# Bayesian Decision Theory

Gilles Gasso

INSA Rouen - ASI Departement
Laboratory LITIS

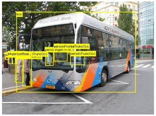
October 12, 2019

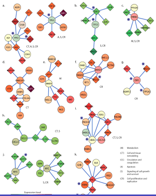
# Plan

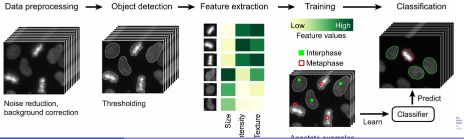
# Classification problems

## Applications

- Object detection
- Protein classification, Medical imaging
- Intrusion detection, fraud detection
- ...



# Classification: taxonomy and formulation

- Data : $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$
- $\boldsymbol{x}$ : sample belonging to the space $\mathcal{X}$ ($\mathcal{X} = \mathbb{R}^d$)
- $y \in \mathcal{Y}$ : associated label. $\mathcal{Y}$ : discrete finite set

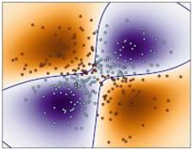
## Taxonomy

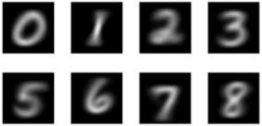
- Binary : $\mathcal{Y} = \{-1, 1\}$ ou $\mathcal{Y} = \{0, 1\}$

  Anomaly detection, Fraud detection ...
- Multi-class : $\mathcal{Y} = \{1, 2, \cdots, K\}$

  Objects or speakers recognition ...
- Multi-label : $\mathcal{Y} = 2^{\{1,2,\cdots,K\}}$

  Recognition of the topic of documents ...



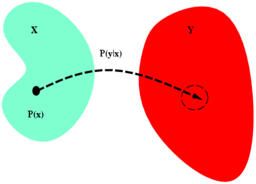


# Classification: taxonomy and formulation

- Data : $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$
- $\boldsymbol{x}$ : sample belonging to the space $\mathcal{X}$ ($\mathcal{X} = \mathbb{R}^d$)
- $y \in \mathcal{Y}$ : associated label. $\mathcal{Y}$ : discrete finite set

## Principle

- Learn a mathematical function $f : \mathcal{X} \to \mathcal{Y}$ able to predict the label of $\boldsymbol{x}$
- Example: $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$




## Different approaches and algorithms

- Bayesian decision, Logistic regression
- SVM, k-nearest neighbors, random forest, XGBoost ...

## This lecture: Bayesian Decision Theory

- Probabilistic decision-making
- $(\boldsymbol{x}, y)$ is considered as a random variable

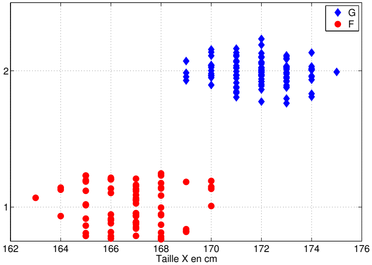
## Pre-requisites

Basics of probability and statistics

# Introduction

## Example

- Let $x \in \mathbb{R}$ the height of a student
- Given $x$ predict the gender of the person: F (class $\mathcal{C}_1$) or M (class $\mathcal{C}_2$ )




Problem

- Find a statistical model (of each class)
- Infer a classification rule

# Formulation

## Elements of solution

- Consider a given height $x$ (ex : $x = 170$ cm).
- Compute the probabilities $\Pr(\mathcal{C}_1/x)$ and $\Pr(\mathcal{C}_2/x)$,
  - $\Pr(\mathcal{C}_1/x)$: probability that the student is a F **knowing** $x$
  - $\Pr(\mathcal{C}_2/x)$: probability that the student is a G **knowing** $x$
- Assign $x$ to the class with the highest probability,

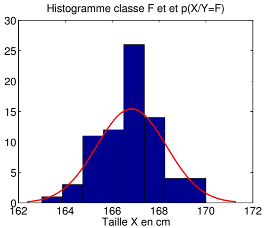
$$x \text{ is assigned to } \mathcal{C}_1 \text{ if } \Pr(\mathcal{C}_1/x) > \Pr(\mathcal{C}_2/x)$$

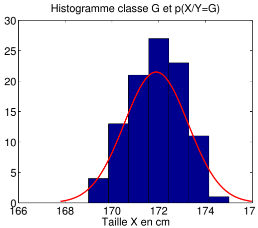
- How to compute the probability $\Pr(\mathcal{C}_k/x)$ ?

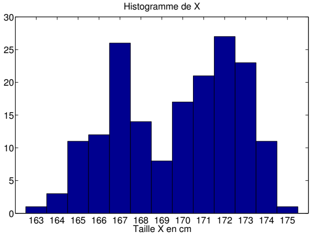
# The training data

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $x_{11}$ | $x_{12}$ | $x_{13}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 | Tota |
| $\mathcal{C}_1 = F$ | 1 | 3 | 11 | 12 | 26 | 14 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 75 |
| $\mathcal{C}_2 = G$ | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 13 | 21 | 27 | 23 | 11 | 1 | 100 |
| Total | 1 | 3 | 11 | 12 | 26 | 14 | 8 | 17 | 21 | 27 | 23 | 11 | 1 | 175 |

Table: training Data statistics




## Notations

- $n_{ik}$: the number of persons of the class $\mathcal{C}_k$ ($k = 1, 2$) with height $x_i$ ($i$= 1 to 13)
- $c_i$: number of persons with height equals to $x_i$
- $N_k$: cardinality of class $\mathcal{C}_k$ with $N = \sum_k N_k$.

# Notions of probability (1)

Random variables : $X$ : size of a person and $\mathcal{C}$ : the category ( $\mathcal{C}_1 = \text{F}$ and $\mathcal{C}_2 = \text{G}$)

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $x_{11}$ | $x_{12}$ | $x_{13}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 | Total |
| $\mathcal{C}_1 = \text{F}$ | 1 | 3 | 11 | 12 | 26 | 14 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 75 |
| $\mathcal{C}_2 = \text{M}$ | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 13 | 21 | 27 | 23 | 11 | 1 | 100 |
| Total | 1 | 3 | 11 | 12 | 26 | 14 | 8 | 17 | 21 | 27 | 23 | 11 | 1 | 175 |

## Joint probability $\text{p}(X, \mathcal{C})$

What is the probability that a student is $x_i = 170$cm tall **and** is a F ?

- Solution: $\text{p}(x_i, \mathcal{C}_1) = \frac{4}{175}$ . Note: we also have $\text{p}(x_i, \mathcal{C}_2) = \frac{13}{175}$
- Joint probability: $\text{p}(x_i, \mathcal{C}_k) = \frac{n_{ik}}{N}$

# Notions of probability (1)

## Joint probability $\mathrm{p}(X, \mathcal{C})$

What is the probability that a student is $x_i = 170$cm tall **and** is a F ?

- Solution: $\mathrm{p}(x_i, \mathcal{C}_1) = \frac{4}{175}$ . Note: we also have $\mathrm{p}(x_i, \mathcal{C}_2) = \frac{13}{175}$
- Joint probability: $\mathrm{p}(x_i, \mathcal{C}_k) = \frac{n_{ik}}{N}$

## Marginal distribution $p_x(X)$

What is the probability that a student is $x_i = 170$cm?

- $p_X(x_i) = \frac{17}{175} = \frac{4}{175} + \frac{13}{175}$: probability to have $(x_i = 170, \mathcal{C}_1)$ or $(x_i = 170, \mathcal{C}_2)$
- Marginal distribution: $p_x(x_i) = \frac{c_i}{N}$

## Probabilities sum : $p_x(\boldsymbol{x}_i) = \sum_k \mathrm{p}(\boldsymbol{x}_i, \mathcal{C}_k)$

# Notions of probability (2)

$x$: height and $\mathcal{C}$: class ($\mathcal{C}_1$ = F and $\mathcal{C}_2$ = M)

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $x_{11}$ | $x_{12}$ | $x_{13}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 | Total |
| $\mathcal{C}_1$ = F | 1 | 3 | 11 | 12 | 26 | 14 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 75 |
| $\mathcal{C}_2$ = M | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 13 | 21 | 27 | 23 | 11 | 1 | 100 |
| Total | 1 | 3 | 11 | 12 | 26 | 14 | 8 | 17 | 21 | 27 | 23 | 11 | 1 | 175 |

## Prior probability $\Pr(\mathcal{C})$

**Without knowing her height** what is the probability that a student is F?

- Solution: $\Pr(\mathcal{C}_1) = \frac{75}{175}$.
- Similarly we have $\Pr(\mathcal{C}_2) = \frac{100}{175}$. Note: $\Pr(\mathcal{C}_1) + \Pr(\mathcal{C}_2) = 1$
- Prior probability of class $\mathcal{C}_k$: $\Pr(\mathcal{C}_k) = \frac{N_k}{N}$
- The sum of prior probabilities is equal to 1

# Notions of probability (2)

$x$: height and $\mathcal{C}$: class ($\mathcal{C}_1$ = F and $\mathcal{C}_2$ = M)

| | $x_1$ 163 | $x_2$ 164 | $x_3$ 165 | $x_4$ 166 | $x_5$ 167 | $x_6$ 168 | $x_7$ 169 | $x_8$ 170 | $x_9$ 171 | $x_{10}$ 172 | $x_{11}$ 173 | $x_{12}$ 174 | $x_{13}$ 175 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{C}_1$ = F | 1 | 3 | 11 | 12 | 26 | 14 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 75 |
| $\mathcal{C}_2$ = M | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 13 | 21 | 27 | 23 | 11 | 1 | 100 |
| Total | 1 | 3 | 11 | 12 | 26 | 14 | 8 | 17 | 21 | 27 | 23 | 11 | 1 | 175 |

## Conditional probability $p(X/\mathcal{C})$

What is the probability that a student is $x_i = 170$cm **knowing that** she is a F ?

- Solution : $p(x_i/\mathcal{C}_1) = \frac{4}{75}$.
- Note: $p(x_i/\mathcal{C}_1) = \frac{4}{175} \times \frac{175}{75} = \frac{\mathrm{p}(x_i, \mathcal{C}_1)}{\Pr(\mathcal{C}_1)}$
- Conditional probability: $p(x_i/\mathcal{C}_k) = \frac{n_{ik}}{N_k}$

Product rule: $\mathrm{p}(\boldsymbol{x}_i, \mathcal{C}_k) = p(\boldsymbol{x}_i/\mathcal{C}_k)\Pr(\mathcal{C}_k)$

# Decision function

Recall: to decide if a person is F or M knowing $x$, we only need to compare $\Pr(\mathcal{C}_1/x)$ and $\Pr(\mathcal{C}_2/x)$

## Posterior probability $\Pr(\mathcal{C}_k/x)$

- Note: $\mathrm{p}(\mathcal{C}_k, x_i) = \mathrm{p}(x_i, \mathcal{C}_k)$.
- Apply the product rule gives $\mathrm{p}(\mathcal{C}_k, x_i)$ : $\mathrm{p}(\mathcal{C}_k, x_i) = \Pr(\mathcal{C}_k/x_i)p_X(x_i)$
- Also it holds $\mathrm{p}(\mathcal{C}_k, x_i) = p(x_i/\mathcal{C}_k)\Pr(\mathcal{C}_k)$, hence we deduce

## Bayesian Rule

$$\Pr(\mathcal{C}_k/\boldsymbol{x}_i) = \frac{p(\boldsymbol{x}_i/\mathcal{C}_k) \times \Pr(\mathcal{C}_k)}{p_X(\boldsymbol{x}_i)}$$

# Decision function

Recall: to decide if a person is F or M knowing $x$, we only need to compare $\Pr(\mathcal{C}_1/x)$ and $\Pr(\mathcal{C}_2/x)$

## Application

What is the assigned label to a student **knowing** $x_i = 170$?

- $\Pr(\mathcal{C}_1/x_i) = \frac{\frac{4}{75} \times \frac{75}{175}}{\frac{17}{175}} \Rightarrow \Pr(\mathcal{C}_1/x_i) = \frac{4}{17}$
- $\Pr(\mathcal{C}_2/x_i) = \frac{\frac{13}{100} \times \frac{100}{175}}{\frac{17}{175}} \Rightarrow \Pr(\mathcal{C}_2/x_i) = \frac{13}{17}$
- $\Pr(\mathcal{C}_2/x_i) > \Pr(\mathcal{C}_1/x_i) \Longrightarrow x$ belongs to $\mathcal{C}_2$

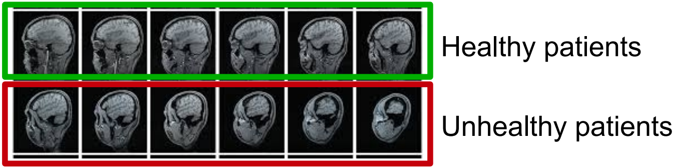
Sum of posterior probabilities is equal to 1 i.e. $\sum_k \Pr(\mathcal{C}_k/\boldsymbol{x}) = 1$

# Bayesian Decision Theory

# Example

## Medical application

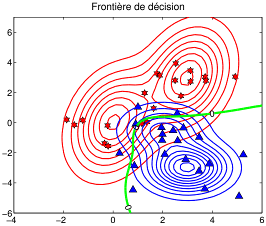
- Inputs: MRI for healthy and non-healthy patients
- Goal: predict based on his MRI if the patient is healthy (no treatment) or unhealthy (treatment)




## Issue

- A bad decision can be catastrophic → associate a cost to each decision
- Take the decision with minimal cost

# Problem formulation for binary classification



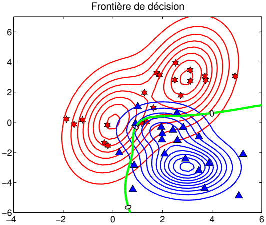

## Training data

- Two classes $\mathcal{C}_1, \mathcal{C}_2$
- Data: $\{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \{\mathcal{C}_1, \mathcal{C}_2\}\}_{i=1}^N$

## Statistical model

- Each class $\mathcal{C}_k$ is characterized by its
  - prior probability $\Pr(\mathcal{C}_k)$ and its conditional distribution $p(x/\mathcal{C}_K)$
- Marginal distribution of data: $p_x(x) = \sum_{k=1}^{2} p(x/\mathcal{C}_k)\Pr(\mathcal{C}_k)$

# Problem formulation for binary classification




## Training data

- Two classes $\mathcal{C}_1, \mathcal{C}_2$
- Data: $\{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \{\mathcal{C}_1, \mathcal{C}_2\}\}_{i=1}^N$

## Cost of a decision

$\ell_{jk}$: cost of predicting class class $\mathcal{C}_j$ to $\boldsymbol{x}$ knowing that $\boldsymbol{x} \in \mathcal{C}_k$

| Decision $a$ \ Truth | Class $\mathcal{C}_1$ | Class $\mathcal{C}_2$ |
|---|---|---|
| Class $\mathcal{C}_1$ | $\ell_{11}$ | $\ell_{12}$ |
| Class $\mathcal{C}_2$ | $\ell_{21}$ | $\ell_{22}$ |

- Right dec. : $\ell_{jk} = 0$ if $j = k$
- Wrong dec. : $\ell_{jk} = 100$ if $j \neq k$

## Problem to solve

Find the classification rule that minimizes the average cost

# More formally

We seek to find a decision function $D$

$$D: \begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathcal{A} \\ \boldsymbol{x} & \longmapsto & a \end{array} \qquad D(\boldsymbol{x}) = a \; (a = \mathcal{C}_1 \text{ or } \mathcal{C}_2)$$

- Classification error
  - Erroneous prediction: $D(\boldsymbol{x}) = \mathcal{C}_1$ while the true label is $\mathcal{C}_2$
  - or the other way around
- Conditional risk

$$\begin{aligned} R(a = \mathcal{C}_1|\boldsymbol{x}) &= \ell_{11}\Pr(\mathcal{C}_1/\boldsymbol{x}) + \ell_{12}\Pr(\mathcal{C}_2/\boldsymbol{x}) \\ R(a = \mathcal{C}_2|\boldsymbol{x}) &= \ell_{21}\Pr(\mathcal{C}_1/\boldsymbol{x}) + \ell_{22}\Pr(\mathcal{C}_2/\boldsymbol{x}) \end{aligned}$$

- What decision to make for $\boldsymbol{x}$ ?

# Bayes' rule

## Overall principle : Minimal risk decision

- predict class $D(\boldsymbol{x}) = \mathcal{C}_1$ to $\boldsymbol{x}$ if $R(a = \mathcal{C}_1|\boldsymbol{x}) < R(a = \mathcal{C}_2|\boldsymbol{x})$
- otherwise predict $D(\boldsymbol{x}) = \mathcal{C}_2$

## Extension to multi-class classification

- Bayes' rule straightforwardly generalizes to multi-class classification problem $\mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_K\}$.
- For all $x \in \mathcal{X}$ Bayes' rule writes:

$$D_{Bayes}(\boldsymbol{x}) = \text{argmin}_{j=1\cdots K} R(a_j|\boldsymbol{x})$$

with $R(a_j|\boldsymbol{x}) = \sum_{k=1}^{K} \ell_{jk} \Pr(\mathcal{C}_k/\boldsymbol{x}) \quad \forall j = 1, \cdots, K$

Concretely : decide $a_r$ if $R(a_r|\boldsymbol{x}) < R(a_j|\boldsymbol{x}) \quad (\forall a_j \neq a_r)$

# Winner takes it all

Let consider 0-1 cost :

$$\ell_{jk} = \begin{cases} 0 & \text{if} \quad j = k \quad \text{(no error)} \\ 1 & \text{if} \quad j \neq k \quad \text{(error)} \end{cases}$$

Conditional risks become

$$\begin{aligned} R(a = \mathcal{C}_1|\boldsymbol{x}) &= \ell_{11}\Pr(\mathcal{C}_1/\boldsymbol{x}) + \ell_{12}\Pr(\mathcal{C}_2/\boldsymbol{x}) = \Pr(\mathcal{C}_2/\boldsymbol{x}) \\ R(a = \mathcal{C}_2|\boldsymbol{x}) &= \ell_{21}\Pr(\mathcal{C}_1/\boldsymbol{x}) + \ell_{22}\Pr(\mathcal{C}_2/\boldsymbol{x}) = \Pr(\mathcal{C}_1/\boldsymbol{x}) \end{aligned}$$

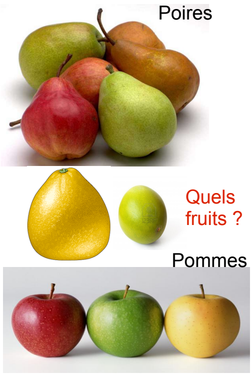
Maximum posterior probability rule

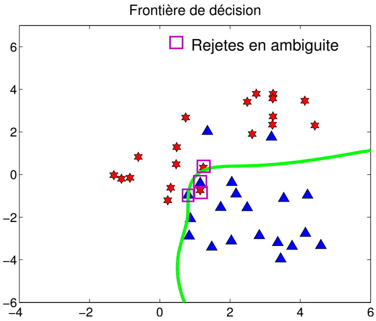
- Predict $D(\boldsymbol{x}) = \mathcal{C}_1$ if $R(a = \mathcal{C}_1|\boldsymbol{x}) < R(a = \mathcal{C}_2|\boldsymbol{x})$
- $\Rightarrow \Pr(\mathcal{C}_1/\boldsymbol{x}) > \Pr(\mathcal{C}_2/\boldsymbol{x})$ or $\Pr(\mathcal{C}_1/\boldsymbol{x}) > 1/2$

Interpretation: predict the class with maximum posterior probability

# Reject option (1)

- Intuition: if the decision may be ambiguous, instead of predicting a class, do not make a decision $\longrightarrow$ call for the reject option



# Reject option (2)

- Binary classification case
  Let $a_3$ be the reject option

- Conditional risks

$$
\begin{aligned}
R(\mathcal{C}_1|\boldsymbol{x}) &= \ell_{11}\mathrm{Pr}(\mathcal{C}_1/\boldsymbol{x}) + \ell_{12}\mathrm{Pr}(\mathcal{C}_2/\boldsymbol{x}) \\
R(\mathcal{C}_2|\boldsymbol{x}) &= \ell_{21}\mathrm{Pr}(\mathcal{C}_1/\boldsymbol{x}) + \ell_{22}\mathrm{Pr}(\mathcal{C}_2/\boldsymbol{x}) \\
R(a_3|\boldsymbol{x}) &= \ell_{31}\mathrm{Pr}(\mathcal{C}_1/\boldsymbol{x}) + \ell_{32}\mathrm{Pr}(\mathcal{C}_2/x) \quad \text{risk related to the reject}
\end{aligned}
$$

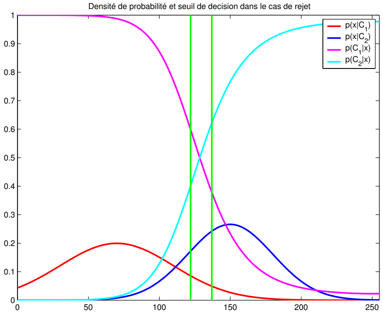
- Let consider he case of 0-1 cost and the reject cost fixed to $\alpha$. The risks read:

$$
\begin{aligned}
R(\mathcal{C}_1|\boldsymbol{x}) &= \mathrm{Pr}(\mathcal{C}_2/\boldsymbol{x}) \\
R(\mathcal{C}_2|\boldsymbol{x}) &= \mathrm{Pr}(\mathcal{C}_1/\boldsymbol{x}) \\
R(a_3|\boldsymbol{x}) &= \alpha
\end{aligned}
$$

# Classification with reject option

The Bayes' rule becomes:

$$D(\boldsymbol{x}) : \begin{cases} \mathcal{C}_1 & \text{if} \quad \Pr(\mathcal{C}_1/\boldsymbol{x}) > \Pr(\mathcal{C}_2/\boldsymbol{x}) \text{ and } \Pr(\mathcal{C}_1/\boldsymbol{x}) > 1 - \alpha \\ \mathcal{C}_2 & \text{if} \quad \Pr(\mathcal{C}_2/\boldsymbol{x}) > \Pr(\mathcal{C}_1/\boldsymbol{x}) \text{ and } \Pr(\mathcal{C}_2/\boldsymbol{x}) > 1 - \alpha \\ \text{reject} & \text{else} \end{cases}$$

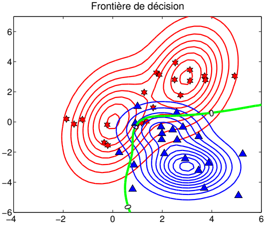



The figure describes conditional distributions and and posterior probabilities. Vertical green lines indicate the reject area

How and when the reject plays:

$$\alpha = 0 \longrightarrow 100\% \text{ reject}$$
$$\alpha = 1/2 \longrightarrow 0\% \text{ reject}$$

# Where is the learning for the machine?




## Available information : data

- $\{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}\}_{i=1}^N$

## Practical procedure

- Fix the costs related to each decision. Default: 0-1 costs
- Find the conditional distributions $p(\boldsymbol{x}/\mathcal{C}_k)$ and prior probability $\Pr(\mathcal{C}_k)$ of each class $\mathcal{C}_k$, $k = 1, \cdots, K$

  $\rightarrow$ Use the data of each $\mathcal{C}_k$ to learn $p(\boldsymbol{x}/\mathcal{C}_k)$ and $\Pr(\mathcal{C}_k)$
- Deduce then the posterior probabilities using Bayes' Th. i.e.

$$\Pr(\mathcal{C}_k/\boldsymbol{x}) = \frac{\Pr(\mathcal{C}_k)p(\boldsymbol{x}/\mathcal{C}_k)}{p_x(\boldsymbol{x})}, \; k = 1, \cdots, K$$

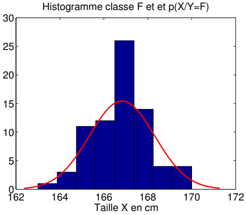
# Practical procedure : example of gender classification

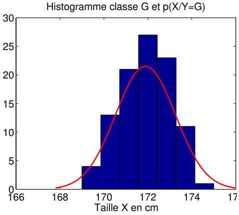
Data of each $\mathcal{C}_k$ follow a Gaussian distribution $p(x/\mathcal{C}_k) = \frac{1}{\sigma_k\sqrt{(2\pi)}} \exp^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$

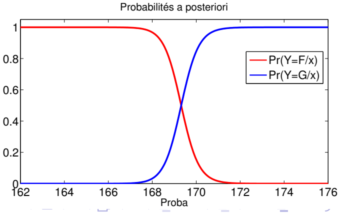
## Determine $\Pr(\mathcal{C}_k)$ and $p(x/\mathcal{C}_k)$

- For each class $\mathcal{C}_k$, select its data $\{(x_i, y_i = \mathcal{C}_k)\}_{i=1,\cdots,N_k}$
- Its prior probability is estimated by: $\Pr(\mathcal{C}_k) = \frac{N_k}{N}$
- Estimation of $\mu_k$ and $\sigma_k$: $\hat{\mu}_k = \frac{\sum_{i\in\mathcal{C}_k} x_i}{N_k}$ and the variance: $\hat{\sigma}_k = \frac{\sum_{i\in\mathcal{C}_k}(x_i-\hat{\mu}_k)^2}{N_k}$
- with $N_k$ the cardinality of class $\mathcal{C}_k$ and $N$: total number of points

Marginal $p_X(x) = p(x/\mathcal{C}_1)\Pr(\mathcal{C}_1) + p(x/\mathcal{C}_2)\Pr(\mathcal{C}_2)$. Deducing posterior probabilities $\Pr(\mathcal{C}_k/x)$



Gaussian conditional distributions case

- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis

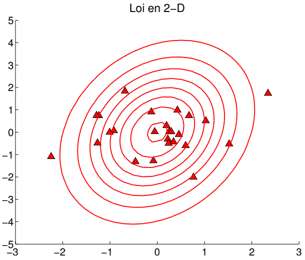
# The Gaussian distribution case

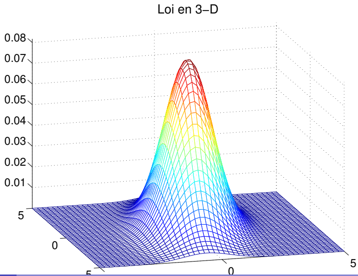
## Gaussian distribution for class $\mathcal{C}_k$

$$p(\boldsymbol{x}/\mathcal{C}_k) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{C}_k|}} \exp^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}, \quad x \in \mathbb{R}^d$$

$\boldsymbol{\mu}_k \in \mathbb{R}^d$ : mean and $\boldsymbol{C}_k \in \mathbb{R}^{d\times d}$ : covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 4 \end{pmatrix}$$



# Case study of Gaussian distributions

### Gaussian distribution for class $\mathcal{C}_k$

$$p(\boldsymbol{x}/\mathcal{C}_k) = \mathcal{N}(\boldsymbol{x}_i, \boldsymbol{\mu}_k, \boldsymbol{C}_k) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{C}_k|}} \exp^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}, \quad x \in \mathbb{R}^d$$

### Decision rule

For $0-1$ costs, we assign $\boldsymbol{x}$ to class $\mathcal{C}_j$ if

$$\begin{array}{llll} & \Pr(\mathcal{C}_j/\boldsymbol{x}) > \Pr(\mathcal{C}_k/\boldsymbol{x}) & \forall k \neq j & \\ \Leftrightarrow & p(\boldsymbol{x}/\mathcal{C}_j)\Pr(\mathcal{C}_j) > p(\boldsymbol{x}/\mathcal{C}_k)\Pr(\mathcal{C}_k) & \forall k \neq j & \\ \Leftrightarrow & \ln p(\boldsymbol{x}/\mathcal{C}_j) + \ln \Pr(\mathcal{C}_j) > \ln p(\boldsymbol{x}/\mathcal{C}_k) + \ln \Pr(\mathcal{C}_k) & \forall k \neq j & (1) \end{array}$$

### Discrimination function

- Let $g_k(\boldsymbol{x}) = \ln p(\boldsymbol{x}/\mathcal{C}_j) + \ln \Pr(\mathcal{C}_j)$ the discrimination function related to $\mathcal{C}_k$
- For $p(\boldsymbol{x}/\mathcal{C}_k) = \mathcal{N}(\boldsymbol{x}_i, \boldsymbol{\mu}_k, \boldsymbol{C}_k)$, we have

$$g_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{C}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) - \frac{d}{2}\ln 2\pi - \frac{1}{2} ln(|\boldsymbol{C}_k|) + \ln \Pr(\mathcal{C}_j)$$

# Linear discriminant analysis (1)

- Assumption
  LDA assumes all classes have the same covariance matrix i.e.

$$\boldsymbol{C}_k = \boldsymbol{C} \quad \forall k = 1 \cdots K$$

- This introduces some simplifications in $g_k(\boldsymbol{x})$

$$g_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{C}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \ln \Pr(\mathcal{C}_j) \underbrace{-\frac{d}{2}\ln 2\pi - \frac{1}{2} ln(|\boldsymbol{C}|)}_{\text{cst}}$$

- Expanding the quadratic term in $g_k(\boldsymbol{x})$, we get

$$g_k(\boldsymbol{x}) \quad = \quad \boldsymbol{\mu}_k^\top \boldsymbol{C}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{C}^{-1}\boldsymbol{\mu}_k + \ln \Pr(\mathcal{C}_j) + \underbrace{\text{cst} - \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{C}^{-1}\boldsymbol{x}}_{\text{cst}}$$

$$g_k(\boldsymbol{x}) = \boldsymbol{w}_k^\top \boldsymbol{x} + w_{jo} + \text{cst}, \text{ with } \boldsymbol{w}_k = \boldsymbol{C}^{-1}\boldsymbol{\mu}_k, \; w_{jo} = \ln \Pr(\mathcal{C}_j) - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{C}^{-1}\boldsymbol{\mu}_k$$

# Linear discriminant analysis (2)

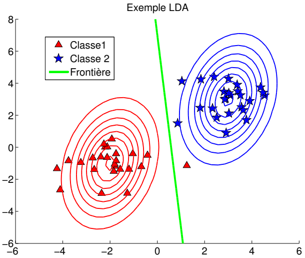
Decision rule: $\boldsymbol{x}$ is predicted the class $\mathcal{C}_j$ if

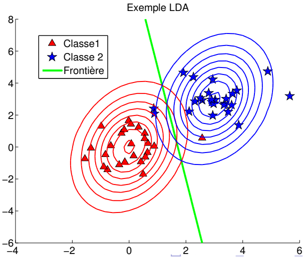
$$g_j(\boldsymbol{x}) = \boldsymbol{w}_j^\top \boldsymbol{x} + w_{jo} > g_k(\boldsymbol{x}) = \boldsymbol{w}_k^\top \boldsymbol{x} + w_{ko} \quad \forall k \neq j$$

**Linear decision function : predict class $\mathcal{C}_j$ if**

$$\boldsymbol{w}^\top(\boldsymbol{x} - \boldsymbol{x}_0) + b > 0 \qquad \forall k \neq j$$

with $\boldsymbol{w} = \boldsymbol{C}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k), \quad \boldsymbol{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_k)$ and $b = \ln \frac{\Pr(\mathcal{C}_j)}{\Pr(\mathcal{C}_k)}$



# Quadratic discriminant analysis (QDA)

General case: the covariance matrices are different i.e. $\boldsymbol{C}_k \neq \boldsymbol{C}_j, \ \forall k \neq j$

## Quadratic discrimination function

$$g_k(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{C}_k^{-1}\boldsymbol{x} + \boldsymbol{w}_k^\top \boldsymbol{x} + w_{ko}$$

$$\boldsymbol{w}_k = \boldsymbol{C}_k^{-1}\boldsymbol{\mu}_k, \ w_{ko} = -\tfrac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{C}_k^{-1}\boldsymbol{\mu}_k - \tfrac{1}{2}\ln(|\boldsymbol{C}_k| \times \Pr(\mathcal{C}_k))$$

## Decision function

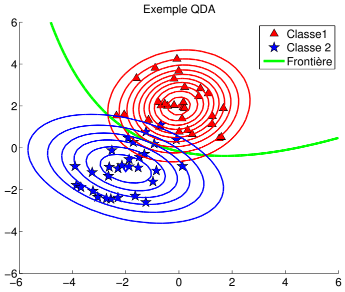
$\boldsymbol{x}$ is predicted the class $\mathcal{C}_j$ if

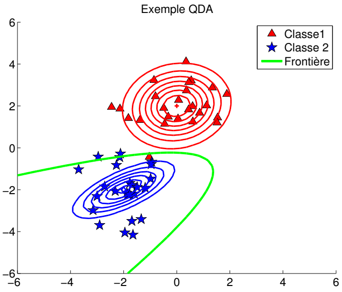
$$\begin{aligned} g_j(\boldsymbol{x}) \quad &> \quad g_k(\boldsymbol{x}) \quad \forall k \neq j \\ \Leftrightarrow -\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{C}_j^{-1}\boldsymbol{x} + \boldsymbol{w}_j^\top \boldsymbol{x} + w_{jo} \quad &> \quad -\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{C}_k^{-1}\boldsymbol{x} + \boldsymbol{w}_k^\top \boldsymbol{x} + w_{ko} \quad \forall k \neq j \end{aligned}$$

$\Longrightarrow$ the decision function is quadratic (hence the name of the method)

# Quadratic discriminant analysis: illustration

For a binary classification problem, the decision boundary is quadratic



# Conclusion: estimation strategies

## Practical implementation

- For each class $\mathcal{C}_k$, $k = 1, \cdots, K$
  - Get the training data-set associated to the class $\mathcal{C}_k$
  - Estimate its prior probability $\Pr(\mathcal{C}_k)$ and its conditional distribution $p(\boldsymbol{x}/\mathcal{C}_k)$
- Estimate the decision rule (by using one of the two methods) :
  1. For each data $\boldsymbol{x}$ compute the posterior probabilities $\Pr(\mathcal{C}_k/\boldsymbol{x})$ and affect $\boldsymbol{x}$ to the class minimizing the conditional risk
  2. Determine the functions of discrimination $g_k(\boldsymbol{x})$ and deduce the rule.
     In the case of binary classification the decision function is often expressed as a sign of $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$

# Estimation strategies : Gaussian case

- LDA : The parameters of each class $\mathcal{C}_k$ and the common covariance matrix $\boldsymbol{C}$ are estimated as:

$$
\begin{aligned}
\boldsymbol{\mu}_k &= \frac{\sum_{i\in\mathcal{C}_k}^{N} \boldsymbol{x}_i}{N_k} \quad \text{with} \quad N_k = \text{card}(\mathcal{C}_k) \\
\Pr(\mathcal{C}_k) &= \frac{N_k}{N} \\
\boldsymbol{C} &= \frac{\sum_{k=1}^{K} \sum_{i\in\mathcal{C}_k}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top}{N-K}
\end{aligned}
$$

- QDA case : we estimate the covariance matrix of each class $\mathcal{C}_k$ by

$$
\boldsymbol{C}_k = \frac{\sum_{i\in\mathcal{C}_k}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(x_i - \boldsymbol{\mu}_k)^\top}{N-1}
$$

The estimation of the prior probability and of $\boldsymbol{\mu}_k$ is similar to LDA.

# Summing up

Bayesian decision theory provides a formal framework for (binary of multi-class) classification which minimizes the generalization risk.