

# Introduction to Machine learning

Sylvain Le Corff

1. Mathematical framework

2. Logistic regression

1. Mathematical framework

2. Logistic regression

## Supervised Learning Framework

- **Input** measurement  $\mathbf{X} \in \mathcal{X}$  (often  $\mathcal{X} \subset \mathbb{R}^d$ ).
- **Output** measurement  $Y \in \mathcal{Y}$ .
- The joint distribution of  $(\mathbf{X}, Y)$  is **unknown**.
- $Y \in \{1, \dots, M\}$  (classification) or  $Y \in \mathbb{R}^m$  (regression).
- A **predictor** is a measurable function  $f : \mathcal{X} \rightarrow \mathcal{Y}$ .

## Training data

- i.i.d. with the same distribution as  $(\mathbf{X}, Y)$ :

$$\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}.$$

## Goal

- Construct a **good** predictor  $\hat{f}_n$  from the training data.
- Need to specify the meaning of good.

## Loss function

- $\ell(Y, f(\mathbf{X}))$ : the goodness of the prediction of  $Y$  by  $f(\mathbf{X})$ .
- **Prediction** loss:  $\ell(Y, f(\mathbf{X})) = \mathbf{1}_{Y \neq f(\mathbf{X})}$ .
- **Quadratic** loss:  $\ell(Y, \mathbf{X}) = \|Y - f(\mathbf{X})\|_2^2$ .

## Risk function

- Risk measured as the average loss:

$$\mathcal{R}(f) = \mathbb{E}[\ell(Y, f(\mathbf{X}))].$$

- **Prediction** loss:  $\mathbb{E}[\ell(Y, f(\mathbf{X}))] = \mathbb{P}(Y \neq f(\mathbf{X}))$ .
- **Quadratic** loss:  $\mathbb{E}[\ell(Y, f(\mathbf{X}))] = \mathbb{E}[\|Y - f(\mathbf{X})\|_2^2]$ .
- **Beware**: As  $\hat{f}_n$  depends on  $\mathcal{D}_n$ ,  $\mathcal{R}(\hat{f}_n)$  is a random variable!

## Bayes classifier

The **Bayes classifier**  $g^*$  is defined as:

$$g^*(X) = \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1|X) > \mathbb{P}(Y = 0|X), \\ 0 & \text{otherwise.} \end{cases}$$

Equivalently,

$$g^*(X) = \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1|X) > 1/2, \\ 0 & \text{otherwise,} \end{cases}$$

## Lemma

*For any classification rule  $g : \mathbb{R}^d \rightarrow \{0, 1\}$ , one has*

$$\mathcal{R}(g^*) \leq \mathcal{R}(g).$$

In practice **we do not know the conditional law of  $Y$  given  $X$** .  
Several solutions to overcome this issue.

## Fully parametric modeling.

Estimate the law of  $(X, Y)$  and use the **Bayes formula** to deduce an estimate of the conditional law of  $Y$ : *LDA/QDA, Naive Bayes...*

## Parametric conditional modeling.

Estimate the conditional law of  $Y$  by a **parametric** law: *linear regression, logistic regression, Feed Forward Neural Networks...*

## Nonparametric conditional modeling.

Estimate the conditional law of  $Y$  by a **non parametric** estimate: *kernel methods, nearest neighbors...*

1. Mathematical framework

2. Logistic regression



- ▶ In regression with  $\mathcal{X} = \mathbb{R}^d$ , the linear model is the **parametric reference model**.
- ▶ This model makes the assumption that the regression function is linear: for  $1 \leq i \leq n$

$$Y = X^\top \beta^\star + \varepsilon,$$

with

$$\mathbb{E}[\varepsilon|X] = 0 \quad \text{and} \quad \mathbb{V}[\varepsilon|X] = \sigma^2.$$

- ▶ Here, estimating the regression function is equivalent to estimate  $\beta^\star \in \mathbb{R}^d$ .

**Finite dimensional parametric model**

- ▶ The least squares estimates, i.e. the optimal solution of

$$\min_{\beta \in \mathbb{R}^d} \widehat{\mathcal{R}}_n(\beta) = \min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2$$

with  $Y \in \mathbb{R}^n$ ,  $X \in \mathbb{R}^{n \times d}$  is given, **if  $X$  has full rank**, by

$$\widehat{\beta}^{LS} = (X^\top X)^{-1} X^\top Y.$$

- ▶ The regression function  $m^* : x \mapsto x^\top \beta^*$  is estimated by  $\widehat{m}_n^{LS} x \mapsto x^\top \widehat{\beta}^{LS}$ .
- ▶ Under some technical assumptions (see lectures of past year)

$$\mathbb{E}[\widehat{\beta}^{LS}] = \beta^* \quad \text{and} \quad \mathbb{V}(\widehat{\beta}^{LS}) = (X^\top X)^{-1} \sigma^2.$$

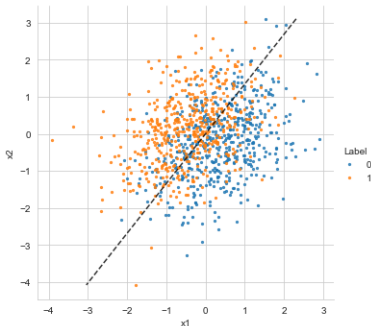
- ▶ We deduce that, see [sylvainlc.github.io](https://github.com/sylvainlc),

$$\mathbb{E} \left[ \|\widehat{\beta} - \beta\|_2^2 \right] = O \left( \frac{1}{n} \right) \quad \text{and} \quad \mathbb{E} \left[ (\widehat{m}_n^{LS}(x) - m^*(x))^2 \right] = O \left( \frac{1}{n} \right)$$

In the LDA case, the classification rule is of the form:

$$g^*(x) = 1 \Leftrightarrow \langle w, x \rangle + b \geq 0,$$

where  $w$  and  $b$  depends on the model parameters.



- Relax the Gaussian assumption ? (logistic model, SVM).
- Design nonlinear classification rules ? (kernels, neural networks).

- ▶ One of the most widely used classification algorithm.
- ▶ It models the distribution of  $Y$  given  $X$ . For  $y \in \{0, 1\}$

$$\mathbb{P}(Y = 1|X) = \sigma(X^T w + b)$$

where  $w \in \mathbb{R}^d$  is a vector of model weights and  $b \in \mathbb{R}$  is the intercept, and where  $\sigma : z \mapsto (1 + e^{-z})^{-1}$  is the **sigmoid** function:

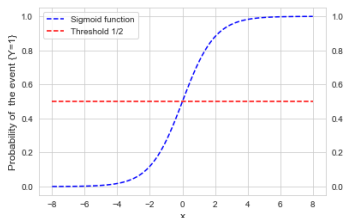


Figure: The sigmoid function

- ▶ The sigmoid is a modelling choice to map  $\mathbb{R} \rightarrow [0, 1]$  (to model a probability).
- ▶ We could also consider

$$\mathbb{P}(Y = 1|X) = F(X^\top w + b)$$

for any distribution function  $F$ .

- ▶ Another popular choice is the Gaussian distribution

$$F(z) = \mathbb{P}(\mathcal{N}(0, 1) \leq z),$$

which leads to another loss called **probit**.

- ▶ In the case of the sigmoid, one has

$$\mathbb{P}(Y = 1|X) = \frac{\exp(b + w^\top X)}{1 + \exp(b + w^\top X)} = \frac{1}{1 + \exp(-(b + w^\top X))}$$
$$\mathbb{P}(Y = 0|X) = \frac{1}{1 + \exp(b + w^\top X)}$$

- ▶ However, the sigmoid choice has the following nice interpretation: an easy computation leads to

$$\log \left( \frac{\mathbb{P}(Y = 1|X)}{\mathbb{P}(Y = 0|X)} \right) = X^\top w + b.$$

- ▶ This quantity is called the **log-odd ratio**.

- ▶ Therefore, this model makes the assumption that (the logit transformation of) the probability  $p(X) = \mathbb{P}(Y = 1|X)$  is linear:

$$\text{logit}(p(X)) := \log\left(\frac{p(X)}{1 - p(X)}\right) = X^\top w + b.$$

- ▶ Note that

$$\mathbb{P}(Y = 1|X) \geq \mathbb{P}(Y = 0|X)$$

if and only if

$$X^\top w + b \geq 0.$$

This is a **linear classification rule**, linear w.r.t. the considered features  $x$ !

## Theorem

Consider that the logit-transformation is linear with parameters  $(b^*, w^*)$ :

$$\text{logit}(p(X)) := \log \left( \frac{p(X)}{1 - p(X)} \right) = f^*(X) = X^\top w^* + b^*.$$

Then  $f^*$  is a minimizer of the risk:  $f \mapsto \mathbb{E} [\log (1 + \exp(-Yf(X)))]$  over all affine functions and

$$g^*: x \in \mathbb{R}^d \mapsto \begin{cases} 1 & \text{if } f^*(x) > 0 \\ 0 & \text{otherwise} \end{cases}$$

is a Bayes classifier.



**Parametric model** for the conditional law of  $Y$  given  $X$ :  $\mathbb{P}_{w,b}(Y|X)$ .

Data  $(X_i, Y_i)_{1 \leq i \leq n}$  i.i.d with the same distribution as  $(X, Y)$ .

Compute estimators  $\hat{w}$  and  $\hat{b}$  by maximum likelihood estimation.

Or equivalently, **minimize the minus log-likelihood**:

$$(w, b) \mapsto -n^{-1} \log \mathbb{P}_{w,b}(Y_{1:n}|X_{1:n}).$$

More generally, when a model is used

$$\text{Goodness-of-fit} = -\log \text{likelihood}$$

The log function is used mainly since averages are easier to study (and compute) than products.

$\rightarrow \{(X_i, Y_i)\}_{1 \leq i \leq n}$  are **i.i.d.** with the same distribution as  $(X, Y)$ .

**Likelihood:**

$$\begin{aligned}\prod_{i=1}^n \mathbb{P}(Y_i|X_i) &= \prod_{i=1}^n \sigma(\langle w, X_i \rangle + b)^{Y_i} (1 - \sigma(\langle w, X_i \rangle + b))^{1-Y_i}, \\ &= \prod_{i=1}^n \sigma(\langle w, x_i \rangle + b)^{Y_i} \sigma(-\langle w, X_i \rangle - b)^{1-Y_i}\end{aligned}$$

and the **normalized negative loglikelihood** is

$$f(w, b) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, \langle w, X_i \rangle + b).$$

Compute  $\hat{w}_n$  and  $\hat{b}_n$  as follows:

$$(\hat{w}_n, \hat{b}_n) \in \operatorname{argmin}_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \left( -Y_i (X_i^\top w + b) + \log(1 + e^{X_i^\top w + b}) \right).$$

→ It is an **average of losses**, one for each sample point.

→ It is a **convex and smooth problem**.

Using the **logistic loss** function

$$\ell : (y, y') \mapsto \log(1 + e^{-yy'})$$

yields

$$(\hat{w}_n, \hat{b}_n) \in \operatorname{argmin}_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, \langle w, X_i \rangle + b).$$

Assume for now that the intercept is 0. Then, the likelihood is,

$$L_n(w) = \prod_{i=1}^n \left( \frac{e^{X_i^T w}}{1 + e^{X_i^T w}} \right)^{Y_i} \left( \frac{1}{1 + e^{X_i^T w}} \right)^{1-Y_i} = \prod_{i=1}^n \left( \frac{e^{X_i^T w Y_i}}{1 + e^{X_i^T w}} \right).$$

And the **negative log-likelihood** is

$$\ell_n(w) = -\log(L_n(w)) = \sum_{i=1}^n \left( -Y_i X_i^T w + \log(1 + e^{X_i^T w}) \right).$$

## Derivatives

$$\begin{aligned} \frac{\partial (\log(L_n(w)))}{\partial w_j} &= \sum_{i=1}^n \left( Y_i X_{ij} - \frac{x_{ij} e^{X_i^T w}}{(1 + e^{X_i^T w})} \right) \\ &= \sum_{i=1}^n X_{ij} (Y_i - \sigma(\langle w, X_i \rangle)). \end{aligned}$$

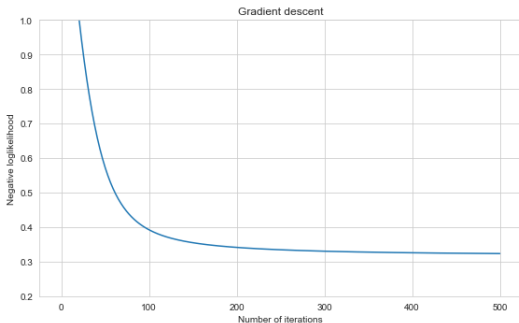
The negative loglikelihood

$$\ell_n(w) = -\log(L_n(w)) = \sum_{i=1}^n \left( -Y_i X_i^T w + \log(1 + e^{X_i^T w}) \right).$$

is minimized using a gradient descent algorithm.

Starting with an **initial estimate**  $w^{(0)}$ , for all  $k \geq 1$ , set

$$w^{(k)} = w^{(k-1)} - \eta_k \nabla \ell_n(w^{(k-1)}).$$



The **gradient of the negative loglikelihood** is,

$$\nabla \ell_n(w) = - \sum_{i=1}^n Y_i X_i + \sum_{i=1}^n \frac{\exp(\langle X_i, w \rangle)}{1 + \exp(\langle X_i, w \rangle)} X_i .$$

On the other hand, for all  $1 \leq i \leq n$  and all  $1 \leq j \leq d$ ,

$$\partial_j \left( \frac{\exp(\langle X_i, w \rangle)}{1 + \exp(\langle X_i, w \rangle)} X_i \right) = \frac{\exp(\langle X_i, w \rangle)}{(1 + \exp(\langle X_i, w \rangle))^2} X_{ij} X_i ,$$

where  $X_{ij}$  is the  $j$ th component of  $X_i$ .

Then, the **Hessian matrix** is

$$H_n(w) = \sum_{i=1}^n \frac{\exp(\langle X_i, w \rangle)}{(1 + \exp(\langle X_i, w \rangle))^2} X_i X_i^T .$$

## Assumptions

→  $\widehat{w}_n \rightarrow w^*$  almost surely.

→ There exists a continuous and nonsingular function  $H$  such that  $n^{-1}H_n(w)$  converges to  $H(w)$ , uniformly in a ball around  $w^*$ .

For all  $t \in \mathbb{R}^d$ , using a Taylor expansion,

$$\mathbb{E} \left[ \exp \left( -\frac{1}{\sqrt{n}} \langle t, \nabla \ell_n(w^*) \rangle \right) \right] \rightarrow_{n \rightarrow \infty} \exp \left( \frac{1}{2} t^T H(w^*) t \right).$$

Therefore,

$$-\nabla \ell_n(w^*) / \sqrt{n} \Rightarrow \mathcal{N}(0, H(w^*)).$$

On the other hand, by Slutsky lemma,

$$\sqrt{n}(\widehat{w}_n - w^*) \Rightarrow \mathcal{N}(0, H(w^*)^{-1}).$$

→  $\sqrt{n}(\hat{w}_j - w_j^*)$  converges in distribution to a centered Gaussian random variable with variance  $(H(w^*)^{-1})_{jj}$ .

Almost surely,  $\hat{\sigma}_{n,j}^2 = (nH_n(\hat{w}_n)^{-1})_{jj} \rightarrow_{n \rightarrow \infty} (H(w^*)^{-1})_{jj}$ .

Then,

$$\sqrt{\frac{n}{\hat{\sigma}_{n,j}^2}}(\hat{w}_{n,j} - \beta_j^*) \rightarrow_{n \rightarrow \infty} \mathcal{N}(0, 1).$$

An asymptotic confidence interval  $\mathcal{I}_{n,\alpha}$  of level  $1 - \alpha$  is then

$$\mathcal{I}_{n,\alpha} = \left[ \hat{w}_{n,j} - z_{1-\alpha/2} \sqrt{\frac{\hat{\sigma}_{n,j}^2}{n}}, \hat{w}_{n,j} + z_{1-\alpha/2} \sqrt{\frac{\hat{\sigma}_{n,j}^2}{n}} \right],$$

where  $z_{1-\alpha/2}$  is the quantile of order  $1 - \alpha/2$  of  $\mathcal{N}(0, 1)$ .