

# Remise à niveau - Statistique

## Introduction au ML / à la classification

Claire Boyer

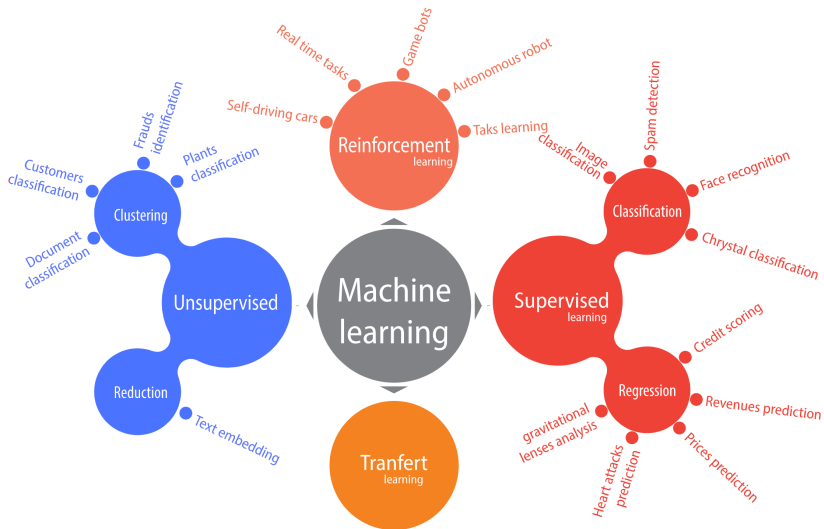
1. Context

2. Discriminant analysis

3. Logistic regression

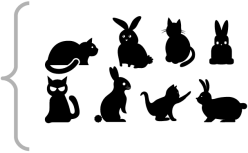
ML develops generic methods for solving different types of problems:

- ▶ **Supervised** learning  
Goal: learn from labeled examples
- ▶ **Unsupervised** learning  
Goal: learn from data alone, extract structure in the data
- ▶ **Reinforcement** learning  
Goal: learn by exploring the environment (e.g. games or autonomous vehicle)



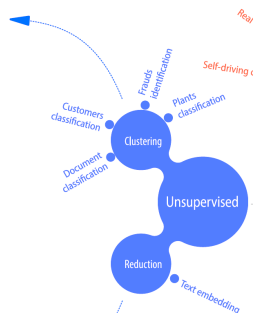

**Clustering :**  
Finding Common Relationships

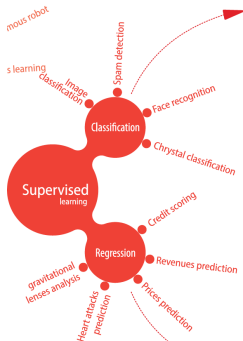
What is the relationship between these data ?



**Reduction :**  
Reduce the number of dimensions

Simplify while keeping meaning





source: fidle-cnrs

## Classification :

Predict qualitative informations



This is a cat



This is a rabbit



Tell me,  
what is it ?



## Régression :

Predict quantitative informations



150 K€



400 K€



120 K€



100 K€



Tell me,  
what's the  
price ?



- ▶ **Supervised learning**: given a training sample  $(X_i, Y_i)_{1 \leq i \leq n}$ , the goal is to “learn” a **predictor**  $f_n$  such that

$$\underbrace{f_n(X_i) \simeq Y_i}_{\text{prediction on training data}}$$

and above all

$$\underbrace{f_n(X_{\text{new}}) \simeq Y_{\text{new}}}_{\text{prediction on test (unseen) data}}$$

Often

- ▶ **(classification)**  $X \in \mathbb{R}^d$  and  $Y \in \{-1, 1\}$
- ▶ **(regression)**  $X \in \mathbb{R}^d$  and  $Y \in \mathbb{R}$

- ▶ **Loss function in general:**  $\ell(Y, f(X))$  measures the goodness of the prediction of  $Y$  by  $f(X)$
- ▶ **Examples:**
  - ▶ **(classification)** Prediction loss:  $\ell(Y, f(X)) = 1_{Y \neq f(X)}$
  - ▶ **(regression)** Quadratic loss:  $\ell(Y, f(X)) = |Y - f(X)|^2$
- ▶ The performance of a predictor  $f$  in regression is usually measured through the risk

$$\text{Risk}(f) = \mathbb{E} \left[ \ell(Y_{\text{new}}, f(X_{\text{new}})) \right]$$

- ▶ A minimizer  $f^*$  of the risk is called a **Bayes predictor**



- ▶ We want to construct a predictor with a small **risk**
- ▶ The **distribution** of the data is in general **unknown**, so is the **risk**
- ▶ Instead, given some **training** samples  $(X_1, Y_1), \dots, (X_n, Y_n)$ , find the best predictor  $f$  that minimizes the **empirical risk**

$$\hat{\mathcal{R}}_n(f) := \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)).$$

- ▶ **Learning** means retrieving information from training data by constructing a predictor that should have good performance on new data

In regression ( $\mathcal{Y} = \mathbb{R}$ ), the **quadratic cost** is often used:

$$\begin{aligned}\ell : \mathbb{R} \times \mathbb{R} &\rightarrow \mathbb{R}^+ \\ (y, y') &\mapsto (y - y')^2.\end{aligned}$$

so that the **quadratic risk** for a machine or regression function  $m : \mathcal{X} \rightarrow \mathbb{R}$ :

$$\mathcal{R}(f) := \mathbb{E} [(Y - f(X))^2].$$

Its Bayes predictor  $f^*$  is  $f^*(x) := \mathbb{E}[Y|X = x]$

Indeed, for any  $f$ , one has

$$\mathcal{R}(f^*) = \mathbb{E} [(Y - f^*(X))^2] \leq \mathbb{E} [(Y - f(X))^2] =: \mathcal{R}(f).$$

**Problem**  $f^*$  is generally unknown, so we have to find an estimate  $\hat{f}_n(x)$  of  $f(x)$  such that  $\hat{f}_n(x) \simeq f^*(x)$

- ▶ Setting: the output can only take 2 values ( $Y \in \{0, 1\}$ )
- ▶ Note that the distribution of  $(X, Y)$  is entirely characterized by  $(\mu_X, r)$  with  $\mu$  the marginal distribution of  $X$  and  $r$  is the **regression function** of  $Y$  on  $X$ . More precisely, for all  $A \in \mathcal{B}(\mathbb{R}^d)$ ,  $\mu_X(A) = \mathbb{P}(X \in A)$ , and

$$r(x) = \mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x)$$

where the last equality comes from  $Y \in \{0, 1\}$

- ▶ There is a classification error (or misclassification) as soon as the prediction  $\hat{Y} \neq Y$

## The error probability or the risk for a classification rule

For a prediction function/rule  $f : \mathbb{R}^d \rightarrow \{0, 1\}$ ,

$$\mathcal{R}(f) = \mathbb{E}[\mathbb{1}_{f(X) \neq Y}] = \mathbb{P}(f(X) \neq Y).$$

## Does an optimum exist?

The Bayes predictor  $f^*$  is

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

the equality favoring 0 by convention. Equivalently,

$$f^*(x) = \begin{cases} 1 & \text{if } r(x) > 1/2, \\ 0 & \text{otherwise,} \end{cases}$$

## Lemma

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

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

**Exercise:** Prove it.

## The Bayes risk

$$\mathcal{R}^* := \mathcal{R}(f^*) = \inf_{f: \mathbb{R}^d \rightarrow \{0,1\}} \mathbb{P}(f(X) \neq Y).$$

**Exercise:** Show that

1.  $\mathcal{R}^* = 1 - \mathbb{E} [\mathbb{1}_{r(X) > 1/2} r(X) + \mathbb{1}_{r(X) \leq 1/2} (1 - r(X))],$
2.  $\mathcal{R}^* = \mathbb{E} [\min(r(X), 1 - r(X))] = \frac{1}{2} - \frac{1}{2} \mathbb{E} |2r(X) - 1|,$
3.  $\mathcal{R}^* = 0 \iff Y = \varphi(X)$  with probability one.

## Problem

$f^*$  depends on the **unknown** distribution of  $(X, Y)$

- ▶ We can use an  $n$ -sample, i.e.  $n$  i.i.d. copies of  $(X, Y)$  to estimate  $f^*$

1. Context

2. Discriminant analysis

3. Logistic regression

## Definition

Let  $\mu \in \mathbb{R}^d$ ,  $\Sigma$  be a positive definite matrix. We write  $X \sim \mathcal{N}(\mu, \Sigma)$  when the Lebesgue density of  $X$  is

$$\begin{aligned} x \in \mathbb{R}^d &\mapsto |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)} \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}, \end{aligned}$$

where  $|\Sigma|$  is the determinant of  $\Sigma$ . In addition, we have

$$\mathbb{E}X = \mu, \quad \mathbb{V}(X) = \Sigma,$$

where  $\mathbb{V}(X)$  is the covariance matrix of  $X$ .

Question: what are the MLEs for the expectation and the covariance matrix of a Gaussian sample?

## Proposition

Let  $\mu^* \in \mathbb{R}^d$ ,  $\Sigma^*$  be a positive definite matrix and  $\{X_1, \dots, X_n\}$  be a sample i.i.d. according to  $\mathcal{N}(\mu^*, \Sigma^*)$ .

Then

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$

and

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})(X_i - \hat{\mu})^\top$$

are maximum likelihood estimators (MLEs) respectively of  $\mu^*$  and  $\Sigma^*$ .



- ▶  $(X, Y) \in \mathbb{R}^d \times \{1, \dots, C\}$  be a pair of r.v.
- ▶  $Y$  is a label characterizing the class of  $X$ .
- ▶ **Goal:** computing the Bayes classifier when each class  $c \in \{1, \dots, C\}$  is **normally distributed**, i.e. there exists a positive definite matrix  $\Sigma_c$  and a vector  $\mu_c \in \mathbb{R}^d$  such that

$$X|Y = c \sim \mathcal{N}(\mu_c, \Sigma_c).$$

Recall: a Bayes classifier

For multiclass

$$\forall x \in \mathbb{R}^d: \quad f^*(x) \in \operatorname{argmax}_{c \in [C]} \mathbb{P}(Y = c | X = x).$$

## Proposition

*Let us assume that each class is normally distributed and let  $\pi_c = \mathbb{P}(Y = c)$  be class prior probabilities, for all  $c \in [C]$ . Then, a Bayes classifier  $f^*$  is defined by:  $\forall x \in \mathbb{R}^d$*

$$f^*(x) \in \operatorname{argmax}_{c \in [C]} \log(\pi_c) - \frac{1}{2} \log |\Sigma_c| - \frac{1}{2} (x - \mu_c)^\top \Sigma_c^{-1} (x - \mu_c).$$

Proof: Compute the log-ratio of the conditional probabilities.

# Linear discriminant analysis (LDA)

- ▶ In the case of  $C=2$  classes
- ▶ LDA model

$$X|Y = c \sim \mathcal{N}(\mu_c, \Sigma), \quad c = 1, 2$$

- ▶ With equal covariance

## Proposition

Let  $\pi_c = \mathbb{P}(Y = c)$  be class prior probabilities, for  $c \in \{1, 2\}$ ,

$$h: x \in \mathbb{R}^d \mapsto (\mu_1 - \mu_2)^\top \Sigma^{-1} x$$

$$b = \frac{1}{2}(\mu_2^\top \Sigma^{-1} \mu_2 - \mu_1^\top \Sigma^{-1} \mu_1) + \log \left( \frac{\pi_1}{\pi_2} \right).$$

Then, a Bayes classifier is

$$f^*: x \in \mathbb{R}^d \mapsto \begin{cases} 1 & \text{if } h(x) + b > 0 \\ 2 & \text{otherwise.} \end{cases}$$

- ▶ Note that the function  $h(x) + b$  is linear in  $x$ .
- ▶ This is a linear classifier!

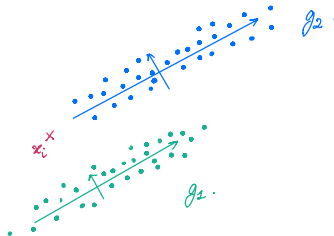
## What happens when $\pi_1 = \pi_2$

- ▶ if  $\pi_1 = \pi_2$ , we have:

$$f^*(x) = 1$$

$$\iff (x - \mu_1)^\top \Sigma^{-1} (x - \mu_1) < (x - \mu_2)^\top \Sigma^{-1} (x - \mu_2),$$

- ▶  $\pi_1 = \pi_2$  if and only if  $x$  is closer to  $\mu_1$  than  $\mu_2$  with respect to the Mahalanobis distance ruled by  $\Sigma$ .



- ▶ Each class is normally distributed
- ▶ But with **different covariances**

## Proposition

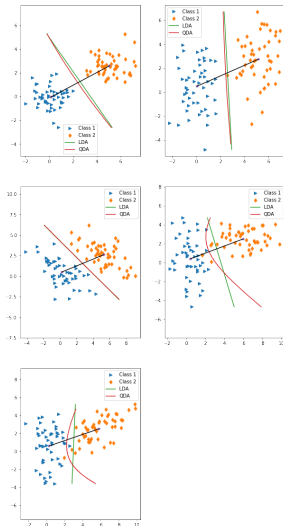
Let  $\pi_c = \mathbb{P}(Y = c)$  be class prior probabilities, for all  $c \in \{1, 2\}$ , and let us denote

$$h: x \in \mathbb{R}^d \mapsto \frac{1}{2}x^\top(\Sigma_2^{-1} - \Sigma_1^{-1})x + (\mu_1^\top \Sigma_1^{-1} - \mu_2^\top \Sigma_2^{-1})x$$
$$b = \frac{1}{2}(\mu_2^\top \Sigma_2^{-1} \mu_2 - \mu_1^\top \Sigma_1^{-1} \mu_1) - \frac{1}{2} \log \left( \frac{|\Sigma_1|}{|\Sigma_2|} \right) + \log \left( \frac{\pi_1}{\pi_2} \right).$$

Then, a Bayes classifier is

$$f^*: x \in \mathbb{R}^d \mapsto \begin{cases} 1 & \text{if } h(x) + b > 0 \\ 2 & \text{otherwise.} \end{cases}$$

Proof: Left as an exercise.



**Figure:** Comparison of linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) on different simulated datasets (Gaussian classes with potentially different covariance matrices).

1. Context

2. Discriminant analysis

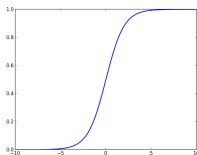
3. Logistic regression

- ▶ One of the most widely used classification algorithm.
- ▶ Logistic model is the "brother" of the linear model in the context of binary classification ( $\mathcal{Y} = \{-1, 1\}$ ).
- ▶ We want to explain the label  $Y$  based on  $X$ , we want to "regress"  $Y$  on  $X$ .
- ▶ It models the distribution of  $Y|X$ . For  $y \in \{-1, 1\}$

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

where  $w \in \mathbb{R}^d$  is a vector of model weights and  $b \in \mathbb{R}$  is the intercept, and where  $\sigma$  is the **sigmoid** function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$





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

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

for any distribution function  $F$ .

- ▶ Another popular choice is the Gaussian distribution function

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

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

- In the case of the sigmoid, one has

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

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

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

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

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

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

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

- ▶ Note that

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

iff

$$x^T w + b \geq 0.$$

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

## Theorem

*Let us consider that  $C = 2$  and that the logit-transformation is linear with parameters  $(b^*, w^*)$ . Let  $f^*: x \in \mathbb{R}^d \mapsto b^* + (w^*)^\top x$ .*

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

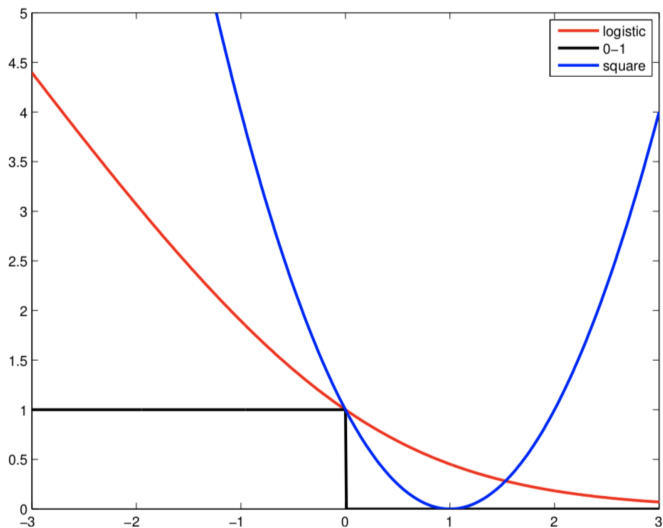
$$g^*: x \in \mathbb{R}^d \mapsto \begin{cases} +1 & \text{if } f^*(x) > 0 \\ -1 & \text{otherwise} \end{cases}$$

*is a Bayes classifier.*

# The logistic regression in a nutshell

29 / 31

- This is a linear classifier chosen for the logistic loss!



- ▶ We have a model for  $Y|X$
- ▶ Data  $(X_i, Y_i)$  is assumed 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.
- ▶ More generally, when a model is used

Goodness-of-fit = -log likelihood

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

By introducing the logistic loss function

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

then

$$\hat{w}, \hat{b} \in \operatorname{argmin}_{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R}}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, x_i^T w + b).$$

- ▶ It is a convex and smooth problem
- ▶ Many ways to find an approximate minimizer
- ▶ Efficient convex optimization algorithms

**Remark.** Careful when **separable data**, i.e.,  $\exists(b_0, w_0)$  such that

$$\forall i = 1, \dots, n, \quad Y_i(w_0^T X_i + b_0) > 0,$$

then there is no minimizer of the negative log-likelihood!