

# Analyse de Données – Développeur

## C5-160512-INFO

### 5. Apprentissage Supervisé

Pr. Carl FRÉLICOT

FST/PAS221 – carl.frelicot@univ-lr.fr



Licence d'Informatique 3ème année

Automne 2021 - La Rochelle



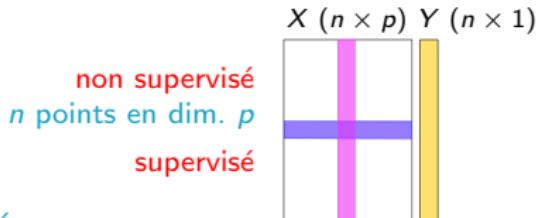
1. Préambule
2. Tableaux et Espaces
3. Réduction de la Dimensionnalité
4. Apprentissage Non Supervisé
5. Apprentissage Supervisée
  - 5.1 Problématique(s)
  - 5.2 Nearest-Prototype
  - 5.3 Nearest-Neighbors
  - 5.4 Évaluation

Chap. 2 : Tableaux et Espaces

- individus vs variables quantitatives

$$X = [X_{ki}]_{k=1,n; i=1,p}$$

+ éventuellement une variable catégorielle  $Y$



Chap. 3 : Réduction de la Dimensionnalité

- ACP non supervisé
  - AFD supervisé

ex. :  $p = 2$  à  $q = 1$

Chap. 4 : Apprentissage Non Supervisé

identifier dans  $X$  une structure de groupes

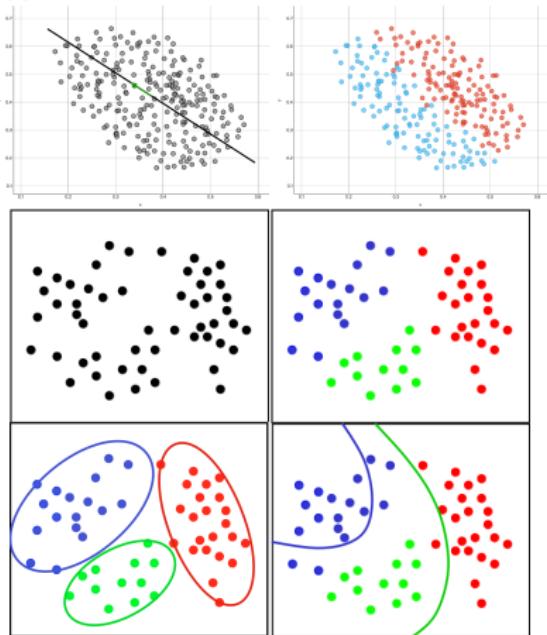
$\rightarrow Y$

## apprentissage

Chap. 5 : Apprentissage Supervisé



par ex. : AFD décisionnelle vue au Chap. 3



$X$  ( $n \times p$ )  $Y$  ( $n \times 1$ )

## Chap. 2 : Tableaux et Espaces

- $X = \{x_1, \dots, x_n\}$  est un ensemble d'exemples  $x_k \in \mathcal{X}$ , où  
 $\mathcal{X}$  est un **espace de représentation** (en général  $\mathcal{X} = \mathbb{R}^p$ )  
 $X$  est parfaitement étiqueté (expertise, *clustering*, etc)
- $Y = \{y_1, \dots, y_n\}$  est l'ensemble d'étiquettes  $y_k \in \mathcal{Y}$  associé,  
 $\mathcal{Y}$  est un **espace d'étiquettes** des classes

## Différents types de Classification Supervisée

- classification binaire
- classification multi-étiquettes  $\supset$  binaire
- **classification multi-classes** (étiquettes singlentons)
- classification sélective

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

$$\mathcal{Y} = 2^{\{1, 2, \dots, c\}}$$

$$\mathcal{Y} = \{1, 2, \dots, c\}$$

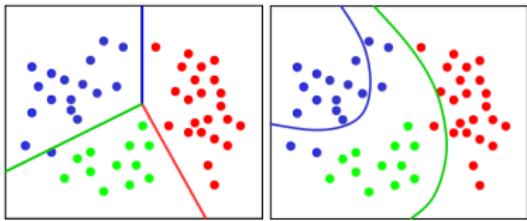
$$y_k \in \mathcal{Y} = \{1, 2, \dots, c\}$$

$$\text{et } y \in \mathcal{Y} = 2^{\{1, 2, \dots, c\}}$$

## Chap. 5 : Apprentissage Supervisé

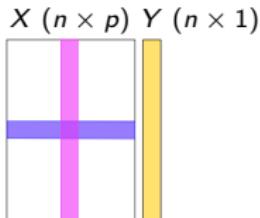
- construire à partir de  $(X, Y)$  une règle de classement ou fonction d'affectation  $f$  :  
 $x \in \mathcal{X} \rightarrow f(x) = y \in \mathcal{Y}$  **apprentissage**  
une fonction linéaire , quadratique , etc
- de sorte de pouvoir **prédir** le groupe d'un nouvel  $x$

par ex. : AFD décisionnelle vue au Chap. 3



## Chap. 2 : Tableaux et Espaces

- $X = \{x_1, \dots, x_n\}$  est un ensemble d'exemples  $x_k \in \mathcal{X}$ , où  $\mathcal{X}$  est un **espace de représentation** (en général  $\mathcal{X} = \mathbb{R}^p$ )  
 $X$  est parfaitement étiqueté (expertise, clustering, etc)
- $Y = \{y_1, \dots, y_n\}$  est l'ensemble d'étiquettes  $y_k \in \mathcal{Y}$  associé,  
 $\mathcal{Y}$  est un **espace d'étiquettes** des classes



## Différents types de Classification Supervisée

- classification binaire
- classification multi-étiquettes  $\supset$  binaire
- **classification multi-classes** (étiquettes singlentons)
- classification sélective

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

$$\mathcal{Y} = 2^{\{1, 2, \dots, c\}}$$

$$\mathcal{Y} = \{1, 2, \dots, c\}$$

$$y_k \in \mathcal{Y} = \{1, 2, \dots, c\}$$

$$\text{et } y \in \mathcal{Y} = 2^{\{1, 2, \dots, c\}}$$

## Étiquetage et Classement

Le plus souvent la classification est constituée de deux étapes distinctes :  $f = (L, A)$

① étiquetage  $L : \mathcal{X} \mapsto \mathcal{U}$ ,  $x \rightarrow u$

probabilités, typicalités, distances

② aggrégation  $A : \mathcal{U} \mapsto \mathcal{Y}$ ,  $u \rightarrow y$

maximum/majorité, top-scores, minimum

Beaucoup de méthodes partagent  $L$  ou (exclusif)  $A$ .

## Évaluation

Quel  $f \in \mathcal{F}$  est le **meilleur prédicteur**  $x$  n'ayant pas participé à sa construction ?

- complexité de la méthode choisie
- variabilité dans  $\mathcal{X}$

capacité en  
généralisation



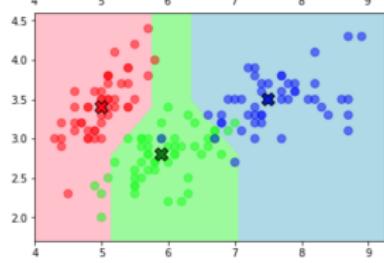
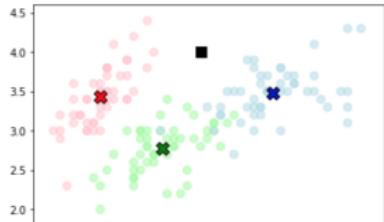
1. Préambule
2. Tableaux et Espaces
3. Réduction de la Dimensionnalité
4. Apprentissage Non Supervisé
5. Apprentissage Supervisée
  - 5.1 Problématique(s)
  - 5.2 Nearest-Prototype
    - 5.2.1 NP Rule
    - 5.2.2 Pros and Cons
    - 5.2.3 Exemple sur Données Réelles
  - 5.3 Nearest-Neighbors
  - 5.4 Évaluation

## Règle du Plus Proche Prototype

- ➊  $(X, Y) \rightarrow G = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_c\}$       **apprentissage expertise, clustering**
- ➋  $L : x \rightarrow u(x) = [u_j(x)]_{j=1,c}$ , où  $u_j(x) = d_j(x, \bar{x}_j)$
- ➌  $A : u \rightarrow y(u) = \operatorname{argmin}_{j=1,c} u_j(x)$       **prédiction**

## Exemple

- ➊  $x = (6.5, 4.0)$
- ➋  $G = \begin{pmatrix} 5.006 & 3.428 \\ 5.936 & 2.77 \\ 7.588 & 3.474 \end{pmatrix}$
- ➌  $u(x) = (1.5998, 1.3531, 1.2085)$
- ➍  $y(u) = 3 \rightarrow \bullet$



pourquoi cette différence ?  
d. euclidienne, Manhattan

## Étiquetage et Classement

Le plus souvent la classification est constituée de deux étapes distinctes :  $f = (L, A)$

- ➊ étiquetage  $L : \mathcal{X} \mapsto \mathcal{U}, x \rightarrow u$       **probabilités, typicalités, distances**
- ➋ aggrégation  $A : \mathcal{U} \mapsto \mathcal{Y}, u \rightarrow y$       **maximum/majorité, top-scores, minimum**

Beaucoup de méthodes partagent  $L$  ou (exclusif)  $A$ .



1. Préambule
2. Tableaux et Espaces
3. Réduction de la Dimensionnalité
4. Apprentissage Non Supervisé
5. Apprentissage Supervisée
  - 5.1 Problématique(s)
  - 5.2 Nearest-Prototype
    - 5.2.1 NP Rule
    - 5.2.2 Pros and Cons
    - 5.2.3 Exemple sur Données Réelles
  - 5.3 Nearest-Neighbors
  - 5.4 Évaluation

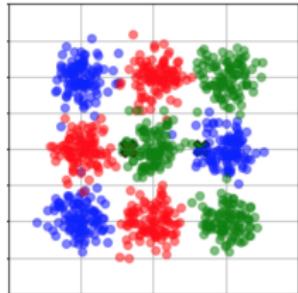
## Règle du Plus Proche Prototype

- ①  $(X, Y) \rightarrow G = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_c\}$  apprentissage  
expertise, clustering
  - ②  $L : x \rightarrow u(x) = [u_j(x)]_{j=1,c}$ , où  $u_j(x) = d_j(x, \bar{x}_j)$
  - ③  $A : u \rightarrow y(u) = \operatorname{argmin}_{i=1,c} u_i(x)$  prédiction

### Avantages / Inconvénients

- + intuitive, simple à implémenter
  - + extensible à plusieurs prototypes par groupe
  - non locale, peu adaptative
  - $\ominus\oplus$  sensible à  $d$
  - + capacité en généralisation ?  
si les classes sont linéairement séparables et/ou uni-modales...

frontières linéaires  
une idée ?



#### **Reclassement :**

```
[[125 75 100]  
 [172 86 42]  
 [ 17 67 216]]
```

$$\begin{aligned}E &= 1 - CA \\&= 1 - \frac{427}{900} \\&= 52.56\%\end{aligned}$$

## Reclassement ?

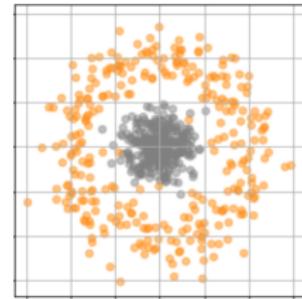
$\begin{bmatrix} 149 & 151 \\ 147 & 153 \end{bmatrix}$

matrice de confusion

$$CA = \frac{\sum_{i=1}^c C_{ii}}{n}$$

$$= \frac{302}{600}$$

$$= 50.33\%$$





1. Préambule
2. Tableaux et Espaces
3. Réduction de la Dimensionnalité
4. Apprentissage Non Supervisé
5. Apprentissage Supervisée
  - 5.1 Problématique(s)
  - 5.2 Nearest-Prototype
    - 5.2.1 NP Rule
    - 5.2.2 Pros and Cons
    - 5.2.3 Exemple sur Données Réelles
  - 5.3 Nearest-Neighbors
  - 5.4 Évaluation

## Exemple : Olivetti Faces DataBase

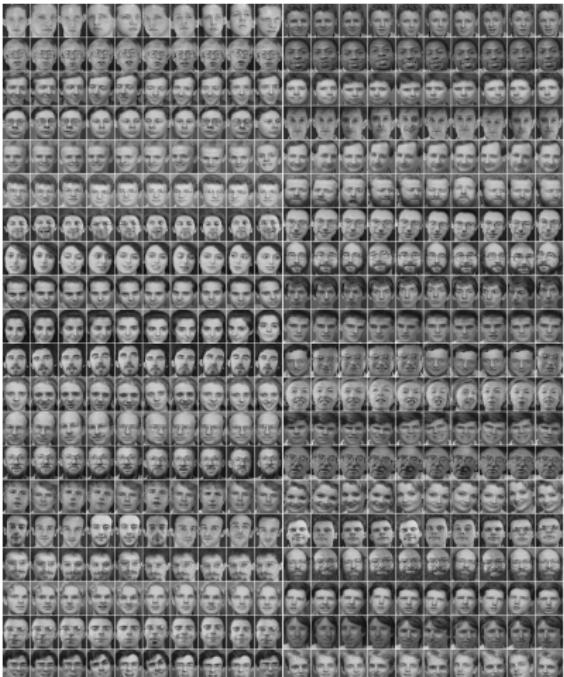
- $c = 10$  personnes
- $n = c \times 40 = 400$  images de taille  $64 \times 64$  pixels
- $p = 64 \times 64 = 4096$
- réduction de la dimensionnalité par ACP  
 $q = 64$  composantes expliquent plus de 90% de la variance totale du nuage

## NP-Rule

- 1 photo au hasard de chaque personne sortie de la base  $\rightarrow C_q$  est  $(360 \times 64)$
- prédiction des 40 images dans l'espace des composantes principales
- reconnaissance des personnes ?
  - dix essais
  - $CA \in [80\%, 95\%]$ ,  $\overline{CA} = 88\%$
- meilleures prédict. dans l'espace initial ? rien ne permet de l'affirmer !

0.95

```
[ True  True  True  False  True  True  True  True  True  True  True  True  True
   True  True  True  True  True  True  True  True  True  True  True  True
   True  True  True  True  False  True  True  True  True  True  True  True
   True  True  True  True  True ]
```

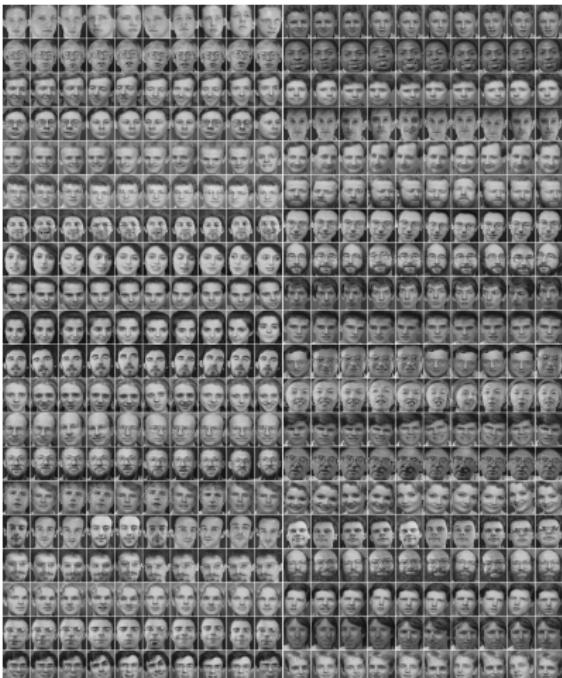


## Exemple : Olivetti Faces DataBase

- $c = 10$  personnes
- $n = c \times 40 = 400$  images de taille  $64 \times 64$  pixels
- $p = 64 \times 64 = 4096$
- réduction de la dimensionnalité par AFD  
 $q$  ? variables discriminantes       $q = 39$

## NP-Rule

- 1 photo au hasard de chaque personne sortie de la base  $\rightarrow C_q$  est  $(360 \times 64)$
- prédiction des 40 images dans l'espace des composantes principales
- reconnaissance des personnes ?
  - dix essais
  - $CA \in [95\%, 100\%]$ ,  $\overline{CA} = 99\%$
- meilleures prédict. dans l'espace initial ? probablement pas !



1.0

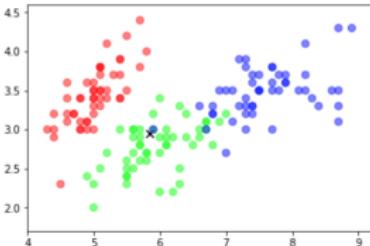
```
[ True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True  True  True  True  True  True  True  True  True
  True  True  True  True ]
```



1. Préambule
2. Tableaux et Espaces
3. Réduction de la Dimensionnalité
4. Apprentissage Non Supervisé
5. Apprentissage Supervisée
  - 5.1 Problématique(s)
  - 5.2 Nearest-Prototype
  - 5.3 Nearest-Neighbors
    - 5.3.1 NN Rule
    - 5.3.2 K-NN Rule
    - 5.3.3 Exemple sur Données Réelles
  - 5.4 Évaluation

## Règle du Plus Proche Voisin

- ①  $X \rightarrow D = [d_k]_{k=1,n}$ , où  $d_k = d(x, x_k)$   
tri dans l'ordre décroissant  $D = [d_{(1)}, d_{(2)}, \dots, d_{(n)}]$
- ②  $L : x \rightarrow u(x) = [u_j(x)]_{j=1,c}$ , où  $u_j(x) = 1$  si  $j = Y_{(1)}$   
et 0 sinon
- ③  $A : u \rightarrow y(u) = \text{argmax}_{j=1,c} u_j(x)$



## Exemple

- $x = (5.85, 2.95)$
- ①  $D = (0.707, 0.707, 1.158, \dots, 3.154, 3.335)$
- ②  $u(x) = (0, 0, 1)$
- ③  $y(u) = 3 \rightarrow \bullet$

## Étiquetage et Classement

Le plus souvent la classification est constituée de deux étapes distinctes :  $f = (L, A)$

- |                                                                      |                                       |
|----------------------------------------------------------------------|---------------------------------------|
| ① étiquetage $L : \mathcal{X} \mapsto \mathcal{U}, x \rightarrow u$  | probabilités, typicalités, distances  |
| ② aggrégation $A : \mathcal{U} \mapsto \mathcal{Y}, u \rightarrow y$ | maximum/majorité, top-scores, minimum |

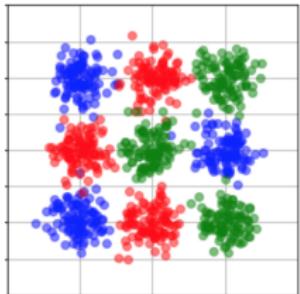
Beaucoup de méthodes partagent  $L$  ou (exclusif)  $A$ .

## Règle du Plus Proche Voisin

- ①  $X \rightarrow D = [d_k]_{k=1,n}$ , où  $d_k = d(x, x_k)$   
tri dans l'ordre décroissant  $D = [d_{(1)}, d_{(2)}, \dots, d_{(n)}]$
- ②  $L : x \rightarrow u(x) = [u_j(x)]_{j=1,c}$ , où  $u_j(x) = 1$  si  $j = Y_{(1)}$   
et 0 sinon
- ③  $A : u \rightarrow y(u) = \text{argmax}_{j=1,c} u_j(x)$

## Avantages / Inconvénients

- ⊕ intuitive, simple à implémenter
- ⊖⊕ locale, comportement adaptatif
- ⊖⊕ sensible à  $d$
- ⊕ capacité en généralisation ?  
si les classes sont linéairement séparables et/ou  
uni-modales...



Reclassement ?

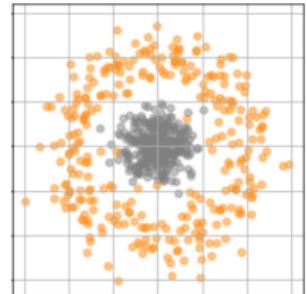
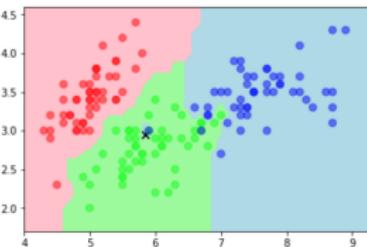
$$\begin{bmatrix} 300 & 0 & 0 \\ 6 & 294 & 0 \\ 3 & 4 & 293 \end{bmatrix}$$

$$\begin{aligned} E &= 1 - CA \\ &= 1 - \frac{887}{900} \\ &= 1.44\% \end{aligned}$$

Reclassement ?

$$\begin{bmatrix} [300 & 0] \\ [5 & 295] \end{bmatrix}$$

$$\begin{aligned} CA &= \frac{595}{600} \\ &= 99.17\% \end{aligned}$$

distance euclidienne  
distance de Manhattan



1. Préambule
2. Tableaux et Espaces
3. Réduction de la Dimensionnalité
4. Apprentissage Non Supervisé
5. Apprentissage Supervisée
  - 5.1 Problématique(s)
  - 5.2 Nearest-Prototype
  - 5.3 Nearest-Neighbors
    - 5.3.1 NN Rule
    - 5.3.2 K-NN Rule
    - 5.3.3 Exemple sur Données Réelles
  - 5.4 Évaluation

## Règle du Plus Proche Voisin

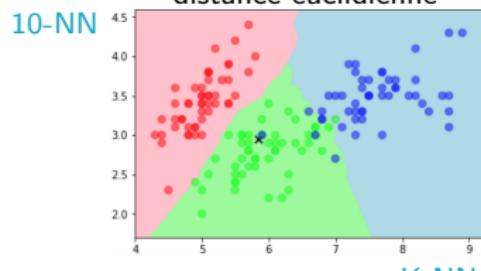
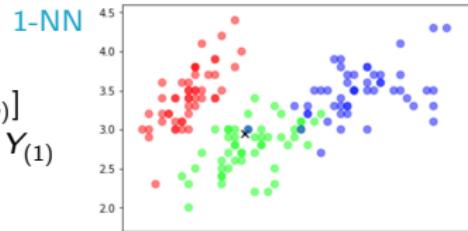
- ①  $X \rightarrow D = [d_k]_{k=1,n}$ , où  $d_k = d(x, x_k)$   
tri dans l'ordre décroissant  $D = [d_{(1)}, d_{(2)}, \dots, d_{(n)}]$
- ②  $L : x \rightarrow u(x) = [u_j(x)]_{j=1,c}$ , où  $u_j(x) = 1$  si  $j = Y_{(1)}$   
et 0 sinon
- ③  $A : u \rightarrow y(u) = \text{argmax}_{j=1,c} u_j(x)$

## Exemple

- $x = (5.85, 2.95)$
- ①  $D = (0.707, 0.707, 1.158, \dots, 3.154, 3.335)$
- ②  $u(x) = (0, \frac{9}{10}, \frac{1}{10})$
- ③  $y(u) = 2 \rightarrow \bullet$

## Règle des K-Plus Proches Voisins

- affecter  $x$  au groupe majoritairement représenté parmi ses  $K$ -PPV
- ①  $X \rightarrow D = [d_k]_{k=1,n}$ , où  $d_k = d(x, x_k)$   
tri dans l'ordre décroissant  $D = [d_{(1)}, d_{(2)}, \dots, d_{(n)}]$   
groupes/étiquettes des  $K$  plus proches  $V_K = [Y_{(1)}, Y_{(2)}, \dots, Y_{(K)}]$   
comptage  $C = [k_j]_{j=1,c}$ , où  $k_j = \text{card}_{V_K}(Y == j)$
- ②  $L : x \rightarrow u(x) = [u_j(x)]_{j=1,c}$ , où  $u_j(x) = \frac{k_j}{K}$
- ③  $A : u \rightarrow y(u) = \text{argmax}_{j=1,c} u_j(x)$



K-NN

 $\in [0, 1]$ 

proba. dans Orange

## Avantages / Inconvénients

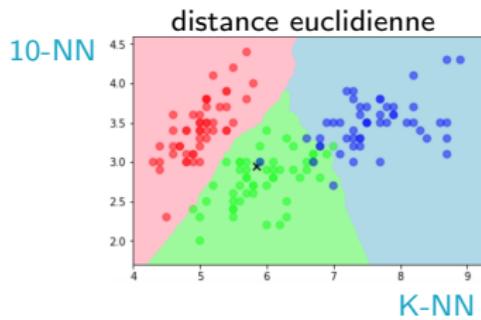
- ⊕ intuitive, simple à implémenter
- ⊖⊕ locale, comportement adaptatif
- ⊖⊕ sensible à  $d$
- ⊕ classes non lin. séparables et/ou multi-modales ?
- ⊕ capacité à prédire ↗ avec  $K$ ; au détriment de CA ?

## Exemple

- $x = (5.85, 2.95)$
- ①  $D = (0.707, 0.707, 1.158, \dots, 3.154, 3.335)$
- ②  $u(x) = (0, \frac{9}{10}, \frac{1}{10})$
- ③  $y(u) = 2 \rightarrow \bullet$

## Règle des K-Plus Proches Voisins

- affecter  $x$  au groupe majoritairement représenté parmi ses  $K$ -PPV
- ①  $X \rightarrow D = [d_k]_{k=1,n}$ , où  $d_k = d(x, x_k)$   
tri dans l'ordre décroissant  $D = [d_{(1)}, d_{(2)}, \dots, d_{(n)}]$   
groupes/étiquettes des  $K$  plus proches  $V_K = [Y_{(1)}, Y_{(2)}, \dots, Y_{(K)}]$   
comptage  $C = [k_j]_{j=1,c}$ , où  $k_j = \text{card}_{V_K}(Y == j)$
- ②  $L : x \rightarrow u(x) = [u_j(x)]_{j=1,c}$ , où  $u_j(x) = \frac{k_j}{K} \in [0, 1]$
- ③  $A : u \rightarrow y(u) = \text{argmax}_{j=1,c} u_j(x)$  proba. dans Orange

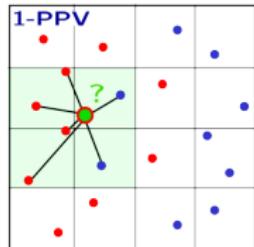


## Avantages / Inconvénients

- ⊕ intuitive, simple à implémenter
- ⊖⊕ locale, comportement adaptatif
- ⊖⊕ sensible à  $d$
- ⊕ classes non lin. séparables et/ou multi-modales
- ⊕ capacité à prédire  $\nearrow$  avec  $K$ ; au détriment de CA ?
- ⊖ coûteuse en temps et espace      comparé à NP

## Réduire le coût

- l'espace de recherche
  - *bucketing*
  - *kd-tree*
- la taille de l'ensemble d'apprentissage, donc  $D$ 
  - *condensing* : supprimer les points mal reclassés
  - *editing* : supprimer ceux entourés de points de la même classe



## Règle des K-Plus Proches Voisins

K-NN

- affecter  $x$  au groupe majoritairement représenté parmi ses  $K$ -PPV
- ①  $X \rightarrow D = [d_k]_{k=1,n}$ , où  $d_k = d(x, x_k)$   
tri dans l'ordre décroissant  $D = [d_{(1)}, d_{(2)}, \dots, d_{(n)}]$   
groupes/étiquettes des  $K$  plus proches  $V_K = [Y_{(1)}, Y_{(2)}, \dots, Y_{(K)}]$   
comptage  $C = [k_j]_{j=1,c}$ , où  $k_j = \text{card}_{V_K}(Y == j)$
- ②  $L : x \rightarrow u(x) = [u_j(x)]_{j=1,c}$ , où  $u_j(x) = \frac{k_j}{K} \in [0, 1]$
- ③  $A : u \rightarrow y(u) = \text{argmax}_{j=1,c} u_j(x)$       proba. dans Orange



1. Préambule
2. Tableaux et Espaces
3. Réduction de la Dimensionnalité
4. Apprentissage Non Supervisé
5. Apprentissage Supervisée
  - 5.1 Problématique(s)
  - 5.2 Nearest-Prototype
  - 5.3 Nearest-Neighbors
    - 5.3.1 NN Rule
    - 5.3.2 K-NN Rule
    - 5.3.3 Exemple sur Données Réelles
  - 5.4 Évaluation

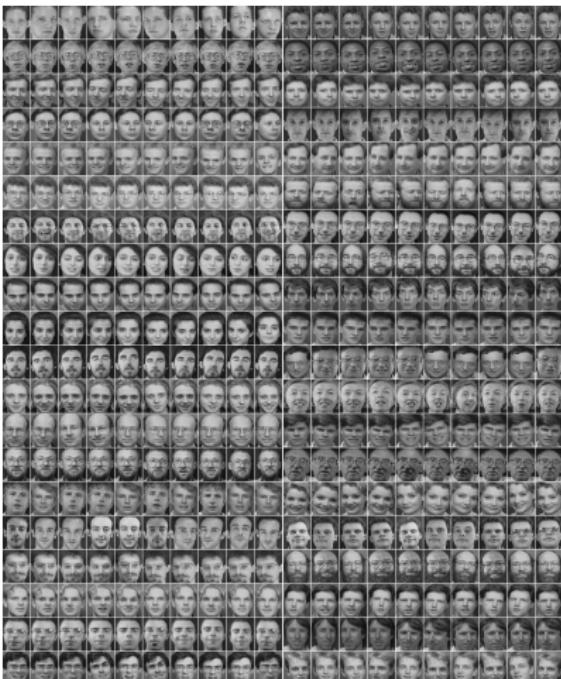
## Exemple : Olivetti Faces DataBase

- $c = 10$  personnes
- $n = c \times 40 = 400$  images de taille  $64 \times 64$  pixels
- $p = 64 \times 64 = 4096$
- réduction de la dimensionnalité par ACP  
 $q = 64$  composantes expliquent plus de 90% de la variance totale du nuage

## KNN-Rule

## 1-NN

- 1 photo au hasard de chaque personne sortie de la base  $\rightarrow C_q$  est  $(360 \times 64)$
- prédiction des 40 images dans l'espace des composantes principales
- reconnaissance des personnes ?
  - dix essais
  - $CA \in [90\%, 100\%]$ ,  $\overline{CA} = 97\%$
- meilleures prédict. dans l'espace initial ? rien ne permet de l'affirmer !



## KNN-Rule et AFD ?

## 1-NN

- une alternative à l'AFD décisionnelle (NP-Rule)
  - $CA \in [95\%, 100\%]$ ,  $\overline{CA} = 98.5\%$

## Exemple : Olivetti Faces DataBase

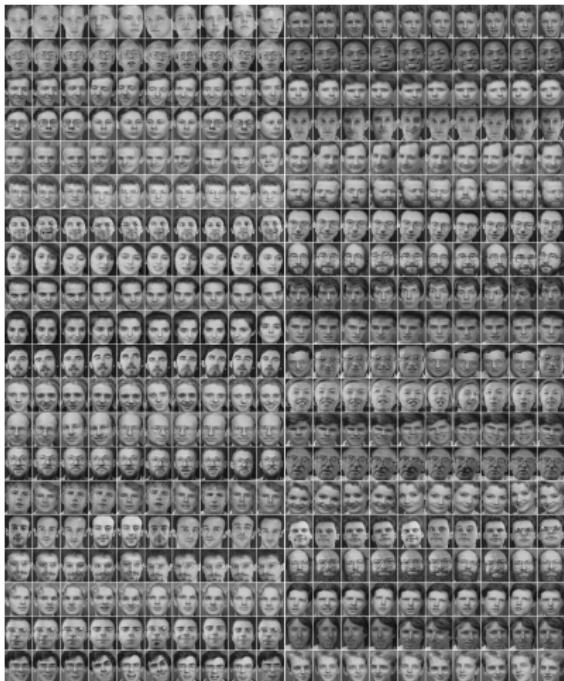
- $c = 10$  personnes
- $n = c \times 40 = 400$  images de taille  $64 \times 64$  pixels
- $p = 64 \times 64 = 4096$
- réduction de la dimensionnalité par
  - ACP,  $q = 64$
  - AFD,  $q = 39$
- 1 photo au hasard de chaque personne sortie de la base  $\rightarrow C_q$  est  $(360 \times 64)$
- prédiction des 40 images dans l'espace de dim.  $q$
- reconnaissance des personnes ? dix essais

## NP-Rule

- $CA \in [80\%, 95\%]$ ,  $\overline{CA} = 88\%$
- $CA \in [95\%, 100\%]$ ,  $\overline{CA} = 99\%$

## 1NN-Rule

- $CA \in [90\%, 100\%]$ ,  $\overline{CA} = 97\%$
- $CA \in [95\%, 100\%]$ ,  $\overline{CA} = 98.5\%$

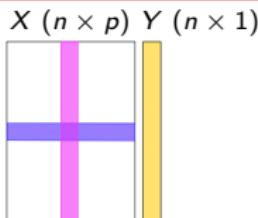




1. Préambule
2. Tableaux et Espaces
3. Réduction de la Dimensionnalité
4. Apprentissage Non Supervisé
5. Apprentissage Supervisée
  - 5.1 Problématique(s)
  - 5.2 Nearest-Prototype
  - 5.3 Nearest-Neighbors
  - 5.4 Évaluation

## Nécessité d'Évaluer

- $X = \{x_1, \dots, x_n\}$  est un ensemble d'exemples  $x_k \in \mathcal{X}$ , où  $\mathcal{X}$  est un **espace de représentation** (en général  $\mathcal{X} = \mathbb{R}^p$ )  
 $X$  est parfaitement étiqueté (expertise, *clustering*, etc)
- $Y = \{y_1, \dots, y_n\}$  est l'ensemble d'étiquettes  $y_k \in \mathcal{Y}$  associé,  
 $\mathcal{Y}$  est un **espace d'étiquettes** des classes



## Différents types de Classification Supervisée

- classification binaire
- classification multi-étiquettes  $\supset$  binaire
- **classification multi-classes** (étiquettes singlentons)
- classification sélective

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

$$\mathcal{Y} = 2^{\{1, 2, \dots, c\}}$$

$$\mathcal{Y} = \{1, 2, \dots, c\}$$

$$y_k \in \mathcal{Y} = \{1, 2, \dots, c\}$$

$$\text{et } y \in \mathcal{Y} = 2^{\{1, 2, \dots, c\}}$$

## Étiquetage et Classement

Le plus souvent la classification est constituée de deux étapes distinctes :  $f = (L, A)$

① étiquetage  $L : \mathcal{X} \mapsto \mathcal{U}$ ,  $x \rightarrow u$       probabilités, typicalités, distances

② aggrégation  $A : \mathcal{U} \mapsto \mathcal{Y}$ ,  $u \rightarrow y$       maximum/majorité, top-scores, minimum

Beaucoup de méthodes partagent  $L$  ou (exclusif)  $A$ .

## Évaluation

Quel  $f \in \mathcal{F}$  est le **meilleur prédicteur**  $x$  n'ayant pas participé à sa construction ?

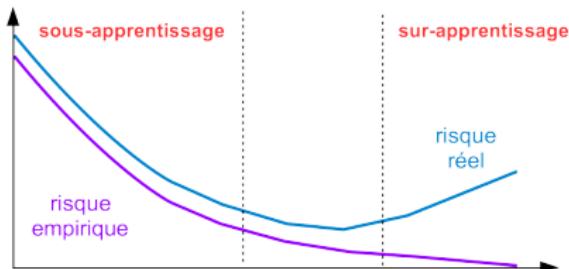
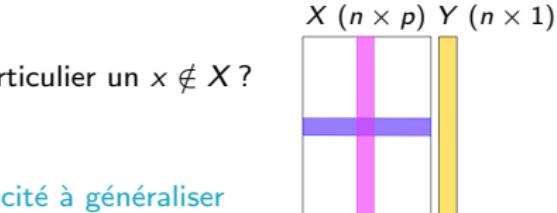
- complexité de la méthode choisie
- variabilité dans  $\mathcal{X}$

capacité en  
généralisation

## Nécessité d'Évaluer

Quel  $f \in \mathcal{F}$  est le meilleur prédicteur  $\forall x$ , en particulier un  $x \notin X$  ?

- complexité de la méthode choisie
- variabilité dans  $\mathcal{X}$ 
  - $X$  ne reflète pas tout  $\mathcal{X}$       capacité à généraliser
  - $(X, Y)$  n'est qu'une observation d'un  $(\mathbf{X}, \mathbf{Y}) \sim \text{loi } P$  sur  $\mathcal{X} \times \mathcal{Y}$  inconnue
- le risque réel (probabilité d'erreur)  $R(f)$  n'est pas accessible, on n'accède qu'à  $\widehat{R}(f)$ , son estimation sur  $(X, Y)$
- risque empirique; par ex.  $E = 1 - CA$
- chercher à diminuer  $\widehat{R}(f)$  à tout prix nuit à la capacité à généraliser
- compromis...
- dans le choix de  $f$  et ses hyper-paramètres  
par ex. NP vs KNN et  $d$ ,  $K$ , etc



## Évaluation

Quel  $f \in \mathcal{F}$  est le meilleur prédicteur  $x$  n'ayant pas participé à sa construction ?

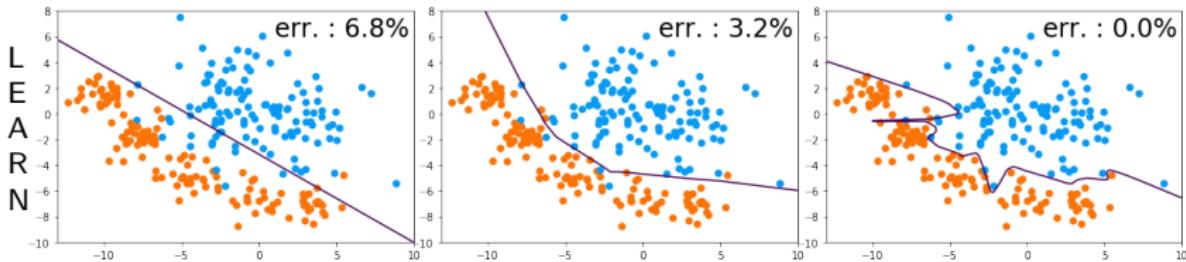
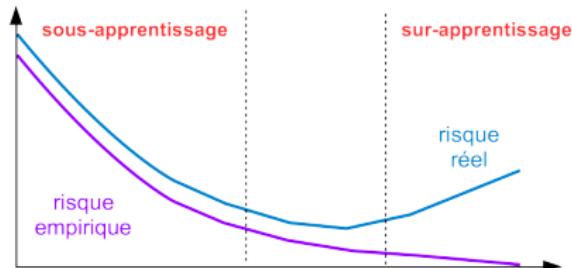
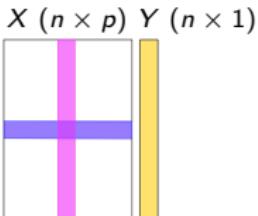
- complexité de la méthode choisie
- variabilité dans  $\mathcal{X}$

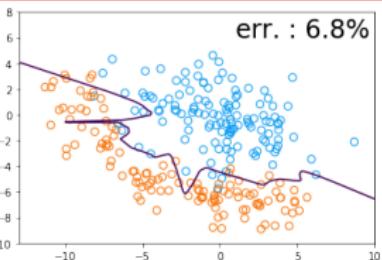
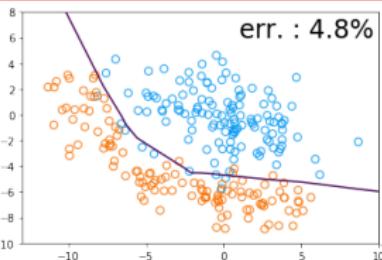
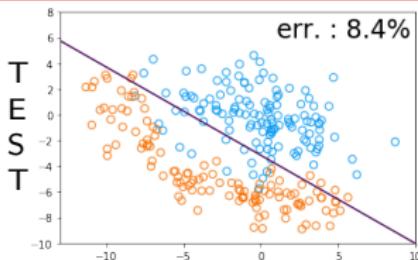
capacité en  
généralisation

## Nécessité d'Évaluer

Quel  $f \in \mathcal{F}$  est le meilleur prédicteur  $\forall x$ , en particulier un  $x \notin X$  ?

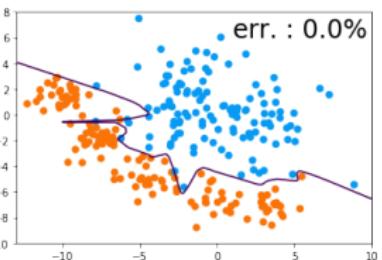
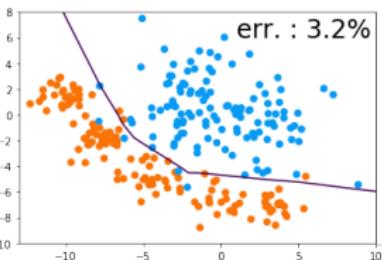
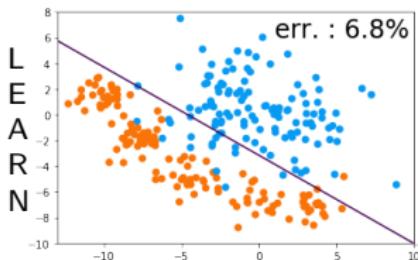
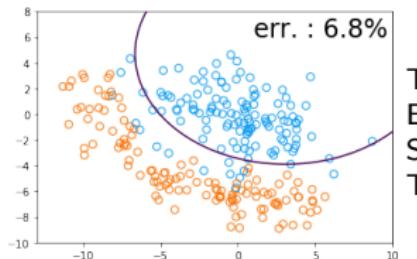
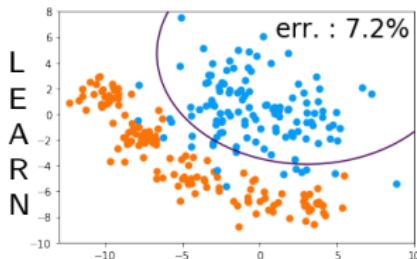
- complexité de la méthode choisie
- variabilité dans  $\mathcal{X}$ 
  - $X$  ne reflète pas tout  $\mathcal{X}$       capacité à généraliser
  - $(X, Y)$  n'est qu'une observation d'un  $(\mathbf{X}, \mathbf{Y}) \sim \text{loi } P$  sur  $\mathcal{X} \times \mathcal{Y}$  inconnue
- le risque réel (probabilité d'erreur)  $R(f)$  n'est pas accessible, on n'accède qu'à  $\hat{R}(f)$ , son estimation sur  $(X, Y)$
- chercher à diminuer  $\hat{R}(f)$  à tout prix nuit à la capacité à généraliser
- compromis...  
dans le choix de  $f$  et ses hyper-paramètres  
par ex. NP vs KNN et  $d$ ,  $K$ , etc





- le risque empirique augmente toujours sur l'ensemble test ?

non !



### Comment Évaluer ?

- partage de  $(X, Y) = (X, Y)_{learn} \cup (X, Y)_{valid} \cup (X, Y)_{test}$ 
  - $(X, Y)_{learn}$  pour construire  $f$
  - $(X, Y)_{valid}$  pour régler  $f$
  - $(X, Y)_{test}$  pour évaluer  $f$

- souvent  $(X, Y)_{learn} = (X, Y)_{learn} \cup (X, Y)_{valid}$
- si  $(X, Y)_{test} = (X, Y)_{learn}$  ?
- échantillonnage **aléatoire**, et **stratifié**

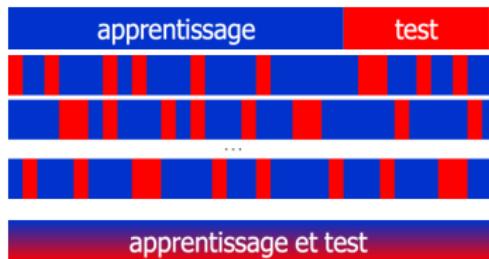
- simulation statistique (validation croisée, *bootstrap*)

- holdout** : par ex. 30% en test
  - problème de représentativité
  - moyenner sur plusieurs tirages
  - 50%, échanger les deux et moyenner

hyper-paramètres  
 $E = \hat{R}(f)$

reclassement (ou *resubstitution*)

si  $n$  est petit



- k-fold cross validation** : par ex.  $k = 5, 10$ 
  - $k$  échantillons de taille  $n/k$  en test, le reste en apprentissage, et on moyenne
  - meilleure représentativité
  - plus  $k$  est grand, plus c'est long... ; diminue le **biais**, augmente la **variance**
  - deux cas particuliers :
    - $k = 2$  (sous-apprentissage)
    - $k = n$  (sur-apprentissage)

recommandé si  $n$  est petit (tous les exemples sont appris et testés)

leave-one-out

## Scores

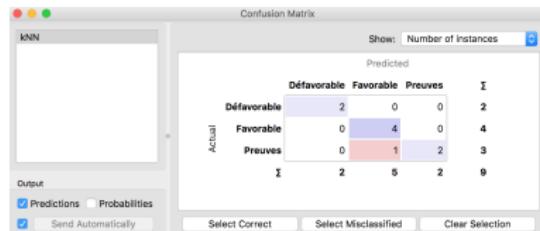
- à partir de la **matrice de confusion**  $C = [c_{ij}]_{i,j=1,m}$  ;  
table de contingence croisant  
 $Y_{test}$  (*actual*) et  $f(X_{test})$  (*predicted*)

$$\text{Classification Accuracy } CA = \frac{\text{trace}(C)}{n_{test}}$$

Exemple : Données *terminale.xlsx*

2-NN, *holdout* 0%

$$CA = 8/9 = 0.889$$



- issus de la classif. binaire, étendus au cas multi-classes originellement, classification binaire pour la séparation signal/bruit :  $f(x) = 1_{x>s}$  ; puis en diagnostic médical

- pour chaque classe  $i = 1, m$ , compter les *True Positives*, *False Pos.*, *True Negatives*, *False Neg.*.

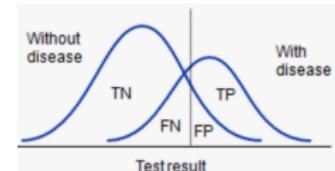
$$\text{Recall } R_j = \frac{TP}{TP+FN}$$

$$\text{Precision } P_j = \frac{TP}{TP+FP}$$

$$\text{False Alarm } FA_j = \frac{FP}{TP+FP} = 1 - P_j$$

$$\text{F1 score } F1_j = 2 \frac{R_j \times P_j}{R_j + P_j}$$

$$\text{Accuracy } A_j = \frac{TP+TN}{TP+FN+FP+TN}$$



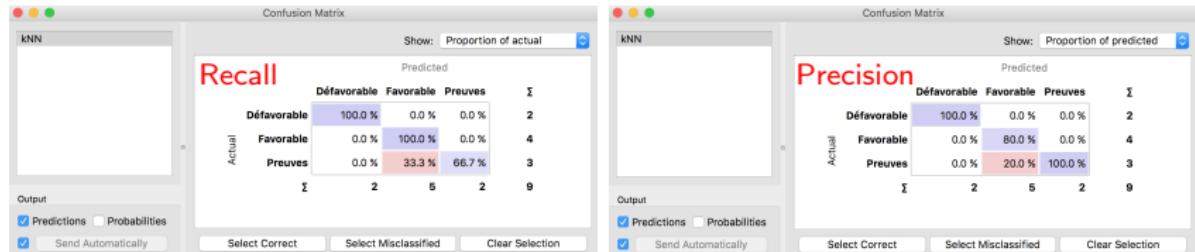
actu./ pred.	positifs	négatifs
positifs (€)	TP	FN
négatifs (€)	FP	TN

## Scores

globaux  $S = R, P, FA, F1, A?$ 

mesures de performance

$$S = \frac{1}{n} \sum_{i=1}^n n_i S_i$$



- issus de la classif. binaire, étendus au cas multi-classes originellement, classification binaire pour la séparation signal/bruit :  $f(x) = 1_{x>s}$  ; puis en diagnostic médical

- pour chaque classe  $i = 1, m$ , compter les *True Positives*, *False Pos.*, *True Negatives*, *False Neg.*.

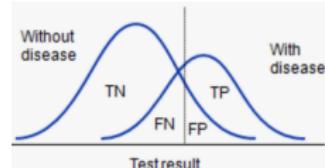
$$\text{Recall } R_j = \frac{TP}{TP+FN}$$

$$\text{Precision } P_j = \frac{TP}{TP+FP}$$

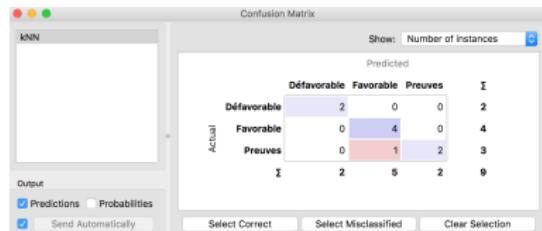
$$\text{False Alarm } FA_j = \frac{FP}{TP+FP} = 1 - P_j$$

$$\text{F1 score } F1_j = 2 \frac{R_j \times P_j}{R_j + P_j}$$

$$\text{Accuracy } A_j = \frac{TP+TN}{TP+FN+FP+TN}$$



actu./ pred.	positifs	négatifs
positifs ( $\in$ )	TP	FN
négatifs ( $\notin$ )	FP	TN

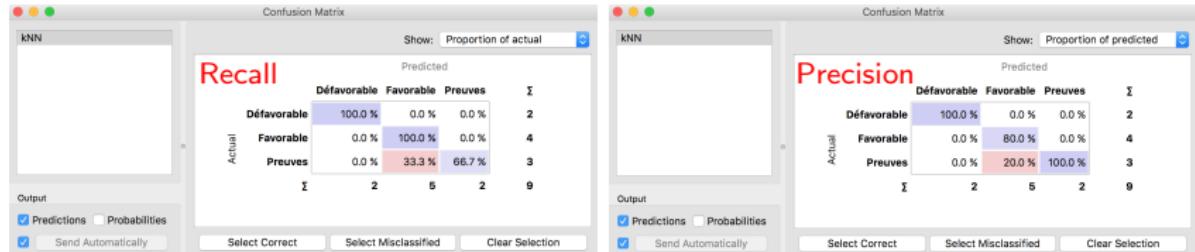


### Scores

globaux  $S = R, P, FA, F1, A?$

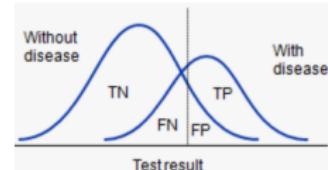
mesures de performance

$$S = \frac{1}{n} \sum_{i=1}^n n_i S_i$$

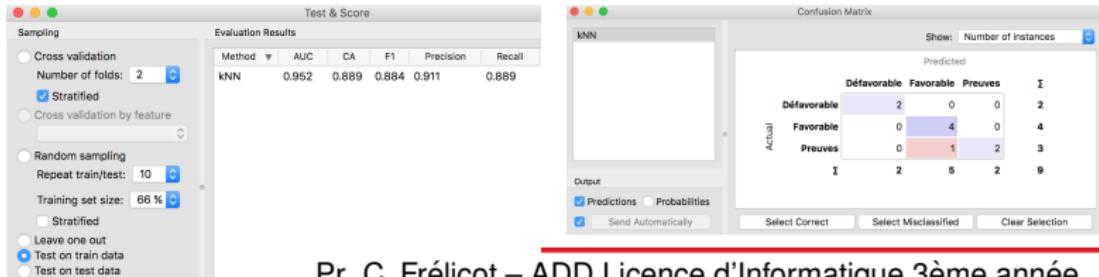


- issus de la classif. binaire, étendus au cas multi-classes originellement, classification binaire pour la séparation signal/bruit :  $f(x) = 1_{x>s}$  ; puis en diagnostic médical

- pour chaque classe  $i = 1, m$ , compter les *True Positives*, *False Pos.*, *True Negatives*, *False Neg.*.



actu./ pred.	positifs	négatifs
positifs ( $\in$ )	TP	FN
négatifs ( $\notin$ )	FP	TN



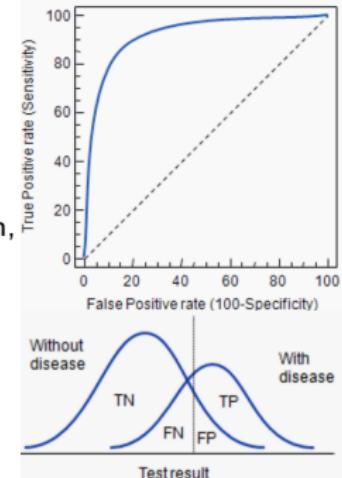
## Courbe ROC

- *TPRate vs FPRate (FARate)* pour diverses valeurs de  $s$  ;
- recherche d'un bon point de fonctionnement du dispositif de détection, par ex : quel  $s$  pour  $TPR > 80\%$

## Area Under Curve (AUC)

- *the more AUC, the more performance*
- comparaison des performances entre règles de classification, ou différents hyper-paramètres de la même règle
- réduction de la dimensionnalité par sélection de variables
- issus de la classif. binaire, étendus au cas multi-classes originellement, classification binaire pour la séparation signal/bruit :  $f(x) = 1_{x>s}$  ; puis en diagnostic médical
- pour chaque classe  $i = 1, m$ , compter les *True Positives, False Pos., True Negatives, False Neg.*

Receiver Operating Characteristics



actu./ pred.	positifs	négatifs
positifs ( $\in$ )	TP	FN
négatifs ( $\notin$ )	FP	TN

Test & Score

**Sampling**

- Cross validation
- Number of folds: 2
- Stratified
- Cross validation by feature

- Random sampling
- Repeat train/test: 10
- Training set size: 66 %
- Stratified
- Leave one out
- Test on train data
- Test on test data

**Evaluation Results**

Method	AUC	CA	F1	Precision	Recall
KNN	0.952	0.889	0.884	0.911	0.889

kNN

**Confusion Matrix**

Show: Number of Instances

		Predicted		$\Sigma$
		Défavorable	Favorable	
Actual	Défavorable	2	0	2
	Favorable	0	4	4
	Preuves	0	1	2
$\Sigma$	2	5	9	

**Output**

Prediction  Probabilities

Send Automatically

Select Correct Select Misclassified Clear Selection



1. Préambule
2. Tableaux et Espaces
3. Réduction de la Dimensionnalité
4. Apprentissage Non Supervisé
5. Apprentissage Supervisée
  - 5.1 Problématique(s)
  - 5.2 Nearest-Prototype
    - 5.2.1 NP Rule
    - 5.2.2 Pros and Cons
    - 5.2.3 Exemple sur Données Réelles
  - 5.3 Nearest-Neighbors
    - 5.3.1 NN Rule
    - 5.3.2 K-NN Rule
    - 5.3.3 Exemple sur Données Réelles
  - 5.4 Évaluation