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 :*

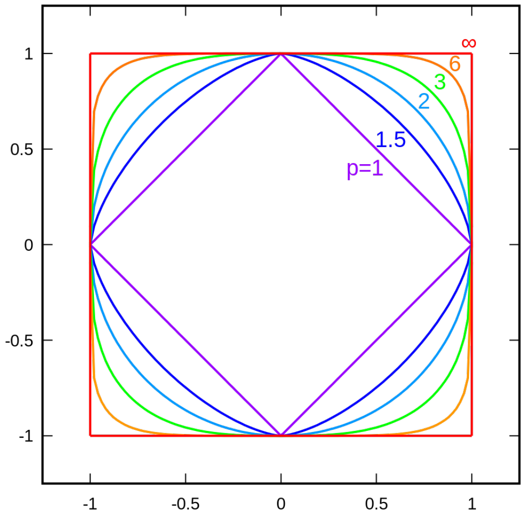
1. $f(\mathbf{x}, \mathbf{y}) = x_1y_1 + x_2y_2$.
2. $f(\mathbf{x}, \mathbf{y}) = x_1y_1 + x_2y_2 - x_3 + y_3$.
3. $f(\mathbf{x}, \mathbf{y}) = x_1y_1 + 2x_2y_2 + 3x_3y_3$.
4. $f(\mathbf{x}, \mathbf{y}) = x_1^2y_1^2 + x_2^2y_2^2 + x_3^3y_3^3$.

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* $\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\|_F\|\mathbf{x}\|_2$ *where* $\mathbf{A} \in \mathcal{M}_{m,d}(\mathbb{R})$ *and* $\mathbf{x} \in \mathbb{R}^d$.
4. *Show that* $\|\mathbf{A}\mathbf{B}\|_F \leq \|\mathbf{A}\|_F\|\mathbf{B}\|_F$ *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:*
$$\mathbf{A} = \begin{pmatrix} 1 & -3 \\ -3 & 1 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 3 & -2 & 3 \\ -2 & 1 & -2 \\ -3 & 2 & 3 \end{pmatrix}$$

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

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 $\frac{1}{p} + \frac{1}{q} = 1$*

   (a) *Show that $\ln(ab) = \frac{\ln(a^p)}{p} + \frac{\ln(b^q)}{q}$ 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 $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$ 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 $x_i = \frac{x_i}{\|\mathbf{x}\|_p^p}$ and $y_i = \frac{y_i}{\|\mathbf{y}\|_q^q}$

3. *Let us now prove 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.*

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* $f(\mathbf{x}) = \ln\left(\sum_{j=1}^d \exp\left(x_j + b_j\right)\right)$ *for all* $\mathbf{x} \in \mathbb{R}^d$

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. $f(x, y) = \exp(x^2 + y^2)$.

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

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.*

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)*

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.

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].

2.3 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* $y < -1$.

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.*

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).*

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

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

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.

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.*
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* $h(\mathbf{x}) = max\,(f(\mathbf{x}), g(\mathbf{x}))$ *is also convex.*

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?*

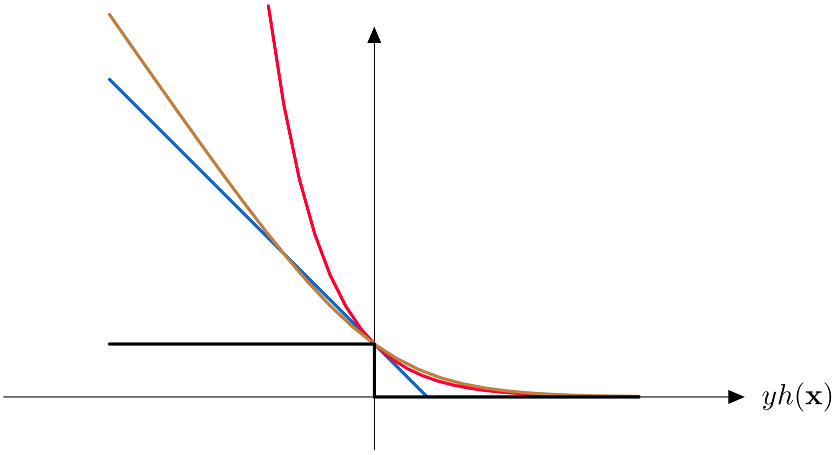
3.2 Analysis of the Algorithm

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* $\ell(h_{\boldsymbol{\theta}}(\mathbf{x}), y) = \max(0, 1 - yh_{\boldsymbol{\theta}}(\mathbf{x}))$
2. *the exponential loss defined by* $\ell(h_{\boldsymbol{\theta}}(\mathbf{x}), y) = \exp(-yh_{\boldsymbol{\theta}}(\mathbf{x}))$

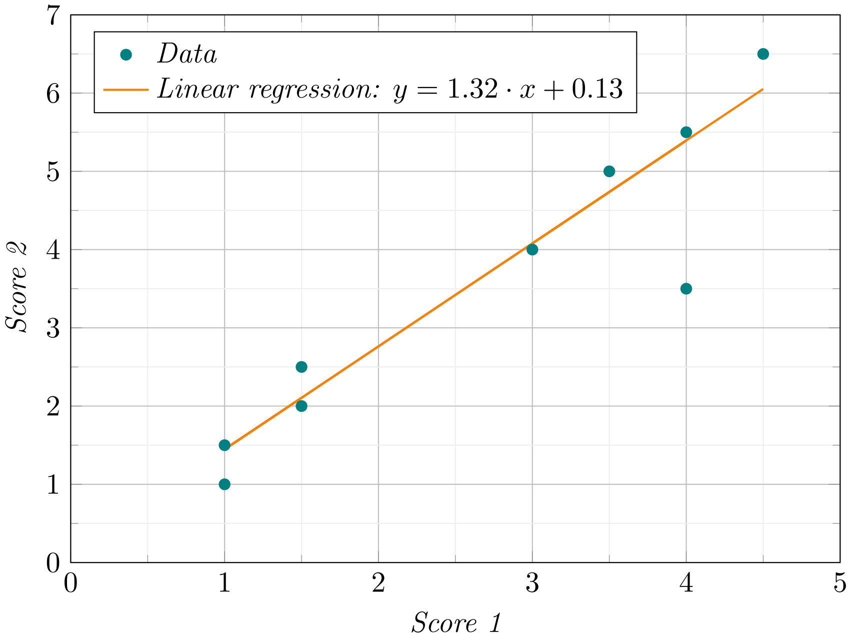
3. *the logistic loss defined by* $\ell(h_{\boldsymbol{\theta}}(\mathbf{x}), y) = \dfrac{1}{\ln(2)} \ln\left(1 + \exp(-y h_{\boldsymbol{\theta}}(\mathbf{x}))\right)$



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.