theta_i = 1\}$. We have to show that $\mathbf{y} = \sum_{i=1}^n \theta_i x_i \in C$.

   Without loss of generality we can consider that $\theta_n \neq 1$. The point $\mathbf{y}$ can be rewritten as follow:

$$\mathbf{y} = \theta_n x_n + (1-\theta_n)(\lambda_1 x_1 + ... + \lambda_{n-1} x_{n-1}),$$

   with $\lambda_i = \dfrac{\theta_i}{1-\theta_n}$. We have $\sum_{i=1}^{n-1} \lambda_i = 1$ so the point $\lambda_1 \mathbf{x}_1 + ... + \lambda_{n-1}\mathbf{x}_{n-1}$ belongs to C using the fact that the definition holds for $k = n-1$. And finally the point $\mathbf{y}$ belongs to C using the case $k = 2$.

Exercice 2.6. *Show that the following sets are convex*

1. *Let C be a set defined by:*

$$C = \{x \in \mathbb{R} \mid 3x^2 - 6x + 2 \leq 0\}$$

Show that C is convex.

2. *In general, consider the set C defined by:*

$$C = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}^{\mathbf{x}} - \mathbf{b}^T\mathbf{x} + \mathbf{c} \leq 0\},$$

where $\mathbf{A} \in S^d(\mathbb{R}), \mathbf{b} \in \mathbb{R}^d$ and $c \in \mathbb{R}$ is convex if $\mathbf{A}$ is a PSD matrix.

Correction.

1. Let C be a set defined by:

$$C = \{x \in \mathbb{R} \mid 3x^2 - 6x + 2 \leq 0\}.$$

We are going to deal with some tools in analysis. Let us consider the function $f : x \mapsto 3x^2 - 6x + 2$. This function is convex and the two roots of this function are

$$x_{\pm} = \frac{6 \pm \sqrt{12}}{6}.$$

Thus, for all $x \in [x_-; x_+]$, the function takes negative values and $[x_-; x_+]$ is convex.

2. For the set C defined:

$$C = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}^{\mathbf{x}} - \mathbf{b}^T\mathbf{x} + \mathbf{c} \leq 0\},$$

We have previously shown that a set is convex if and only if its intersection with any line segment is convex. So let us set $\mathbf{x} = \mathbf{u} + t\mathbf{v}$ where $t \in [0, 1]$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, we have:

$$\begin{aligned}
\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} + \mathbf{c} &= (\mathbf{u} + t\mathbf{v})^T\mathbf{A}(\mathbf{u} + t\mathbf{v}) - \mathbf{b}^T(\mathbf{u} + t\mathbf{v}) + \mathbf{c}, \\
&= (\mathbf{v}^T\mathbf{A}\mathbf{v})t^2 + (2\mathbf{u}^T\mathbf{A}\mathbf{v} - \mathbf{b}^T\mathbf{v})t + (\mathbf{c} - \mathbf{b}^T\mathbf{u} + \mathbf{u}^T\mathbf{A}\mathbf{u}), \\
&= \alpha t^2 + \beta t + \gamma.
\end{aligned}$$

The set $\{t \in \mathbb{R} \mid \alpha t^2 + \beta t + \gamma \leq 0\}$ is convex if $\alpha \geq 0$ (using the previous question). So this set is convex because $\mathbf{A}$ is PSD.

Exercice 2.7. *Show that the hyperbolic set $\{\mathbf{x} \in \mathbb{R}^d_+ \mid \prod_{i=1}^n x_i \geq 1\}$ is convex*[2].

[2] *Hint:* you can first show that, for all $x, y > 0$ and $\theta \in [0, 1]$ we have : $x^\theta y^{1-\theta} \leq \theta x + (1 - \theta)y$.

Correction.

We first begin by proving the hint using the convexity of the exponential function. For all $\theta \in [0,1]$ and for all $x, y \in \mathbb{R}_{++}$ we have:

$$x^\theta y^{1-\theta} = \exp(\theta \ln(x) + (1-\theta)\ln(y)) \leq \theta \exp(\ln(x)) + (1-\theta)\exp(\ln(y)) = \theta x + (1-\theta)y.$$

We now have to show that, for all $\mathbf{x} \in \mathbb{R}_+^d$ $\prod_{i=1}^n (\theta x_i + (1-\theta)y_i) \geq 1$. The result is a consequence of the hint:

$$\prod_{i=1}^n (\theta x_i + (1-\theta)y_i) \geq \prod_{i=1}^n x_i^\theta y_i^{1-\theta} = \left(\prod_{i=1}^n x_i\right)^\theta \left(\prod_{i=1}^n y_i\right)^{1-\theta} \geq 1.$$

2.3 Convex Function

Exercice 2.8. *Explain why the following functions are convex:*

1. $f : \mathbb{R}^d \to \mathbb{R}, f(\mathbf{x}) = \sum_{i=1}^n x_i^2$.
2. *For all* $x, y \in \mathbb{R}$, $f(x,y) = 3x^2 + (y-3)^2 + 4x + 6y + 5$
3. *For all* $x, y \in \mathbb{R}$, $f(x,y) = x^4 + 6y^4 + 2y^2 + 9x^2 + 3$.
4. *For all* $x, y \in \mathbb{R}$, $f(x,y) = 6x^2 + 5y^2 + 6xy$.
5. *For all* $x, y \in \mathbb{R}$, $f(x,y) = \exp(xy)$ *such that* $x > 1$ *and* $y < -1$.

Correction.

1. The first function is convex as the positive sum of convex functions.
2. It is convex as a positive sum of quadratic and linear convex functions.
3. It is convex as the sum of convex functions. Indeed, for all $n \in \mathbb{N}$, x^{2n} is convex.
4. You can compute the trace and the determinant of the Hessian matrix which is given by:

$$\begin{pmatrix} 12 & 6 \\ 6 & 10 \end{pmatrix}$$

These lasts are respectively equal to 22 and 84. So the function f is convex. It is also possible to directly compute the eigenvalues of the hessian matrix.

5. Let us compute the Hessian matrix also:

$$\begin{pmatrix} y^2\exp(xy) & (1+xy)\exp(xy) \\ (1+xy)\exp(xy) & x^2\exp(xy) \end{pmatrix}$$

The trace is equal to $(x^2+y^2)\exp(xy) > 0$ and the determinant is equal to $(-1-2xy)\exp(xy) > 0$ because $y < -1$ and $x > 1$ so $-2xy > 2$.

Exercice 2.9. *We are still working on convex functions.*

1. *Which of the following functions are convex?*
   (a) $f(x,y) = (1-x)^2 + 4(y-x^2)^2$.
   (b) $f(x,y) = (x+2y-7)^2 + (2x+y-5)^2$.
   (c) $f(x,y) = 2x^2 - 1.05x^4 + xy + y^2$.
   (d) $f(x,y) = \sin(x+y) + (x-y)^2 - 1.5x + 2.5y + 1$.
   (e) $f(x,y) = 10 + (x^2 - \cos(2\pi x)) + (y^2 - \cos(2\pi y))$.
2. *Find the local or global minima of the two first functions.*

Correction.

We will not provide the graph for each function in order to avoid a heavy document.

1. We can show that this function is not convex:

   The first term is a convex function in x. But second one is not a convex function. If we compute the Hessian matrix of the second part of the function, we have:

$$\begin{pmatrix} 48x^2 - 16y & -16x \\ -16x & 8 \end{pmatrix},$$

   and the trace can be negative for high values of y, which means that the function is not convex for all y.
2. This function is also convex as the sum of two convex quadratic terms.
3. This function is not convex because of the term $-1.05x^4$
4. This function is not convex because the function g defined by $g(x) = \sin(x)$ is not convex.

5. Same reason as before with the function cos.

Exercice 2.10 (A Result). *Try to prove the following result*[3]*:*

Proposition 2.1: Convexity and Restriction to a Segment

A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only its restriction to a line is always convex, i.e. if the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(t) = f(\mathbf{x} + t\mathbf{y})$ is convex, for all $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x} + t\mathbf{y}$ belongs to the domain of definition of f (f is concave if and only if g is concave).

Correction.

Exercice 2.11. *Let $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ be n vectors of $\mathbb{R}^d$, we denote by $\mathbf{X} \in \mathbb{R}^{n \times d}$ the matrix where each the i^{th} row is the vector x_i.*
We consider the matrix $G \in \mathbb{R}^{d \times d}$ defined by $\mathbf{G} = \mathbf{X}\mathbf{X}^T$. The matrix $\mathbf{G}$ is called the Gram Matrix.

Show that the Gram Matrix is a PSD matrix using the definition of a PSD matrix. In other words, the associated function is convex.

Correction.

The aim is to show that the eigenvalues of this matrix are non-negative. Let us consider $(\lambda, \mathbf{u})$ the couple eigenvalue, eigenvector of the matrix $\mathbf{G}$ such that $\mathbf{G}\mathbf{u} = \lambda\mathbf{u}$.
We now want to show that $\lambda \geq 0$. The previous equation can be rewritten as follows:

$$\mathbf{G}\mathbf{u} = \mathbf{X}\mathbf{X}^T\mathbf{u} = \lambda\mathbf{u}.$$

We then take the inner product with $\mathbf{u}$ on both sides:

$$\mathbf{u}^T\mathbf{X}\mathbf{X}^T\mathbf{u} = (\mathbf{X}^T\mathbf{u})^T(\mathbf{X}^T\mathbf{u}) = \|\mathbf{X}^T\mathbf{u}\|^2 = \lambda\|\mathbf{u}\|^2.$$

This last equality, combined with the positiveness of the norm imply the non-negativity of the eigenvalue λ.

Exercice 2.12. *Prove the following statements*

1. *Given two real convex functions f and g, the sum $f + g$ is also a convex function.*

[3] *Hint*: you just have to apply (write) the definition of convex function.

2. *If f is an increasing and real convex function, g a real convex function, then $f \circ g(\mathbf{x})$ is convex.*

3. *If f and g are two real convex functions, then h defined by $h(\mathbf{x}) = max\left(f(\mathbf{x}), g(\mathbf{x})\right)$ is also convex.*

Correction.

3 Optimization

3.1 Applications

Exercice 3.1 (A Quadratic function: Matyas function)**.** *We consider the function $f : [-10, 10]^2 \to \mathbb{R}$ defined by:*

$$f(x, y) = 0.26(x^2 + y^2) - 0.48yx$$

1. *Is the function f convex or not ?*

2. *Find the solution(s) of the equation $\nabla f(x, y) = 0$.*

3. *What is the global minimum of the function ?*

4. *We set $\mathbf{u}_0 = (x, y)^{(0)} = (1, 1)$, the initial point of the gradient descent with the optimal learning rate (or optimal step)*

   (a) *First recall what the gradient descent with optimal step consists of.*

   (b) *Compute $\mathbf{u}_1$ and $\mathbf{u}_2$.*

Exercice 3.2 (The Rosenbrock function)**.** *We consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by:*

$$f(x, y) = (1 - x)^2 + 10(y - x^2)^2.$$

1. *Is the function f convex or not ?*

2. *Find the solution(s) of the equation $\nabla f(x, y) = 0$.*

3. *What is the global minimum of the function ?*

4. *We set $\mathbf{u}_0 = (x, y)^{(0)} = (2, 2)$, the initial point of the gradient descent with a learning rate $\rho = 0.5$.*

   (a) *First recall what the gradient descent consists of.*

(b) Compute $\mathbf{u}_1$ *and* $\mathbf{u}_2$.

Exercice 3.3 (The Rastrigin function)**.** *We consider the function* $f : [-\pi, \pi]^2 \to \mathbb{R}$ *defined by:*

$$f(x, y) = 20 + (x^2 - 10\cos(2\pi x)) + (y^2 - 10\cos(2\pi y))$$

1. *Is the function* f *convex or not ?*
2. *Find the solution(s) of the equation* $\nabla f(x, y) = 0$.
3. *We assume that this function is positive for all* x, y. *What is the global minimum of the function ?*
4. *We set* $\mathbf{u}_0 = (x, y)^{(0)} = (2, 2)$, *the initial point of the gradient descent with a learning rate* $\rho = 0.5$.
   (a) *First recall what the gradient descent consists of.*
   (b) *Compute* $\mathbf{u}_1$ *and* $\mathbf{u}_2$.
5. *Are we sure that the algorithm will reach the global minimum ? Why ?*

Exercice 3.4 (A quadratic function)**.** *We consider the function* $f : \mathbb{R}^2 \to \mathbb{R}$ *defined by:*

$$f(x, y) = 7y^2 + 4x^2 - 5xy + 2x - 7y + 32.$$

1. *Is the function* f *convex or not ?*
2. *Find the solution(s) of the equation* $\nabla f(x, y) = 0$.
3. *What is the global minimum of the function ?*
4. *We set* $\mathbf{u}_0 = (x, y)^{(0)} = (1, 1)$, *the initial point of the Newton's Method*
   (a) *First recall what is the Newton's Method.*
   (b) *Calculate* $\mathbf{u}_1$ *and* $\mathbf{u}_2$.

Exercice 3.5. *We consider the function* $f : \mathbb{R}^2 \to \mathbb{R}$ *defined by:*

$$f(x, y) = 2x^2 - 1.05x^4 + \frac{x^6}{6} + xy + y^2.$$

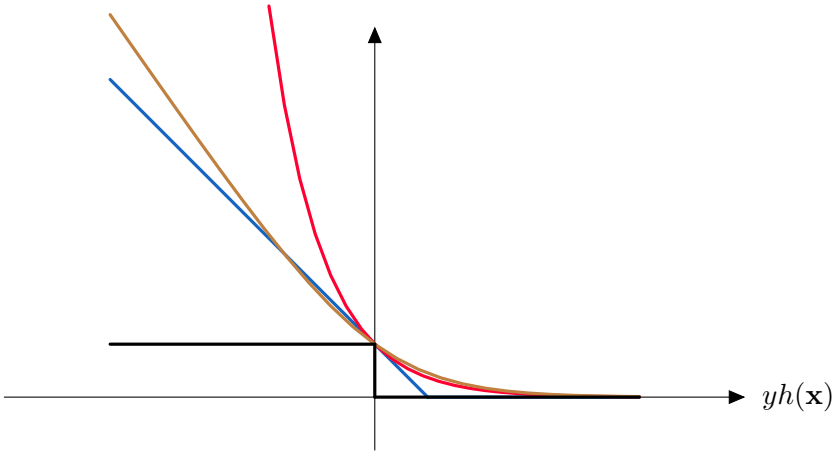
1. *Compute the Hessian Matrix.*
2. *What are the quantities we have to compute to prove that a* 2×2 *matrix is PSD? Compute them.*
3. *We assume that the function* f *is non-negative and non-convex, i.e* $f(x, y) \geq 0$. *Show that* $(0, 0)$ *is a solution of* $\nabla f(x, y) = 0$.
4. *What is the global minimum of the function ?*
5. *We set* $\mathbf{u}_0 = (x, y)^{(0)} = (1, 1)$, *the initial point of the gradient descent with a learning rate* $\rho = 0.5$.
   (a) *First recall what is the gradient descent.*
   (b) *Compute* $\mathbf{u}_1$ *and* $\mathbf{u}_2$.
6. *Are we sure that the algorithm will reach the global minimum? Why?*

3.2 Analysis of the Algorithm

4 About Supervised Algorithms

Exercice 4.1 (Loss functions)**.** *Show that the following loss functions are convex upper-bounds of the* $0 - 1$ *loss, where* $h_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x}$, $\mathbf{x}, \boldsymbol{\theta} \in \mathbb{R}^d$ *and* $y \in \{-1, +1\}$*:*

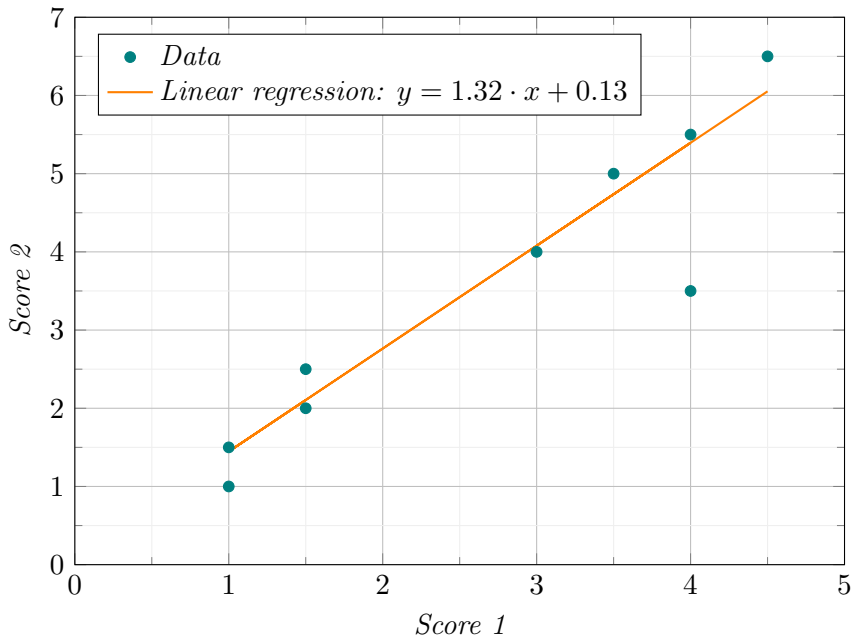
1. *the hinge loss defined by* $\ell(h_{\boldsymbol{\theta}}(\mathbf{x}), y) = \max(0, 1 - y h_{\boldsymbol{\theta}}(\mathbf{x}))$
2. *the exponential loss defined by* $\ell(h_{\boldsymbol{\theta}}(\mathbf{x}), y) = \exp(-y h_{\boldsymbol{\theta}}(\mathbf{x}))$
3. *the logistic loss defined by* $\ell(h_{\boldsymbol{\theta}}(\mathbf{x}), y) = \dfrac{1}{\ln(2)} \ln\left(1 + \exp(-y h_{\boldsymbol{\theta}}(\mathbf{x}))\right)$



Exercice 4.2 (The Linear Regression). *We consider the following dataset in which we aim to predict the score y obtained at a second exam accorind the schore obtained at the first exam x*

$\mathbf{x}$	4.0	3.0	3.5	1.0	1.5	1.0	1.5	4.0	3.5	4.5
$\mathbf{y}$	3.5	4.0	5.0	1.5	2.0	1.0	2.5	5.5	6.0	6.5

We focus on the gaussian linear regression model $Y = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, i.e. for all i, $y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$.



Exercice 4.3 (The Logistic Regression). *Let us consider $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$, where each $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y} = (y_1, y_2, ..., y_n) \in \{0, 1\}^n$ be respectively the matrix of the feature vector of n instances and their label.*

The logistic regression is used as a binary classification (it can be extended to multiclass classification problem) where the classifier returns the probability of an example to belong to a class of reference (let us say the class 1).

The logistic regression is based on the following model:

$$\ln\left(\frac{Pr(y=1 \mid \mathbf{x})}{Pr(y=0 \mid \mathbf{x})}\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_d x_d.$$

*In other words we estimate the log of the ratio of the probabilities of being in the class 1 with the one being in the class 0. This model is called a **LOGIT** model. The quantity $p(y = 1 \mid x)$ is called the posterior probability of being in the class 1.*

1. *Using the above equation, give an expression of* $Pr(y = 1 \mid \mathbf{x})$ *which depends on the vector of parameters* $\boldsymbol{\theta} \in \mathbb{R}^{d+1}$. *We will note* g *the obtained function, this function is called the* ***logistic*** *function.*

2. *Show that, for any* $\theta \in \mathbb{R}$, *we have* $\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) = g(\boldsymbol{\theta})(1 - g(\boldsymbol{\theta}))$.

3. *Study the convexity of function* g.

4. *What about* $\ln(g(\boldsymbol{\theta}))$ *?*

A classical method to estimate the parameters of a logistic regression model is to find the ones that maximize the likelihood of your data. The likelihood of an instance $\mathbf{x}_i$ *under this model is given by:*

$$Pr(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = g(w, \mathbf{x}_i)^{y_i}(1 - g(\boldsymbol{\theta}, \mathbf{x}_i))^{1-y_i}.$$

The law is the same as the Bernoulli law $\mathcal{B}(p)$ *with probability* $p = g(\boldsymbol{\theta}, \mathbf{x}_i)$ *where* p *is probability of being in the class* 1.

5. *We denote by* $\mathfrak{L}$ *the likelihood of our data and* ℓ *the log-likelihood of the data. Determine the expression of* $-\ell$, *the opposite of the log-likelihood.*

6. *Study the convexity of such problem.*

7. *Write the Newton's method to solve the minimization problem:*

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} -\ell(\boldsymbol{\theta}, \mathbf{X}, \mathbf{y})$$

Exercice 4.4.

Exercice 4.5.

Exercice 4.6.

5 Linear Algebra

Exercice 5.1. *Prove that hinge loss, logistic loss and exponential loss are convex upper bounds on 0-1 loss (consider that $h(x) = \beta^t \mathbf{x}$)*

Exercice 5.2. *Prove that this function $\mathbf{x} \longmapsto \|\mathbf{x}\|_2^2$ is convex.*

Exercice 5.3. *Prove that p-norm is convex. Recall the definition of p-norm $\|\mathbf{x}\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{\frac{1}{p}}$.*

Exercice 5.4. *Recall the assumptions of the Gaussian linear model.*

Correction1.1

Recall the *0-1 loss* :

$$\ell(h(\mathbf{x}), y) = \begin{cases} 1 & \text{if } h(x) \times y < 0 \\ 0 & \text{else} \end{cases} \tag{1}$$

Proof for *hinge loss* :

$$\begin{aligned} \ell(h(\mathbf{x}), y) &= \max(0, 1 - yh(\mathbf{x})) && (2)\\ &= \begin{cases} 1 - yh(\mathbf{x}) & \text{if } yh(\mathbf{x}) \leq 1 \\ 0 & \text{if } yh(\mathbf{x})) > 1 \end{cases} && (3)\\ &= \begin{cases} 1 - yh(\mathbf{x}) & \text{if } yh(\mathbf{x}) < 0 \\ 1 - yh(\mathbf{x}) & \text{if } 0 \leq yh(\mathbf{x}) \leq 1 \\ 0 & \text{if } yh(\mathbf{x}) > 1 \end{cases} && (4)\\ &\geq \begin{cases} 1 & \text{if } yh(\mathbf{x})) < 0 \\ 0 & \text{if } 0 \leq yh(\mathbf{x}) \leq 1 \\ 0 & \text{if } yh(\mathbf{x}) > 1 \end{cases} && (5)\\ &= \textit{0-1 loss} && (6) \end{aligned}$$

So *hinge loss* is an upper bound on *0-1 loss*, we'll see now it is convex.

Since $h(x) = \beta^t \mathbf{x}$, $1 - yh(x)$ is a linear function. The hinge loss is the maximum of two linear functions. And we know that :

1. Any linear function is convex.
2. The maximum of two convex functions is convex

Thus, *hinge loss* is convex with respect to β.

Proof for *logistic loss* :

$$\ell(h(\mathbf{x}), y) = \frac{1}{\ln(2)} \ln(1 + \exp(-yh(\mathbf{x}))) \tag{7}$$

if $yh(x) < 0$, then

$$\exp(-yh(\mathbf{x})) > 1 \Rightarrow \frac{\ln(1 + \exp(-yh(\mathbf{x})))}{\ln(2)} \tag{8}$$

if $yh(x) \geq 0$, then

$$0 < \exp(-yh(\mathbf{x})) \leq 1 \Rightarrow 0 < \frac{\ln(1 + \exp(-yh(\mathbf{x})))}{\ln(2)} \leq 1 \tag{9}$$

So *logistic loss* is an upper bound on *0-1 loss*. To prove its convexity with respect to β, we need to look at its hessian matrix. Firstly, we calculate the gradient :

$$\nabla f(\beta) = \frac{1}{\ln 2} \times \frac{-y \exp(-y\beta^t \mathbf{x})}{1 + \exp(-y\beta^t \mathbf{x})}\mathbf{x} = \frac{-y}{\ln 2} \times \frac{1}{1 + \exp(y\beta^t \mathbf{x})}\mathbf{x} \tag{10}$$

Then we calculate the hessian matrix :

$$\nabla^2 f(\beta) = \frac{y^2}{ln2} \times \frac{\exp(y\beta^t \mathbf{x})}{(1 + \exp(y\beta^t \mathbf{x}))^2}\mathbf{x}\mathbf{x}^t \tag{11}$$

Which is a positive semi-definite matrix (i.e.$\mathbf{x}\mathbf{x}^t$) multiplied by a positive scalar, is therefore also positive semi-definite. We can conclude that *logistic loss* is convex.

Proof for *exponential loss* :

$$\ell(h(\mathbf{x}), y) = \exp(-yh(\mathbf{x})) \tag{12}$$

Just like what we have done for *logistic loss*,

if $yh(x) < 0$, then

$$\exp(-yh(\mathbf{x})) > 1 \tag{13}$$

if $yh(x) \geq 0$, then

$$0 < \exp(-yh(\mathbf{x})) \leq 1 \tag{14}$$

The hessian matrix :

$$\nabla^2 f(\beta) = y^2 \exp(-y\beta^t \mathbf{x})\mathbf{x}\mathbf{x}^t \succeq 0 \tag{15}$$

Thus, *exponential loss* is an convex upper bound on *0-1 loss*.

Correction 1.2

The main idea doesn't change, we need to look at the hessian matrix, if it's positive semi-definite, we can conclude the convexity. The gradient of $l2$ norm squared is :

$$\nabla f(\mathbf{x}) = 2\mathbf{x} \tag{16}$$

And its hessian matrix (d is the dimension of x) :

$$\nabla^2 f(\mathbf{x}) = 2I_d \tag{17}$$

which is positive semi-definite. Thus, $\|\mathbf{x}\|_2^2$ is convex.

Correction 1.3

To prove that p-norm is convex, one may want to find the hessian matrix and verify if it is positive semi-definite. However, this approach requests a lot of derivations and the sign of ths hessian matrix is difficult to study. Instead, we look at the problem from another angle.

The Definition of a norm is:

Be V a Vectorspace, $\|\cdot\| : V \to \mathbb{R}$ is a norm : $\iff$

1. $\forall v \in V : \|v\| \geq 0$ and $\|v\| = 0 \iff v = 0$ (positive/definite)
2. $\forall v \in V, \lambda \in \mathbb{R} : |\lambda|\|v\| = \|\lambda v\|$ (absolutely scaleable)
3. $\forall v, w \in V : \|v + w\| \leq \|v\| + \|w\|$ (Triangle inequality)

The Definition of convex is:

$f : V \to \mathbb{R}$ is convex : $\iff \forall v, w \in V, \lambda \in [0, 1] : f(\lambda v + (1-\lambda)w) \leq \lambda f(v) + (1-\lambda)f(w)$

So using the Triangle inequality and the fact that the norm is absolutely scalable, you can see that every Norm is convex:

$$\|\lambda v + (1-\lambda)w\| \leq \|\lambda v\| + \|(1-\lambda)w\| = \lambda\|v\| + (1-\lambda)\|w\| \tag{18}$$

So by definition every norm is convex. But we need to show that the three requirements for a norm hold for p-norm. The first one is pretty easy. We can show the second one very quickly :

$$\|\lambda \mathbf{v}\|_p = \left(\sum_{i=1}^{d} |\lambda v_i|^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{d} |\lambda|^p |v_i|^p\right)^{\frac{1}{p}} = \left(|\lambda|^p \sum_{i=1}^{d} |v_i|^p\right)^{\frac{1}{p}} = |\lambda| \|\mathbf{v}\|_p \tag{19}$$

However, it is hard to show that the triangle inequality holds for p-norm, which means $\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$, AKA Minkowski Inequality. Because the Minkowski Inequality is a result of the Hölder inequality which is a result of Young's Inequality. We'll demonstrate it step by step.

Let p and q be two numbers satisfying

$$\frac{1}{p} + \frac{1}{q} > 1, with\ p > 1\ and\ q > 1 \tag{20}$$

We show firstly Young's inequality :

Let f be a continuous, and strictly increasing function on $[0, c]$ with $c > 0$. If $f(0) = 0, a \in [0, c]$, and $b \in [0, f(c)]$, then

$$\int_0^a f(x)\mathrm{d}x + \int_0^b f^{-1}(x)\mathrm{d}x \geq ab \tag{21}$$

where f^{-1} is the inverse function of f. Equality holds iff $b = f(a)$. Taking the particular function $f(x) = x^{p-1}$ gives the special case

$$\frac{a^p}{p} + \left(\frac{p-1}{p}\right) b^{p/(p-1)} \geq ab \tag{22}$$

Notice that $q = \frac{p}{p-1}$, then we have

$$\frac{a^p}{p} + \frac{b^q}{q} \geq ab, \quad where\ a, b \geq 0 \tag{23}$$

Let

$$a_i = \frac{|x_i|}{\|\mathbf{x}\|_p}, \; b_i = \frac{|y_i|}{\|\mathbf{y}\|_q} \tag{24}$$

Then by Young's inequality we can show Hölder inequality :

$$\sum_{i=1}^{d} \frac{|x_i|^p}{p\,\|\mathbf{x}\|_p^p} + \sum_{i=1}^{d} \frac{|y_i|^q}{q\,\|\mathbf{y}\|_q^q} \geq \sum_{i=1}^{d} \frac{|x_i y_i|}{\|\mathbf{x}\|_p \|\mathbf{y}\|_q} \tag{25}$$

$$\frac{1}{p} + \frac{1}{q} \geq \frac{\|\mathbf{xy}\|_1}{\|\mathbf{x}\|_p \|\mathbf{y}\|_q} \tag{26}$$

$$\|\mathbf{x}\|_p \|\mathbf{y}\|_q \geq \|\mathbf{xy}\|_1 \tag{27}$$

Finally, we show Minkowski inequality :

$$\|\mathbf{x}+\mathbf{y}\|_p^p = \sum_{i=1}^{d} |x_i + y_i|^p \tag{28}$$

$$= \sum_{i=1}^{d} |x_i + y_i|\,|x_i + y_i|^{p-1} \tag{29}$$

$$\leq \sum_{i=1}^{d} (|x_i| + |y_i|)\,|x_i + y_i|^{p-1} \tag{30}$$

$$= \sum_{i=1}^{d} |x_i|\,|x_i + y_i|^{p-1} + |y_i|\,|x_i + y_i|^{p-1} \tag{31}$$

$$\text{(Hölder inequality)} \leq \|\mathbf{x}\|_p \left(\sum_{i=1}^{d} |x_i + y_i|^{(p-1)q}\right)^{1/q} + \|\mathbf{y}\|_p \left(\sum_{i=1}^{d} |x_i + y_i|^{(p-1)q}\right)^{1/q} \tag{32}$$

$$= (\|\mathbf{x}\|_p + \|\mathbf{y}\|_p)\,\|\mathbf{x}+\mathbf{y}\|_p^{p/q} \tag{33}$$

The last equality is obtained by $q = \frac{p}{p-1}$. So we have shown that

$$\|\mathbf{x}+\mathbf{y}\|_p^p \leq (\|\mathbf{x}\|_p + \|\mathbf{y}\|_p)\,\|\mathbf{x}+\mathbf{y}\|_p^{p/q} \tag{34}$$

$$\|\mathbf{x}+\mathbf{y}\|_p^{p-\frac{p}{q}} \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p \tag{35}$$

$$\|\mathbf{x}+\mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p \tag{36}$$

Since $p - \frac{p}{q} = 1$, we have exactly the Minkowski inequality !

Correstion 1.4

1. Homoscedasticity: The variance of residual is the same for any value of X.
2. Independence: Observations are independent of each other.
3. Normality: For any fixed value of X, Y is normally distributed.