DEVLOG RÉSEAU DU DEVELOPPEMENT LOGICIEL Recherche et Enseignement Supérieur

cnrs dépasser les frontières www.cnrs.fr

INSTITUT de MATHEMATIQUES de TOULOUSE

ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

cimi Centre International de Mathématiques et d'Informatique TOULOUSE

Introduction à l'apprentissage automatique

Laurent Risser
Ingénieur de Recherche à l'Institut de Mathématiques de Toulouse et au 3IA ANITI

lrisser@math.univ-toulouse.fr

Exemples introductifs à l'apprentissage automatique

Aide au diagnostic

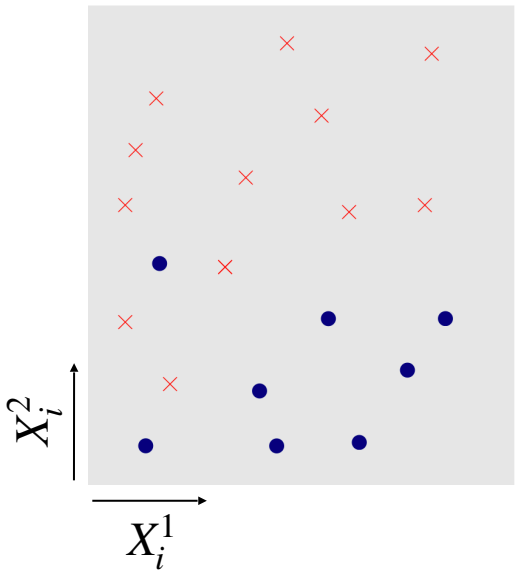
Base d'apprentissage	
Patient 1 : • Age = 40 • Globule Blancs/L = 6	Sain
Patient 2 : • Age = 28 • Globule Blancs/L = 12	Rhume
Patient N : • Age = 57 • Globule Blancs/L = 8	Sain

Nouveau Patient (hors base d'apprentissage) :
- Age = 34
- Globule Blancs/L = 5

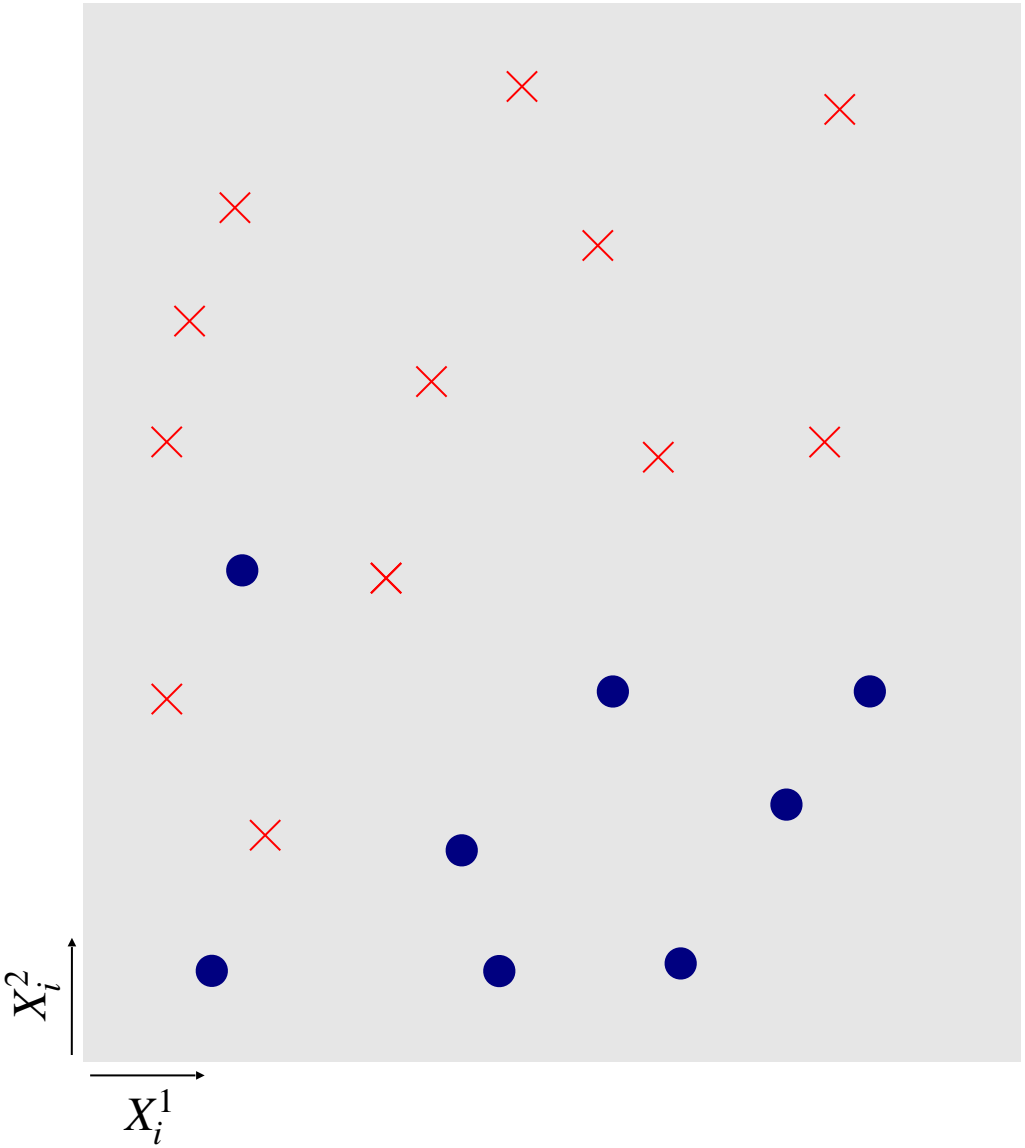
→ Sain ou rhume ???

Base d'apprentissage	
Patient 1 : • Age = 40 • Globule Blancs/L = 6	Sain
Patient 2 : • Age = 28 • Globule Blancs/L = 12	Rhume
Patient n : • Age = 57 • Globule Blancs/L = 8	Sain

	Age	G.B./L.	Etat
Pat. 1	40	6	1
Pat. 2	28	12	0
…	…	…	…
Pat. n	57	8	1



Apprentissage supervisé — classification



Observations d'entrée (*X*) :

- *n* observations $X_i \in \mathbb{R}^p$

Observations de sortie (*Y*) :

- *n* Labels $Y_i \in \{-1,1\}^K$
- × $Y_i = 1$
- ● $Y_i = -1$

- Ici *n = 20* , *p = 2 et K=1*

Dans notre exemple :

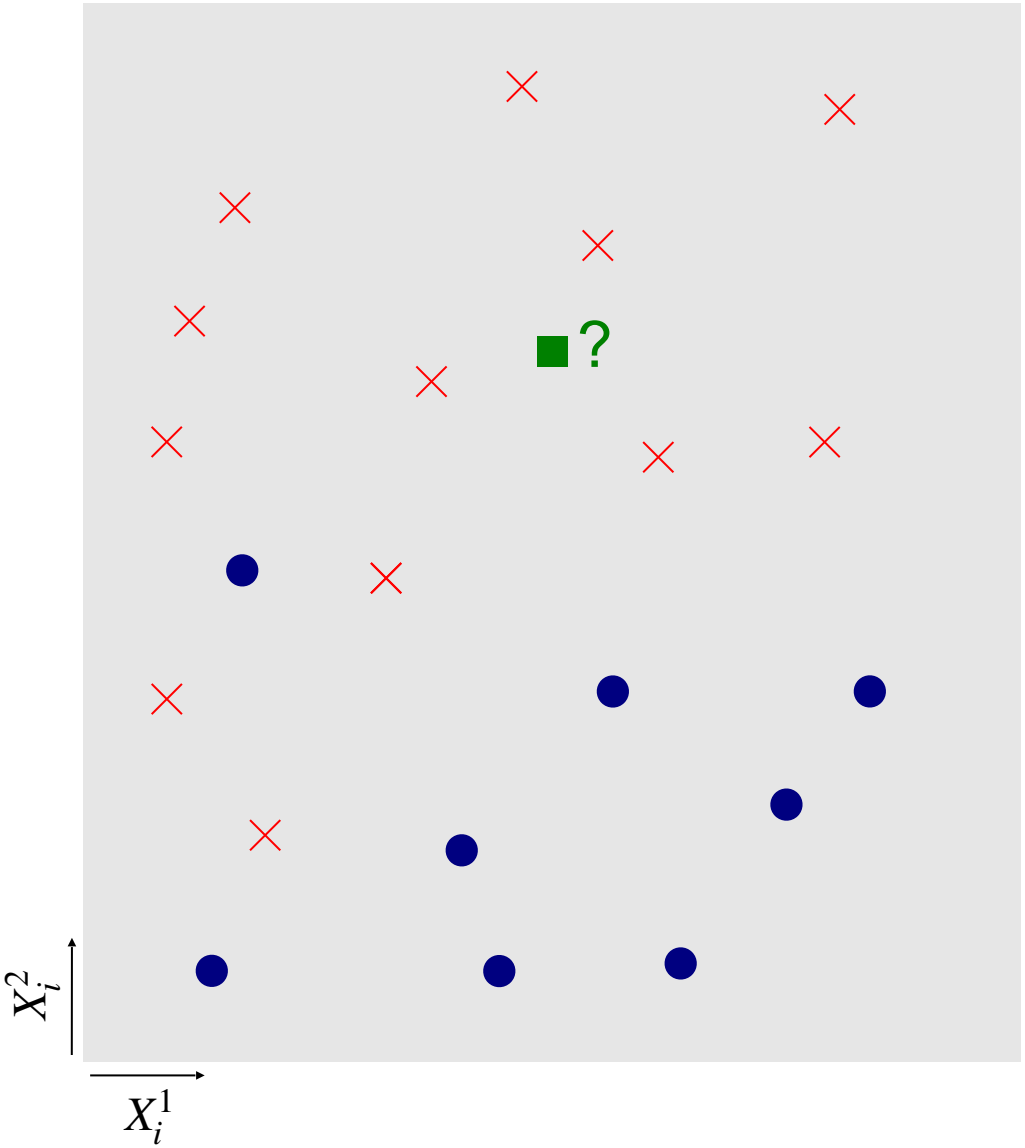
i ⟶ Patient de la base d'apprentissage

X_i^1 ⟶ Age

X_i^2 ⟶ Globule Blancs/L

Y_i ⟶ Sain ou rhume

Apprentissage supervisé — classification



Observations d'entrée (X) :

- n observations $X_i \in \mathbb{R}^p$

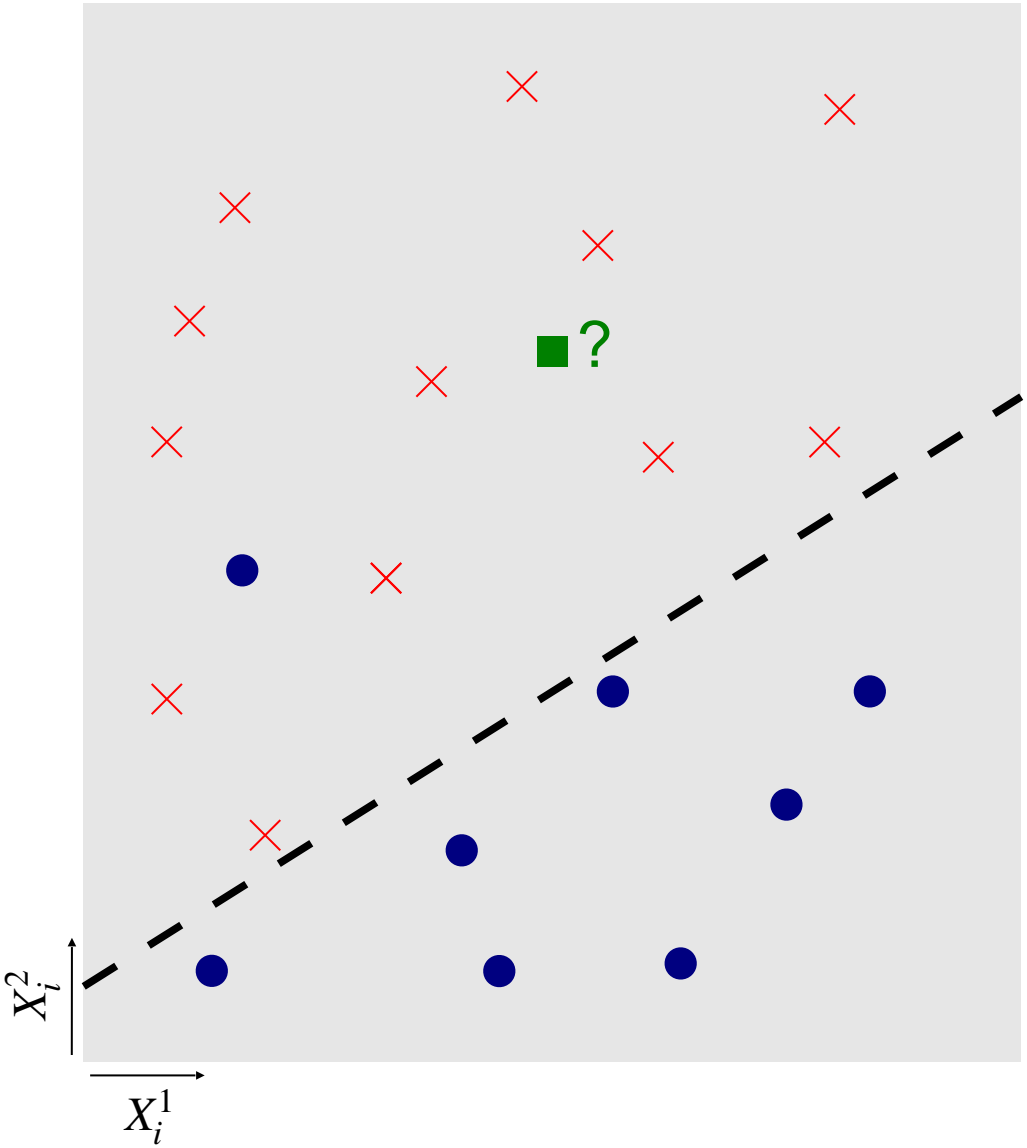
Observations de sortie (Y) :

- n Labels $Y_i \in \{-1,1\}^K$
- × $Y_i = 1$
- ● $Y_i = -1$

- Ici $n = 20$, $p = 2$ et $K=1$

Label le plus probable de ■ ?

Apprentissage supervisé — classification



Observations d'entrée (*X*) :

- *n* observations $X_i \in \mathbb{R}^p$

Observations de sortie (*Y*) :

- *n* Labels $Y_i \in \{-1,1\}^K$
- × $Y_i = 1$
- ● $Y_i = -1$

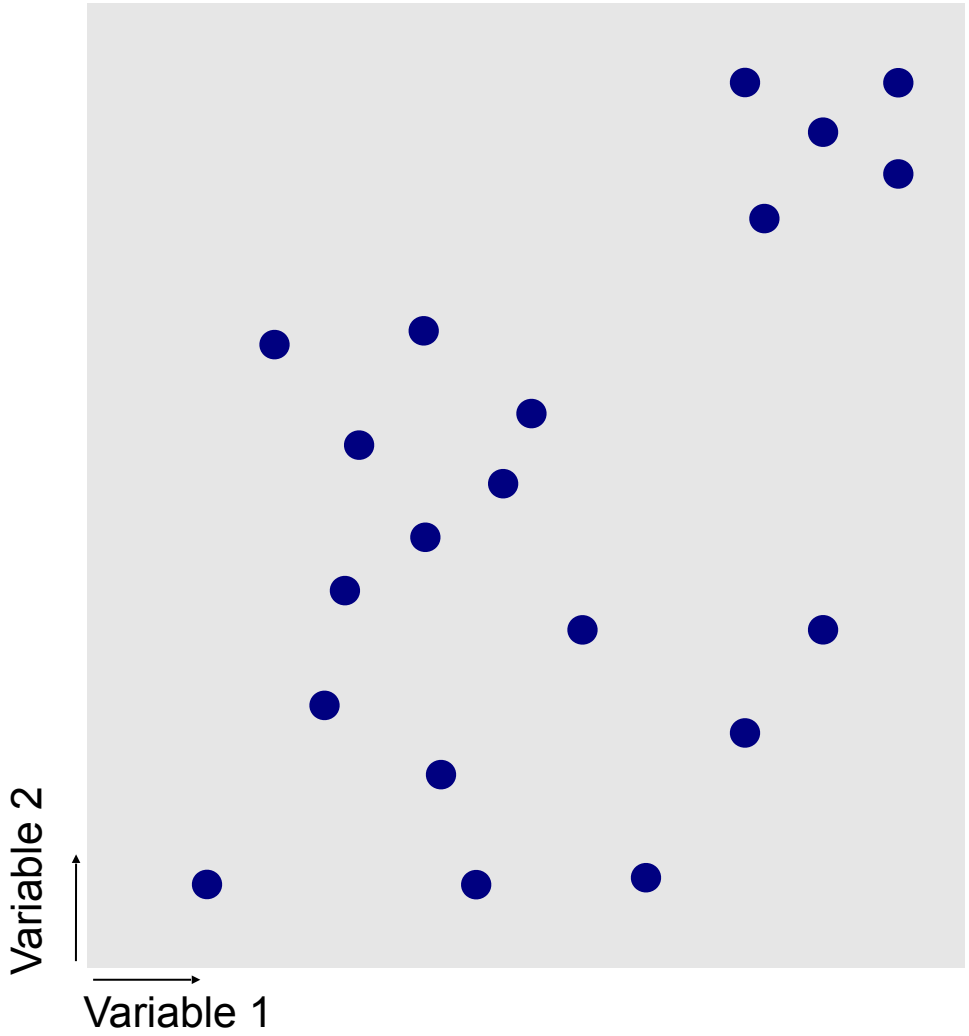
- Ici *n = 20* , *p = 2 et K=1*

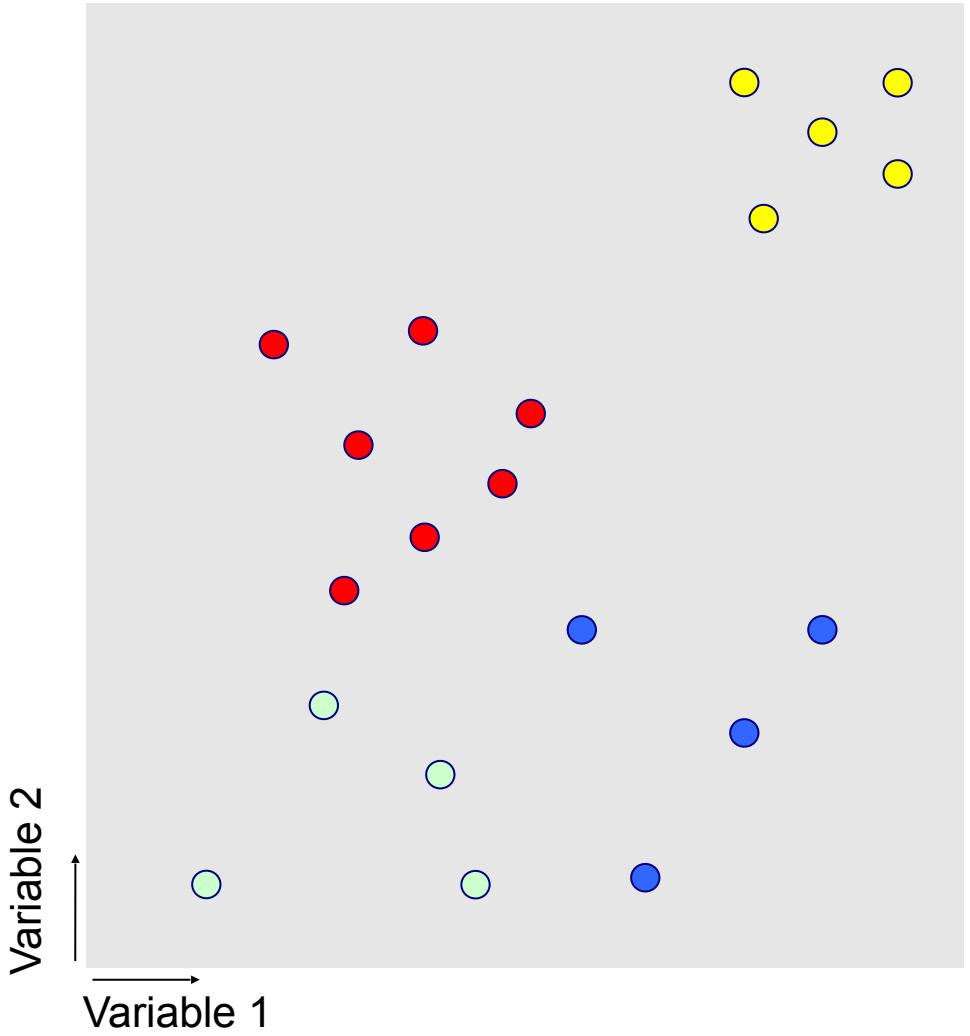
1. **Choix d'un modèle** pour séparer les données d'apprentissage, i.e. les ● et les × .
2. **Apprentissage des paramètres** optimaux
3. Une fois les paramètres du modèle appris, **prédiction** extrêmement simple et rapide de ■ .

Pour résumer

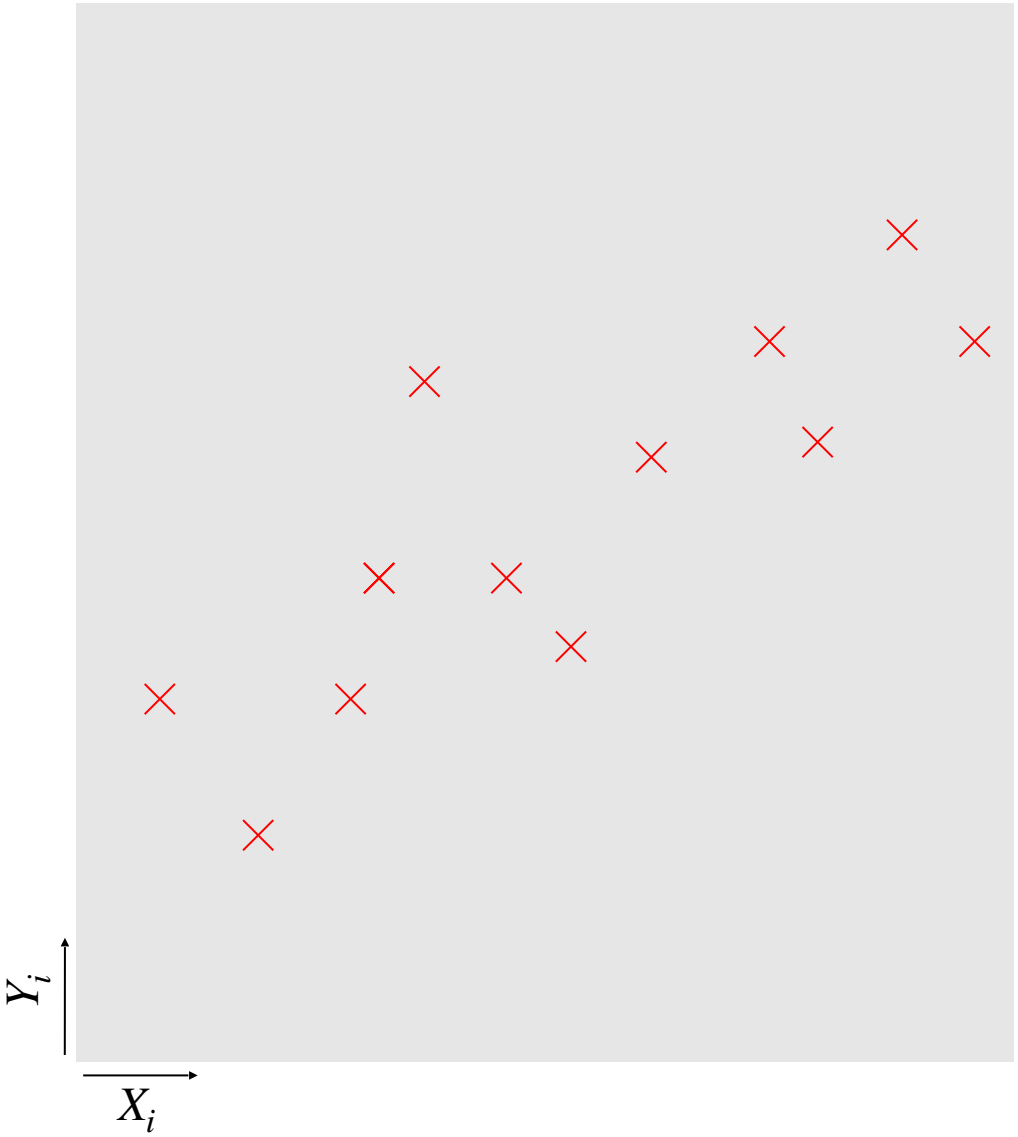
1. On dispose de ***données d'apprentissage annotées***
2. Définition d'un ***modèle*** pour séparer les données
3. Optimisation des paramètres du modèle (***apprentissage***) en fonction d'un critère de *risque empirique* ($\approx$ erreur moyenne)
4. Idéalement : ***validation*** du modèle sur données test
5. ***Prédiction*** sur de nouvelles observation

Imitation game (2014) — Vie d'Alan Turing


Variable 2
Variable 1


Variable 2
Variable 1

Apprentissage supervisé — *Regression*



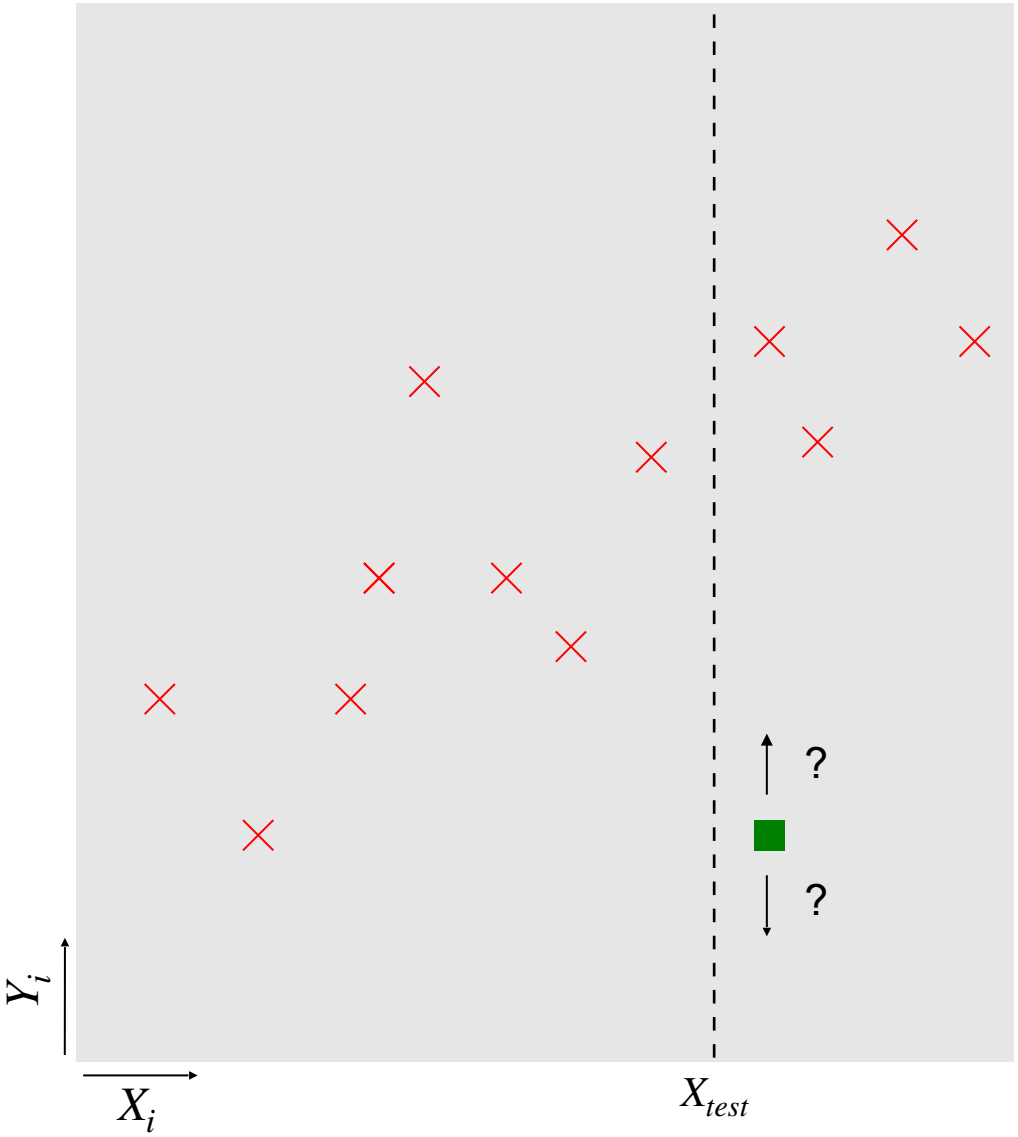
Observations d'entrée (*X*) :

- *n* observations $X_i \in \mathbb{R}^p$

Observations de sortie (*Y*) :

- $Y_i \in \mathbb{R}^K$
- Ici *n* = *20* , *p* = *1* et *K=1*

Apprentissage supervisé — *Regression*



Observations d'entrée (*X*) :

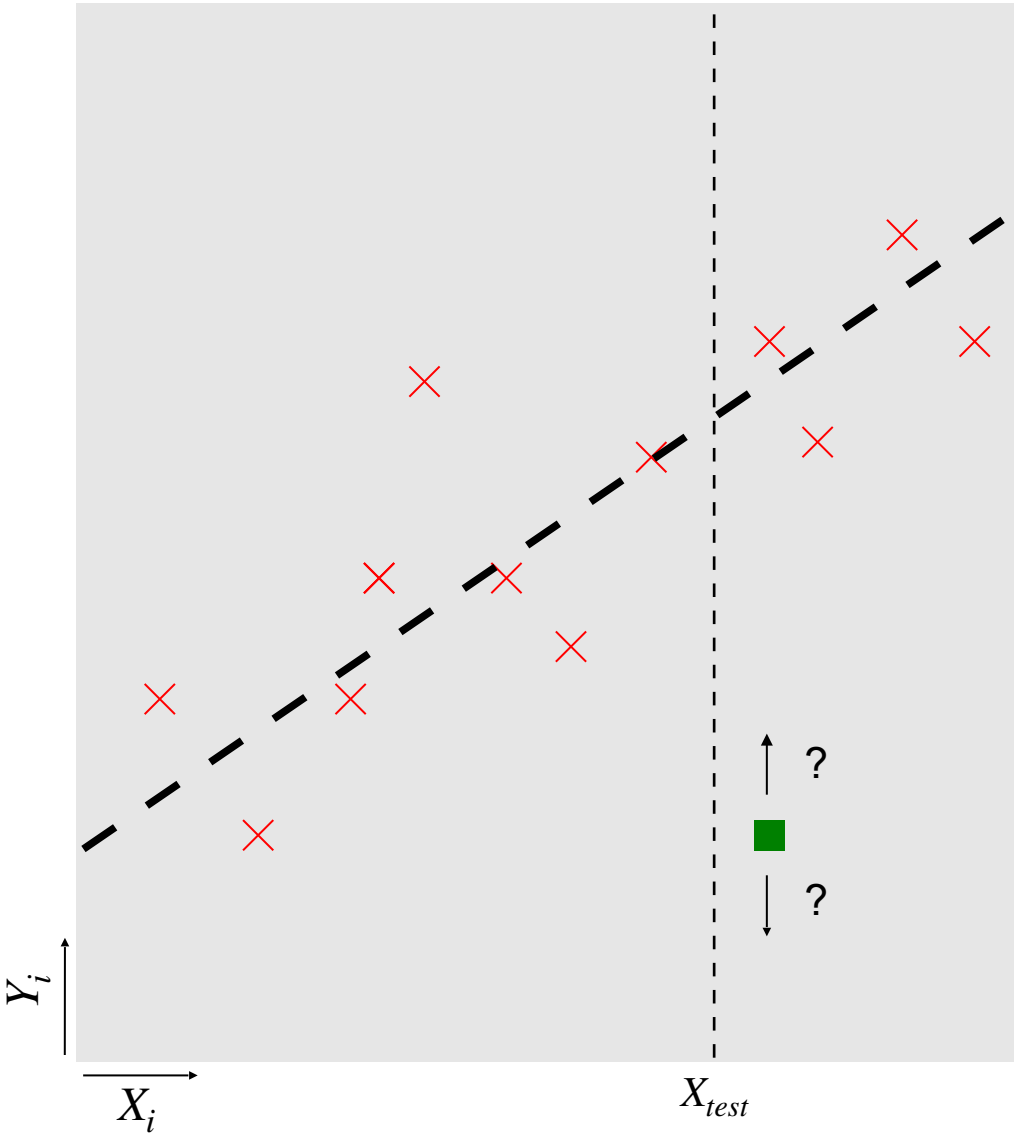
- *n* observations $X_i \in \mathbb{R}^p$

Observations de sortie (*Y*) :

- $Y_i \in \mathbb{R}^K$
- Ici *n* = *20* , *p* = *1* et *K*=*1*

Prédiction de Y_{test} pour un X_{test} donné ?

Apprentissage supervisé — *Regression*



Observations d'entrée (*X*) :

- *n* observations $X_i \in \mathbb{R}^p$

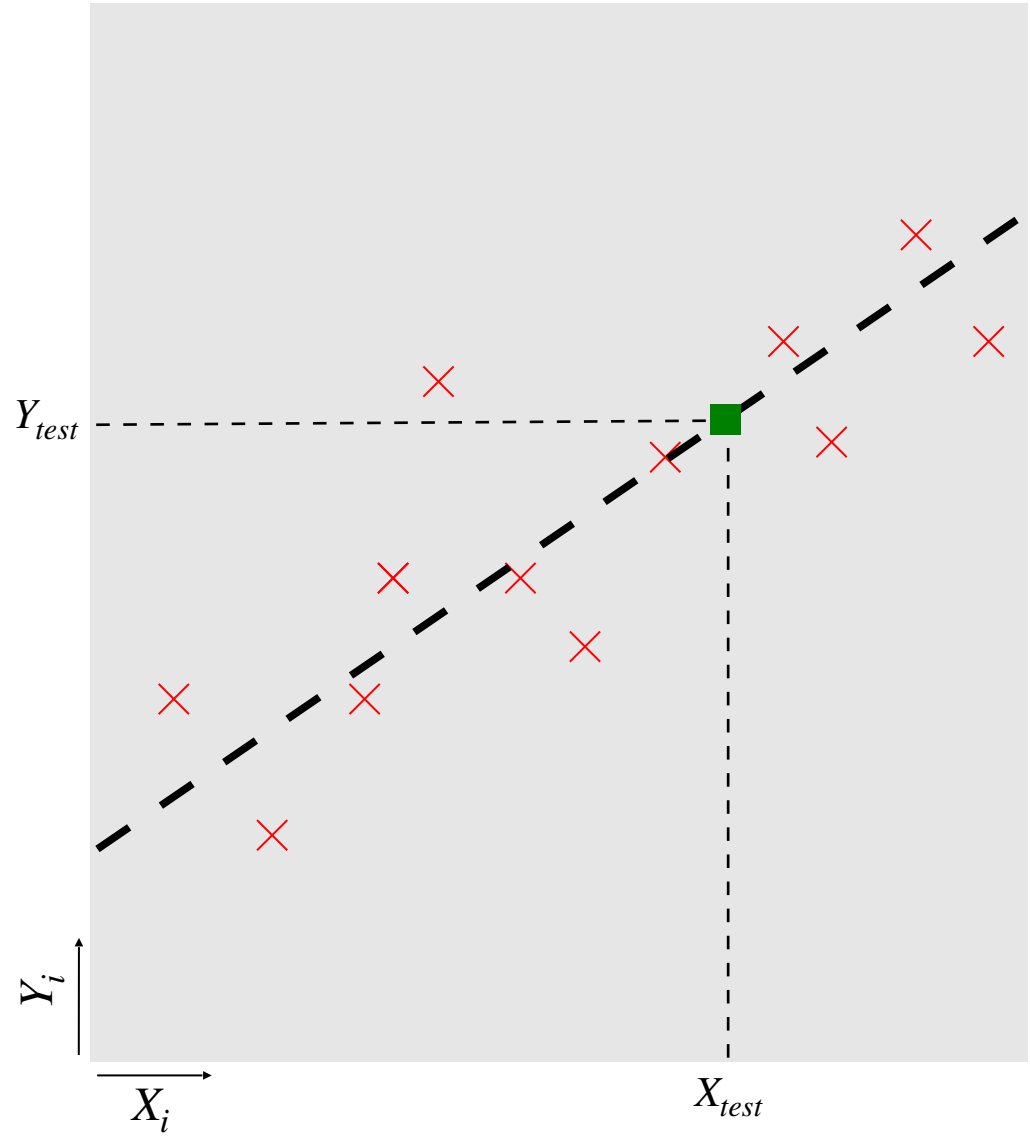
Observations de sortie (*Y*) :

- $Y_i \in \mathbb{R}^K$
- Ici *n* = *20* , *p* = *1* et *K*=*1*

Prédiction de Y_{test} pour un X_{test} donné ?

1. **Choix d'un modèle** pour mettre en lien les X_i et les Y_i.
2. **Apprentissage des paramètres** optimaux

Apprentissage supervisé — *Regression*



Observations d'entrée (*X*) :

- *n* observations $X_i \in \mathbb{R}^p$

Observations de sortie (*Y*) :

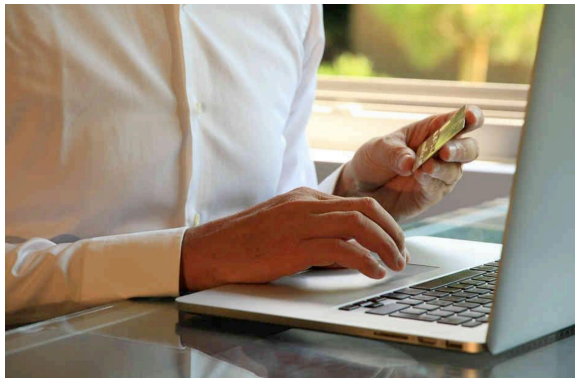
- $Y_i \in \mathbb{R}^K$
- Ici *n* = *20* , *p* = *1* et *K*=*1*

Prédiction de Y_{test} pour un X_{test} donné ?

1. **Choix d'un modèle** pour mettre en lien les X_i et les Y_i.
2. **Apprentissage des paramètres** optimaux
3. Une fois les paramètres du modèle appris, **prédiction** extrêmement simple et rapide.

Vision *actuelle* de l'Intelligence Artificielle :

Applications de *l'apprentissage supervisé* avec des **règles de décisions complexes** et des **bases d'apprentissage massives**.

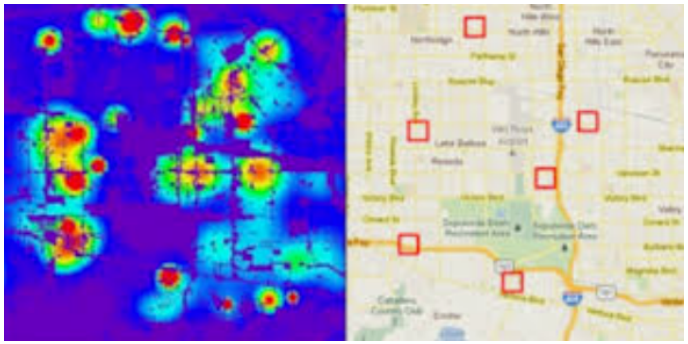


Publicité en ligne

Flux d'information

Aide au diagnostic

Véhicules autonomes



Police prédictive

...

Plan de la présentation

- Exemples introductifs
- Evolution des tendances en science des données
- Algorithmes classiques en apprentissage automatique
  - K-means
  - Arbres de décision et Random forests
  - SVM
  - Régression logistique
  - Réseaux de neurones
- Sur-apprentissage et validation croisée
  - Sur-apprentissage
  - Validation croisée
- Grande dimension et régularisation
  - Modélisation
  - Effet de la régularisation
- Réduction de dimension par ACP
- Conclusion

De la statistique classique à l'apprentissage automatique

Historique[1]:

1940-70 : *Statistiques classiques*. Question associée à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.

1970s : Généralisation des premiers outils informatiques. L'analyse de données explore des données plus volumineuses.

1980s : Les systèmes experts sont supplantés par l'apprentissage automatique et les réseaux de neurones.

1990s : *1er changement de paradigme* : Les données ne sont plus planifiées mais sont préalablement acquises : *From Data Mining to Knowledge Discovery*.

2000s : *2eme changement de paradigme* : Le nombre de variables *p* explose, notamment avec les données omiques où *p>>n*. La qualité de prévision devient plus importante que la réalité du modèle devenu *boîte noire*. Problématique du fléau de la dimension.

2010s : *3eme changement de paradigme* : Le nombre d'observations *n* explose dans le e-commerce, la géo-localisation, … . Bases de données structurées en *cloud* et moyens de calculs regroupés en *clusters (big data)*. La notion de rapidité des algorithmes devient critique.

… 2020s : *4eme changement de paradigme ?* : L'utilisation massive, dans la société et l'industrie, de solutions d'I.A. basées sur de l'apprentissage pose des problématiques sociales et légales. Notions d'explicabilité, de loyauté et de certificabilité de l'I.A.

[1]Cours Apprentissage Statistique de P. Besse (INSA Toulouse): http://www.math.univ-toulouse.fr/~besse/enseignement.html

Historique[1]:

1940-70 : *Statistiques classiques*. Question associée à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.

1970s : Généralisation des premiers outils informatiques. L'analyse de données explore des données plus volumineuses.

1980s : Les systèmes experts sont supplantés par l'apprentissage automatique et les réseaux de neurones.

1990s : *1er changement de paradigme* : Les données ne sont plus planifiées mais sont préalablement acquises : *From Data Mining to Knowledge Discovery*.

2000s : *2eme changement de paradigme* : Le nombre de variables *p* explose, notamment avec les données omiques où *p>>n*. La qualité de prévision devient plus importante que la réalité du modèle devenu *boîte noire*. Problématique du fléau de la dimension.

2010s : *3eme changement de paradigme* : Le nombre d'observations *n* explose dans le e-commerce, la géo-localisation, … . Bases de données structurées en *cloud* et moyens de calculs regroupés en *clusters (big data)*. La notion de rapidité des algorithmes devient critique.

… 2020s : *4eme changement de paradigme ?* : L'utilisation massive, dans la société et l'industrie, de solutions d'I.A. basées sur de l'apprentissage pose des problématiques sociales et légales. Notions d'explicabilité, de loyauté et de certificabilité de l'I.A.

[1]Cours Apprentissage Statistique de P. Besse (INSA Toulouse): http://www.math.univ-toulouse.fr/~besse/enseignement.html

Historique[1]:

1940-70 : *Statistiques classiques*. Question associée à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.

1970s : Généralisation des premiers outils informatiques. L'analyse de données explore des données plus volumineuses.

1980s : Les systèmes experts sont supplantés par l'apprentissage automatique et les réseaux de neurones.

1990s : *1er changement de paradigme* : Les données ne sont plus planifiées mais sont préalablement acquises : *From Data Mining to Knowledge Discovery*.

2000s : *2eme changement de paradigme* : Le nombre de variables *p* explose, notamment avec les données omiques où $p>>n$. La qualité de prévision devient plus importante que la réalité du modèle devenu *boîte noire*. Problématique du fléau de la dimension.

2010s : *3eme changement de paradigme* : Le nombre d'observations *n* explose dans le e-commerce, la géo-localisation, … . Bases de données structurées en *cloud* et moyens de calculs regroupés en *clusters (big data)*. La notion de rapidité des algorithmes devient critique.

… 2020s : *4eme changement de paradigme ?* : L'utilisation massive, dans la société et l'industrie, de solutions d'I.A. basées sur de l'apprentissage pose des problématiques sociales et légales. Notions d'explicabilité, de loyauté et de certificabilité de l'I.A.

[1]Cours Apprentissage Statistique de P. Besse (INSA Toulouse): http://www.math.univ-toulouse.fr/~besse/enseignement.html

Historique[1]:

1940-70 : *Statistiques classiques*. Question associée à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.

1970s : Généralisation des premiers outils informatiques. L'analyse de données explore des données plus volumineuses.

1980s : Les systèmes experts sont supplantés par l'apprentissage automatique et les réseaux de neurones.

1990s : *1er changement de paradigme* : Les données ne sont plus planifiées mais sont préalablement acquises : *From Data Mining to Knowledge Discovery*.

2000s : *2eme changement de paradigme* : Le nombre de variables *p* explose, notamment avec les données omiques où $p>>n$. La qualité de prévision devient plus importante que la réalité du modèle devenu *boîte noire*. Problématique du fléau de la dimension.

2010s : *3eme changement de paradigme* : Le nombre d'observations *n* explose dans le e-commerce, la géo-localisation, … . Bases de données structurées en *cloud* et moyens de calculs regroupés en *clusters (big data)*. La notion de rapidité des algorithmes devient critique.

… 2020s : *4eme changement de paradigme ?* : L'utilisation massive, dans la société et l'industrie, de solutions d'I.A. basées sur de l'apprentissage pose des problématiques sociales et légales. Notions d'explicabilité, de loyauté et de certificabilité de l'I.A.

[1]Cours Apprentissage Statistique de P. Besse (INSA Toulouse): http://www.math.univ-toulouse.fr/~besse/enseignement.html

Historique[1]:

1940-70 : *Statistiques classiques*. Question associée à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.

1970s : Généralisation des premiers outils informatiques. L'analyse de données explore des données plus volumineuses.

1980s : Les systèmes experts sont supplantés par l'apprentissage automatique et les réseaux de neurones.

1990s : *1er changement de paradigme* : Les données ne sont plus planifiées mais sont préalablement acquises : *From Data Mining to Knowledge Discovery*.

2000s : *2eme changement de paradigme* : Le nombre de variables *p* explose, notamment avec les données omiques où $p>>n$. La qualité de prévision devient plus importante que la réalité du modèle devenu *boîte noire*. Problématique du fléau de la dimension.

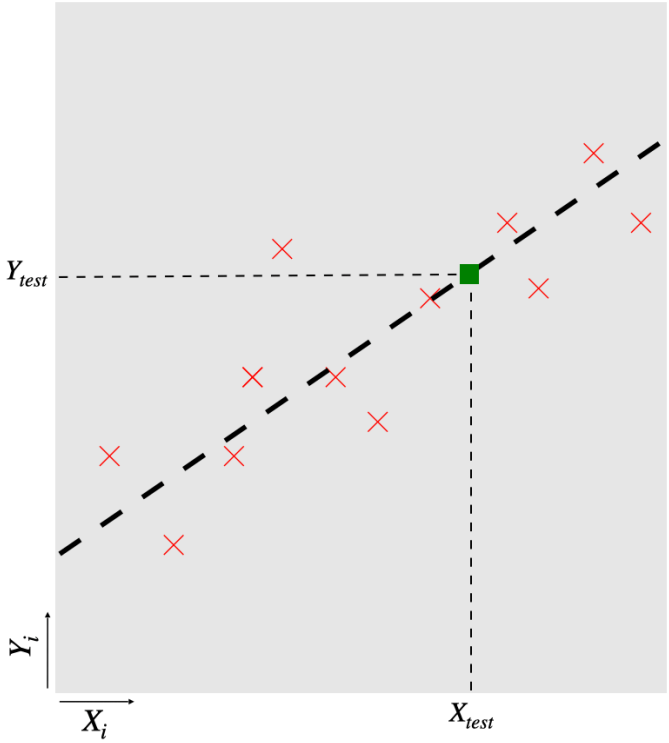
2010s : *3eme changement de paradigme* : Le nombre d'observations *n* explose dans le e-commerce, la géo-localisation, … . Bases de données structurées en *cloud* et moyens de calculs regroupés en *clusters (big data)*. La notion de rapidité des algorithmes devient critique.

… 2020s : *4eme changement de paradigme ?* : L'utilisation massive, dans la société et l'industrie, de solutions d'I.A. basées sur de l'apprentissage pose des problématiques sociales et légales. Notions d'explicabilité, de loyauté et de certificabilité de l'I.A.

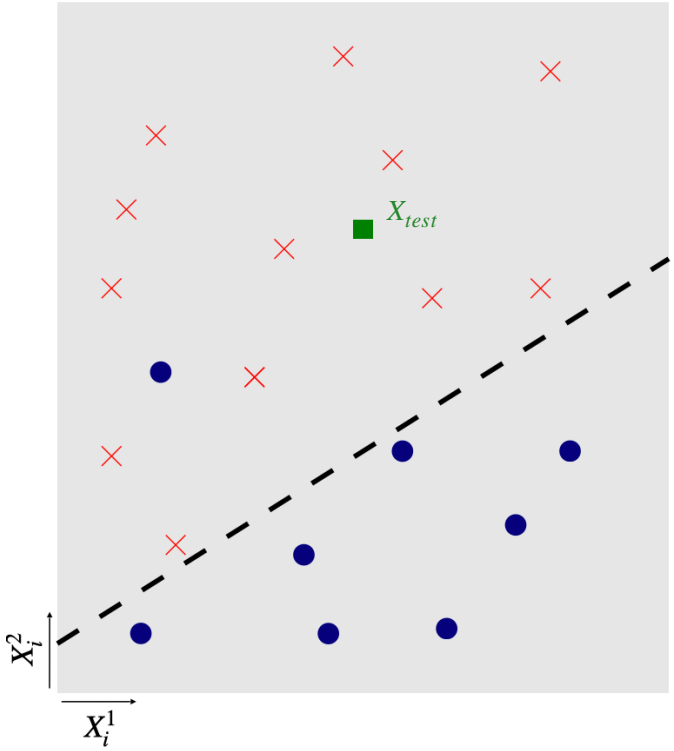
[1]Cours Apprentissage Statistique de P. Besse (INSA Toulouse): http://www.math.univ-toulouse.fr/~besse/enseignement.html

Modèles classiques en apprentissage automatique

Nous avons vu en introduction deux modèles d'apprentissage parmi les plus classiques …



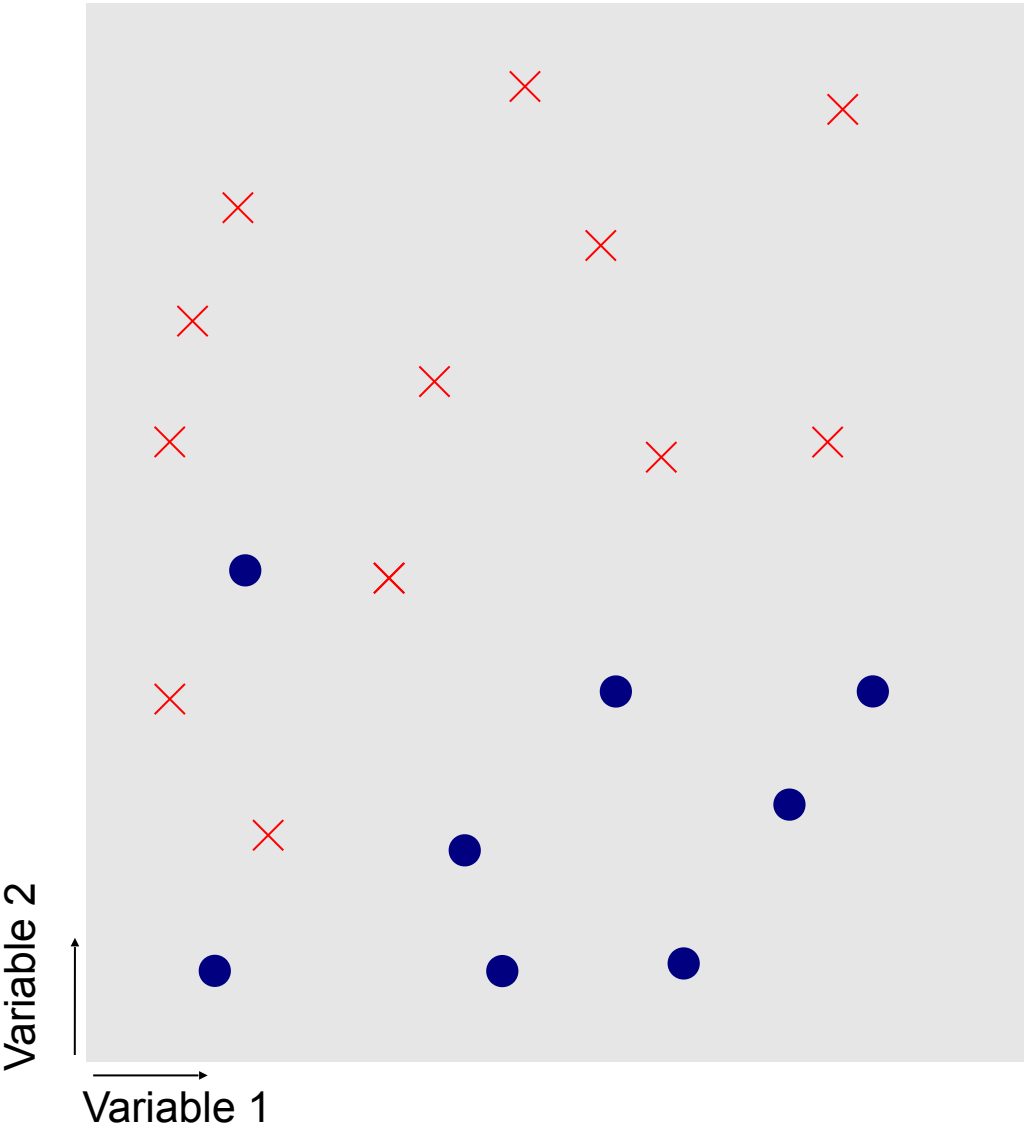
Regression linéaire
(regression - supervisé)



Regression logistique
(classification - supervisé)

$$< x_{test}, w > + b = \left[x_{test}^1, x_{test}^2, \cdots, x_{test}^p\right] \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_p \end{bmatrix} + b = \left(\sum_{j=1}^{p} x_{test}^j w_j \right) + b$$

Arbres de décision (classification - supervisé)

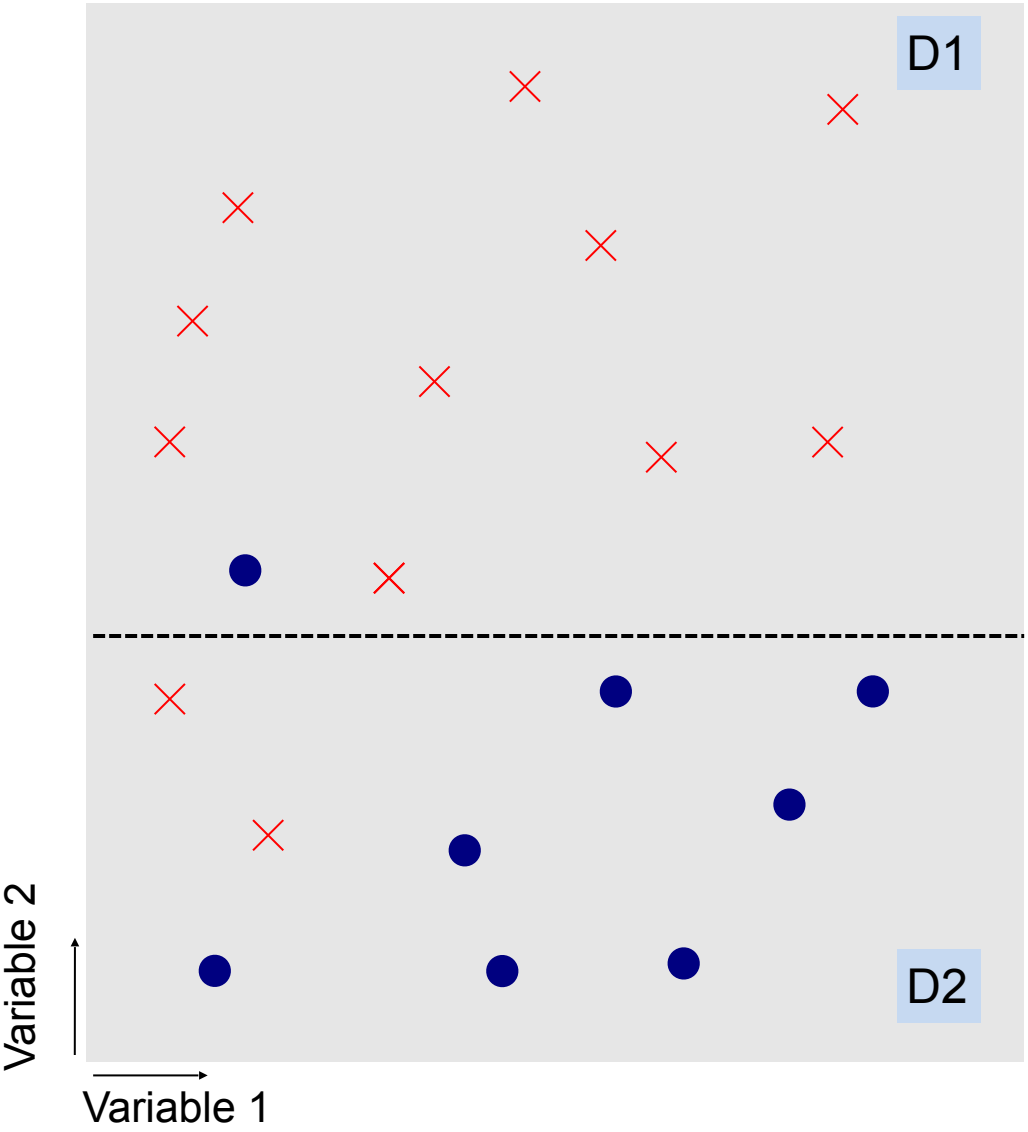


$\boldsymbol{x_1}$, $\boldsymbol{x_2}$, … , $\boldsymbol{x_N}$ les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × ➔ 1 et ● ➔ -1)

On coupe le domaine en *sous domaines* pour minimiser la variance dans chaque sous domaine (CART).

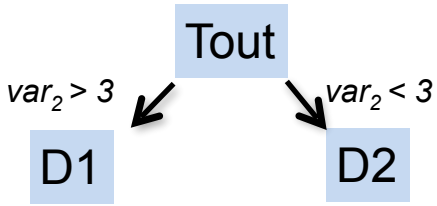
Arbres de décision (classification - supervisé)



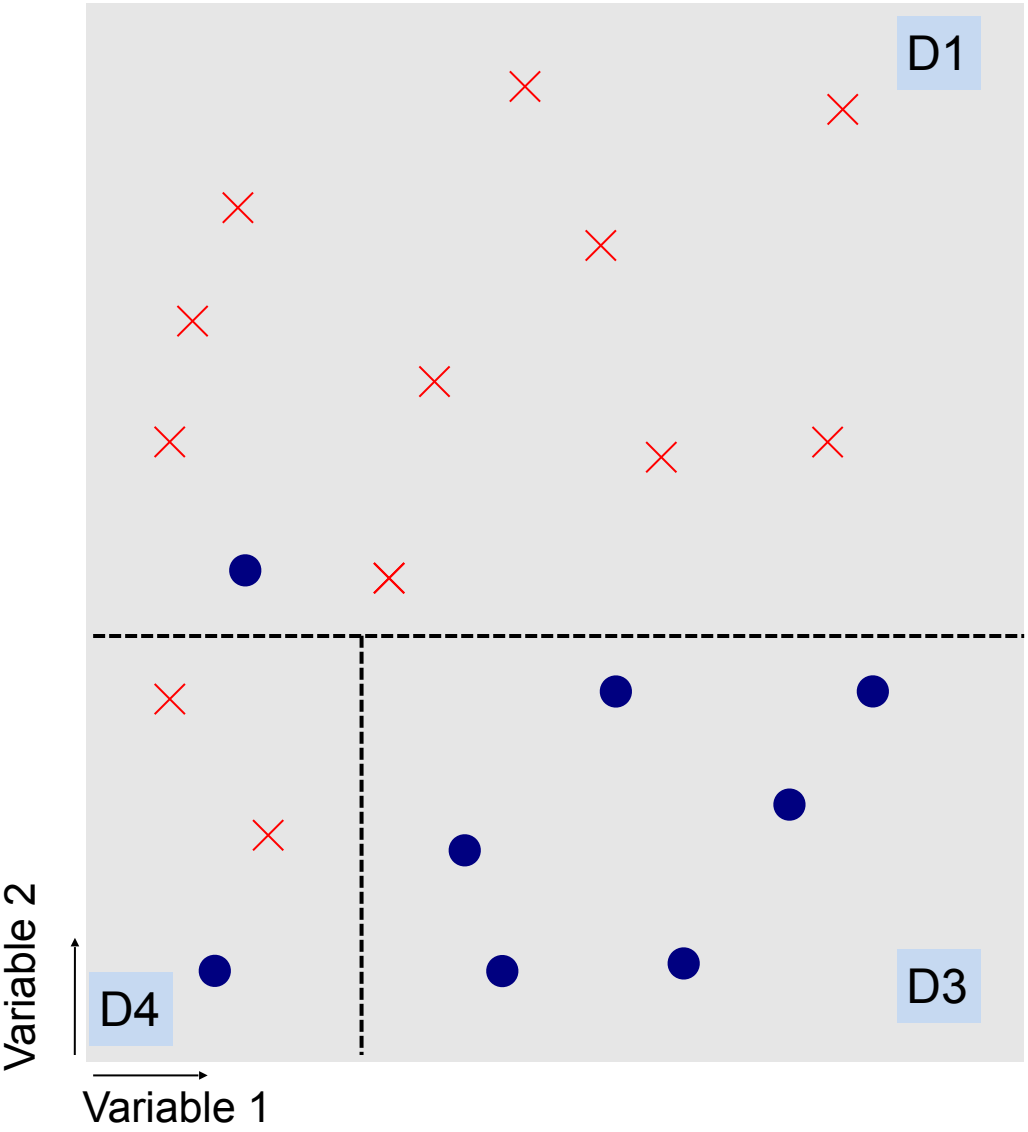
$\boldsymbol{x_1}$, $\boldsymbol{x_2}$, … , $\boldsymbol{x_N}$ les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × → 1 et ● → -1)

On coupe le domaine en *sous domaines* pour minimiser la variance dans chaque sous domaine (CART).



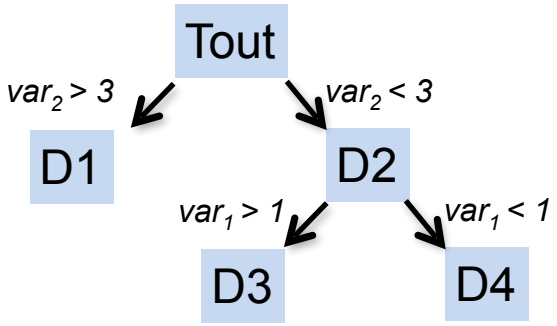
Arbres de décision (classification - supervisé)



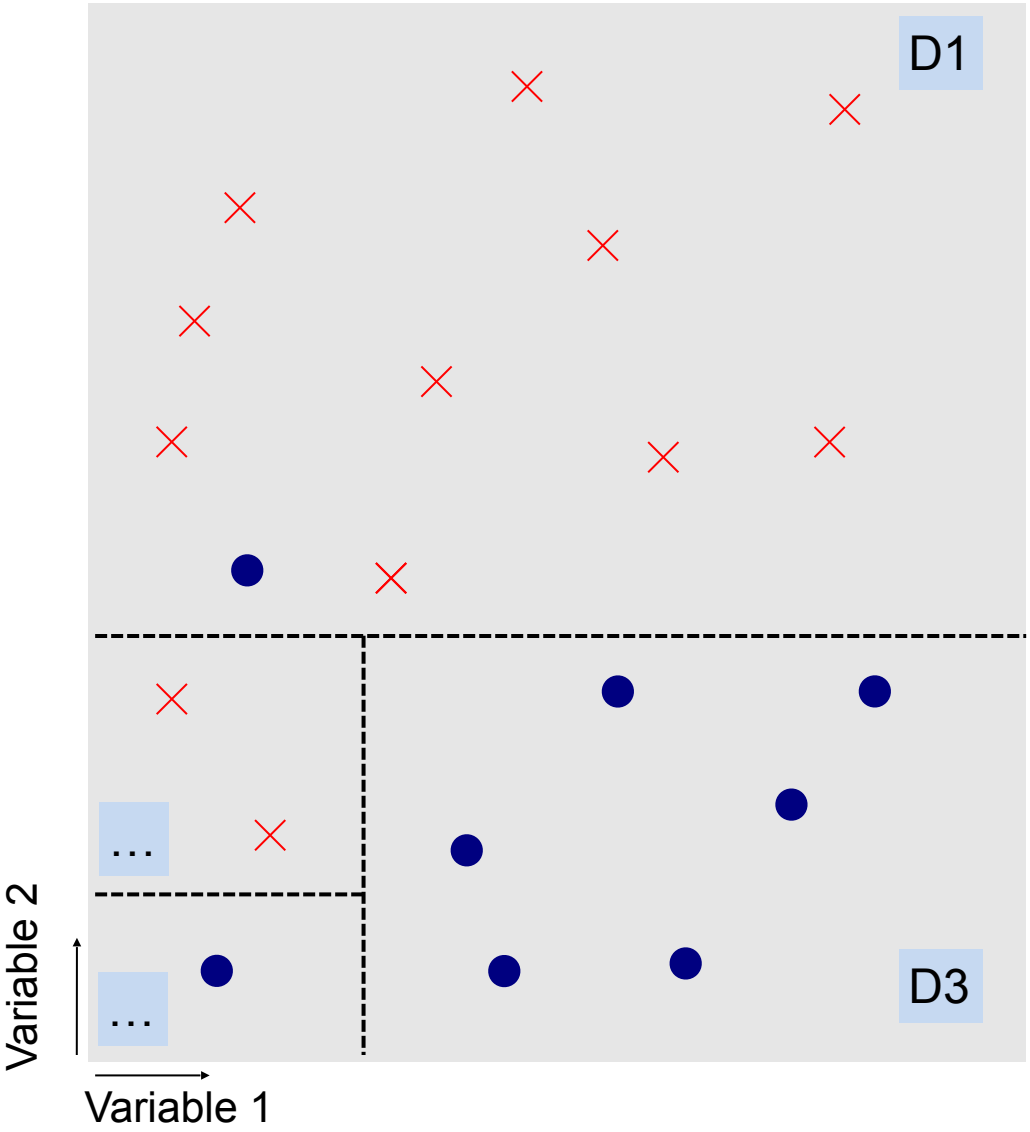
$\boldsymbol{x_1}$, $\boldsymbol{x_2}$, … , $\boldsymbol{x_N}$ les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × → 1 et ● → -1)

On coupe le domaine en *sous domaines* pour minimiser la variance dans chaque sous domaine (CART).



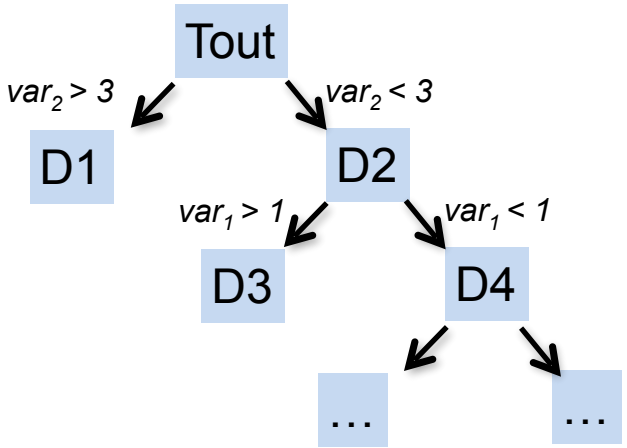
Arbres de décision (classification - supervisé)



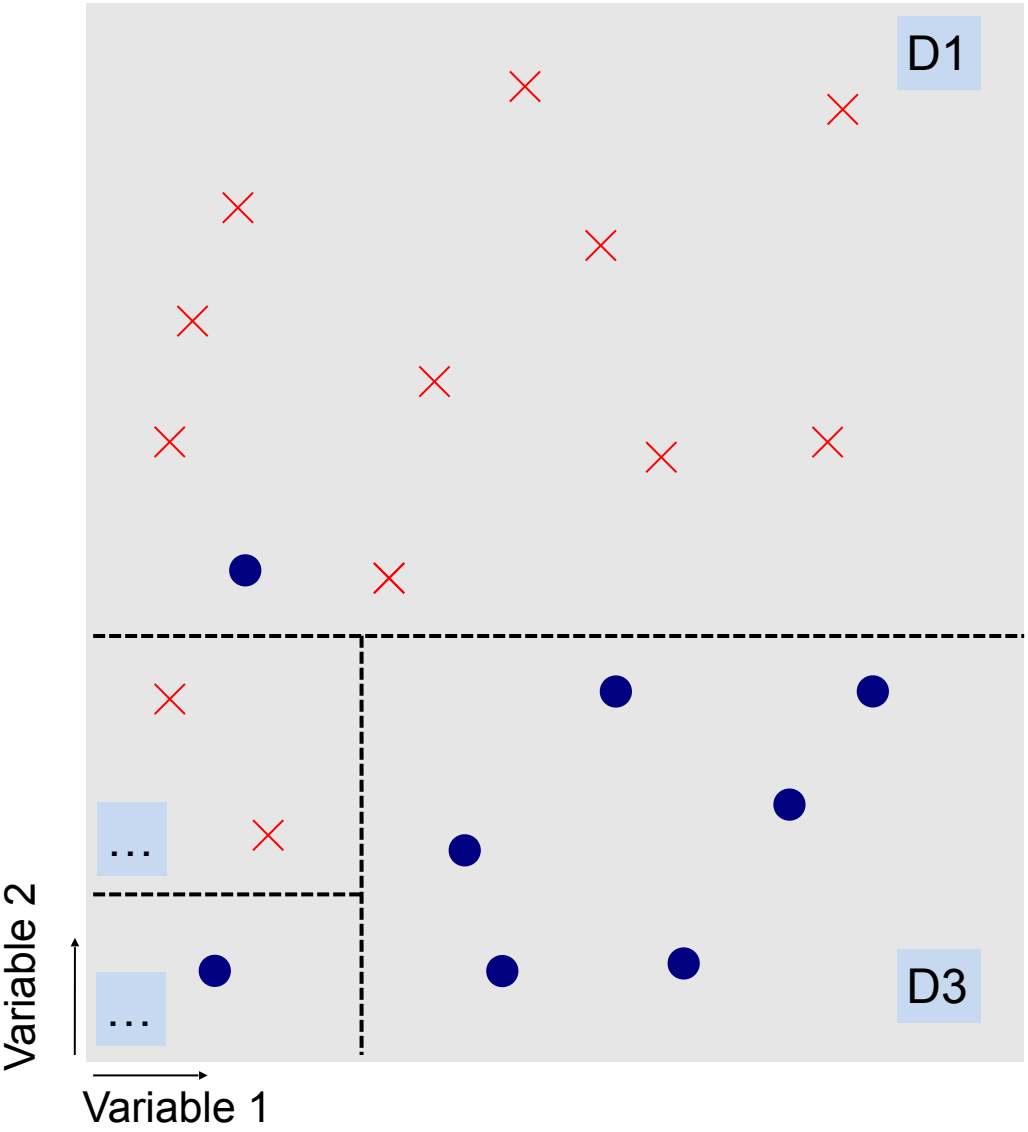
$\boldsymbol{x_1}$, $\boldsymbol{x_2}$, … , $\boldsymbol{x_N}$ les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × → 1 et ● → -1)

On coupe le domaine en *sous domaines* pour minimiser la variance dans chaque sous domaine (CART).



Arbres de décision (classification - supervisé)



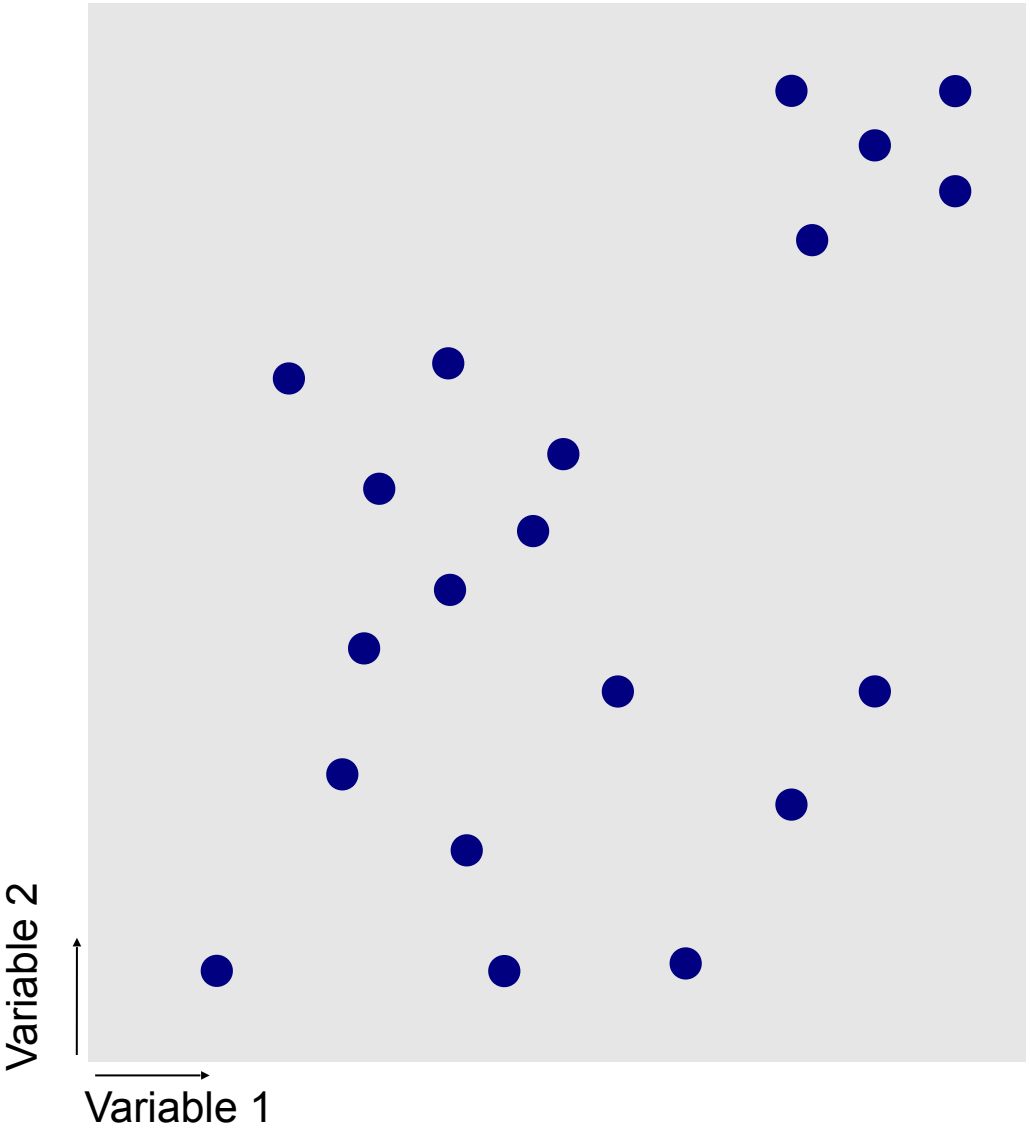
$\boldsymbol{x_1}$, $\boldsymbol{x_2}$, … , $\boldsymbol{x_N}$ les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × ➔ 1 et ● ➔ -1)

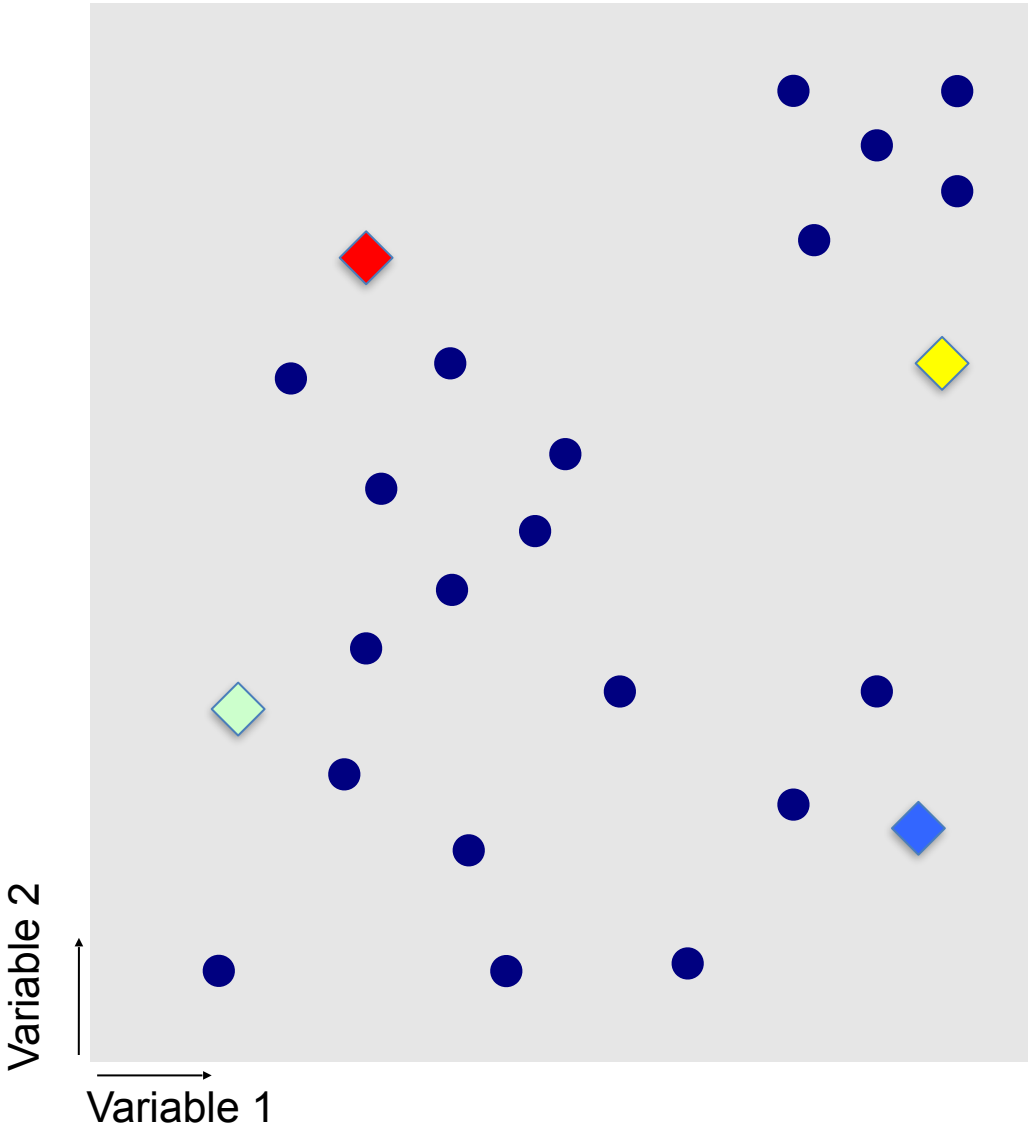
On coupe le domaine en *sous domaines* pour minimiser la variance dans chaque sous domaine (CART).

Principe des random forests :
- Construction de plusieurs arbres
- Variables de coupes tirées au hasard dans chaque arbre
- Agrégation des prédictions

Algorithme des K-means (classification - non supervisé)

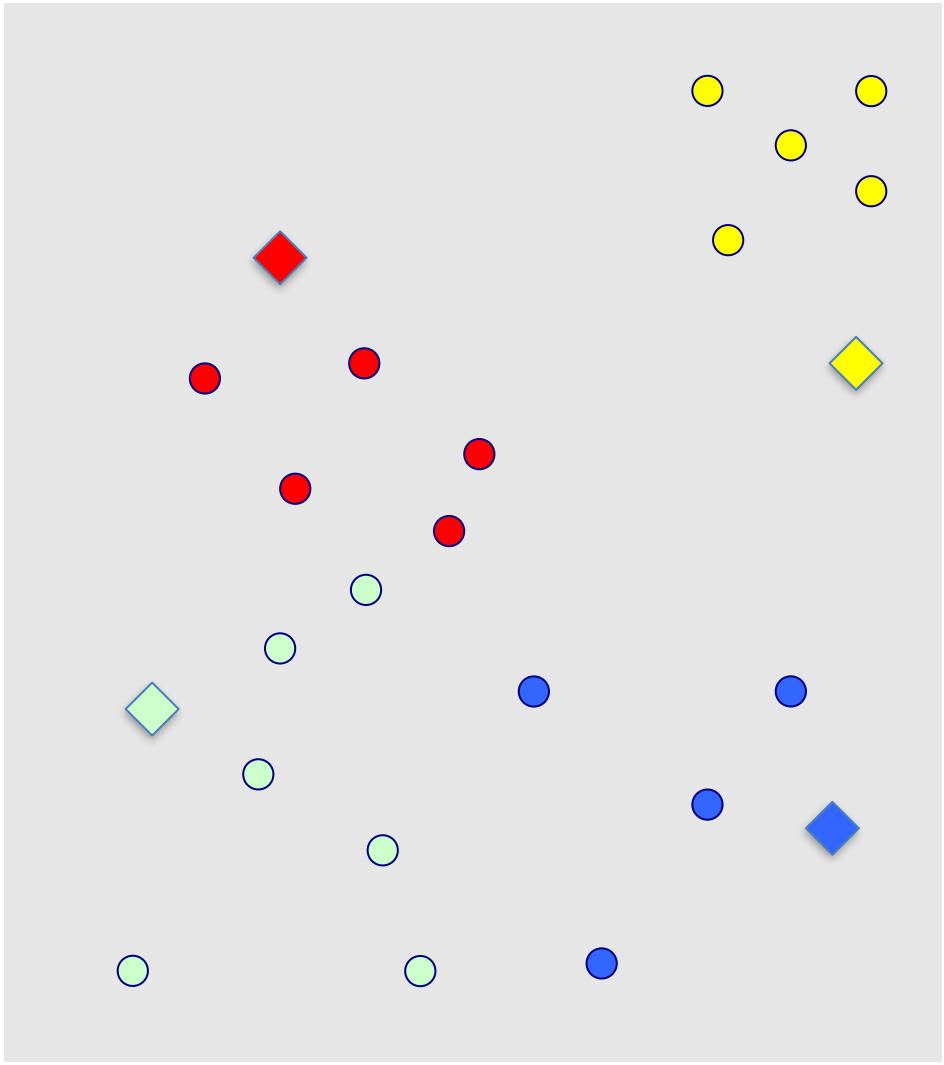


Algorithme des K-means (classification - non supervisé)



On tire K graines au hasard
(pour l'exemple K=4)

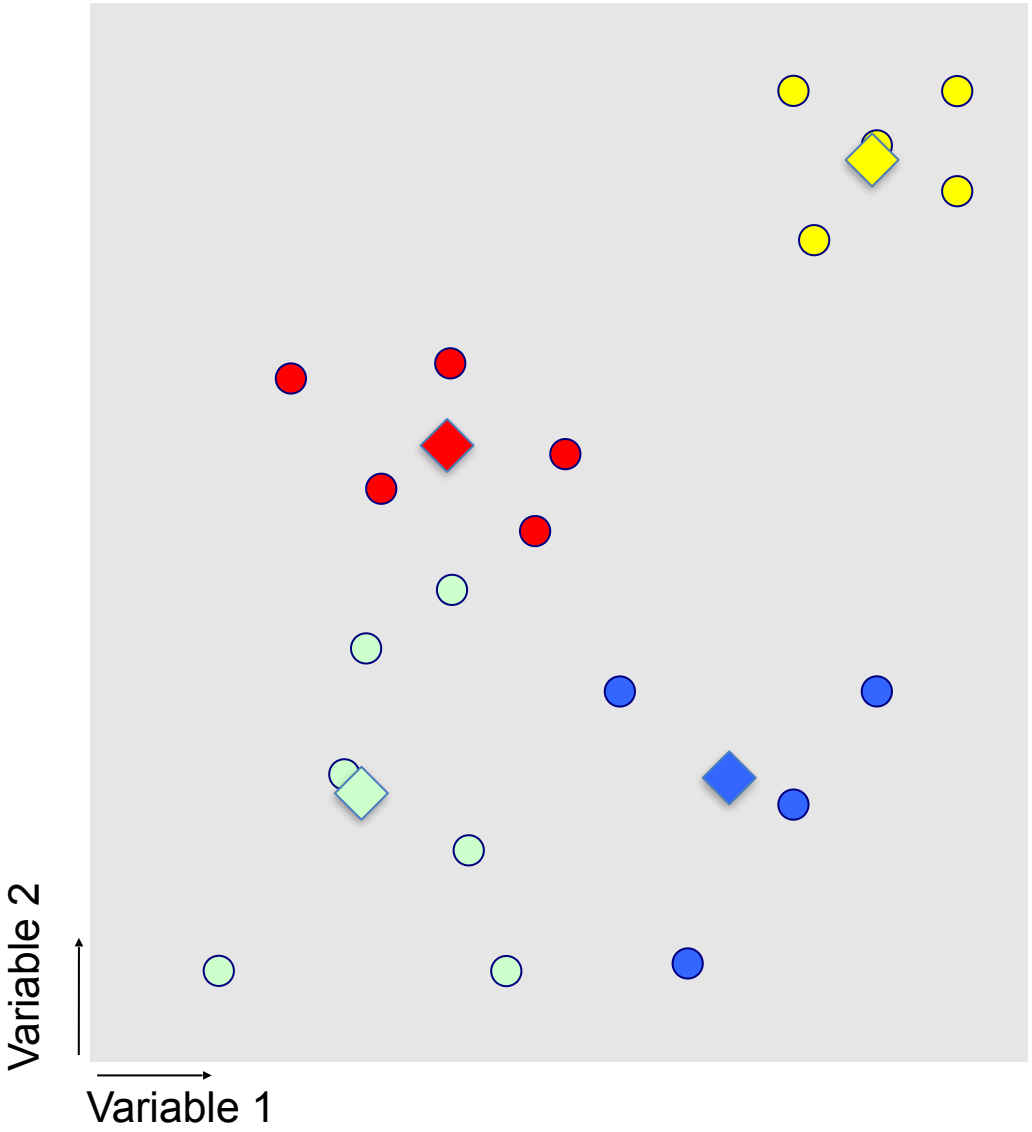
Algorithme des K-means (classification - non supervisé)



Pour chaque observation, on cherche la graine la plus proche.

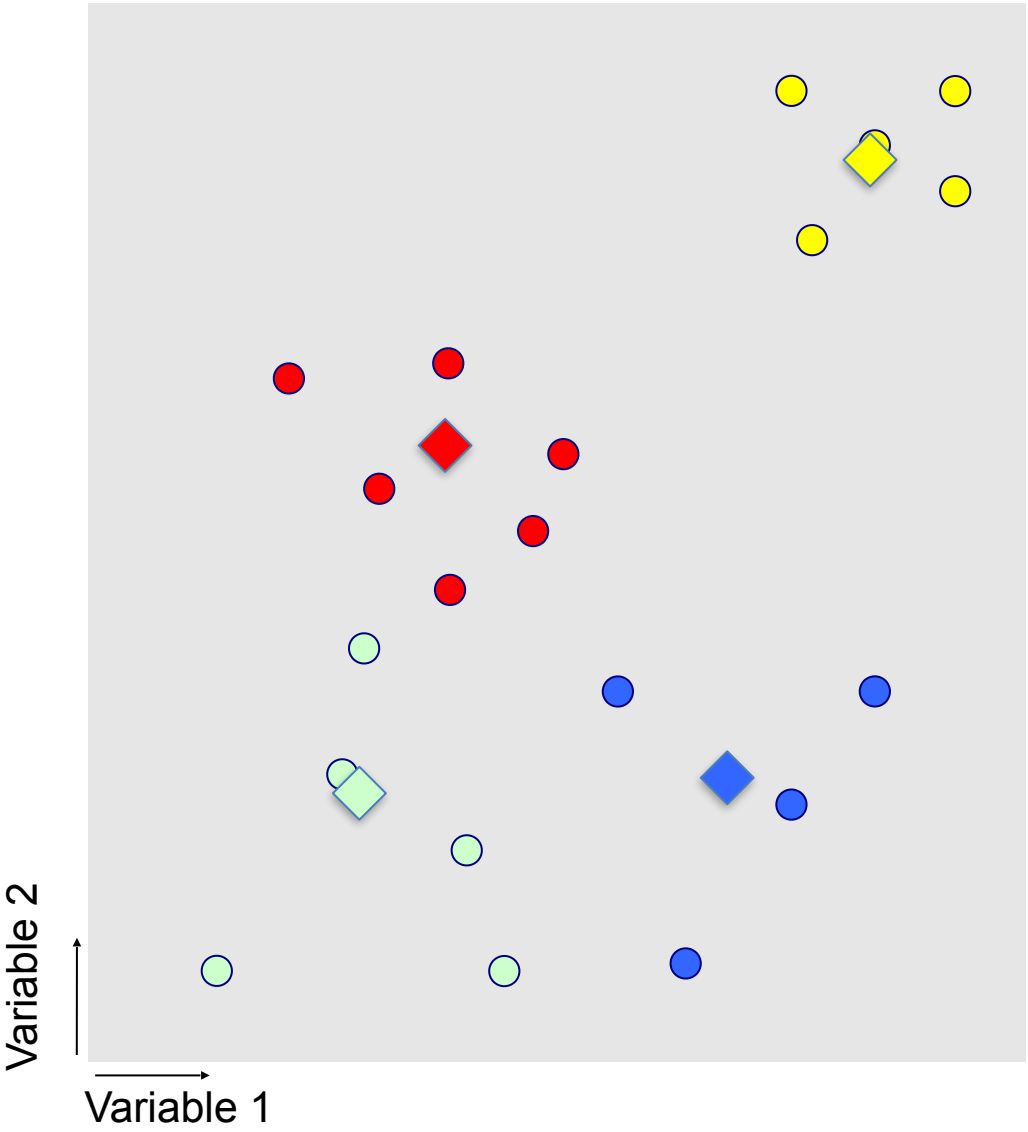
Remarque : Des distances Euclidiennes sont utilisées

Algorithme des K-means (classification - non supervisé)



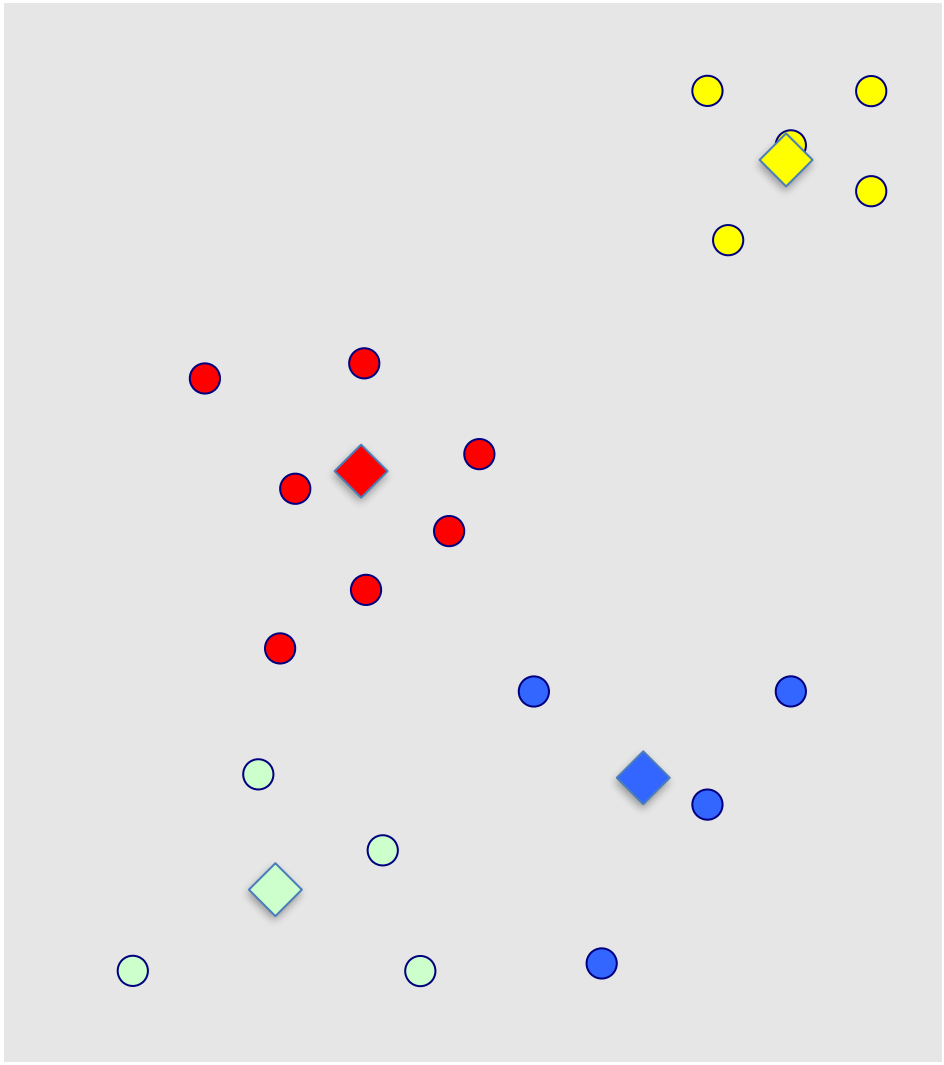
On centre les graines…

Algorithme des K-means (classification - non supervisé)



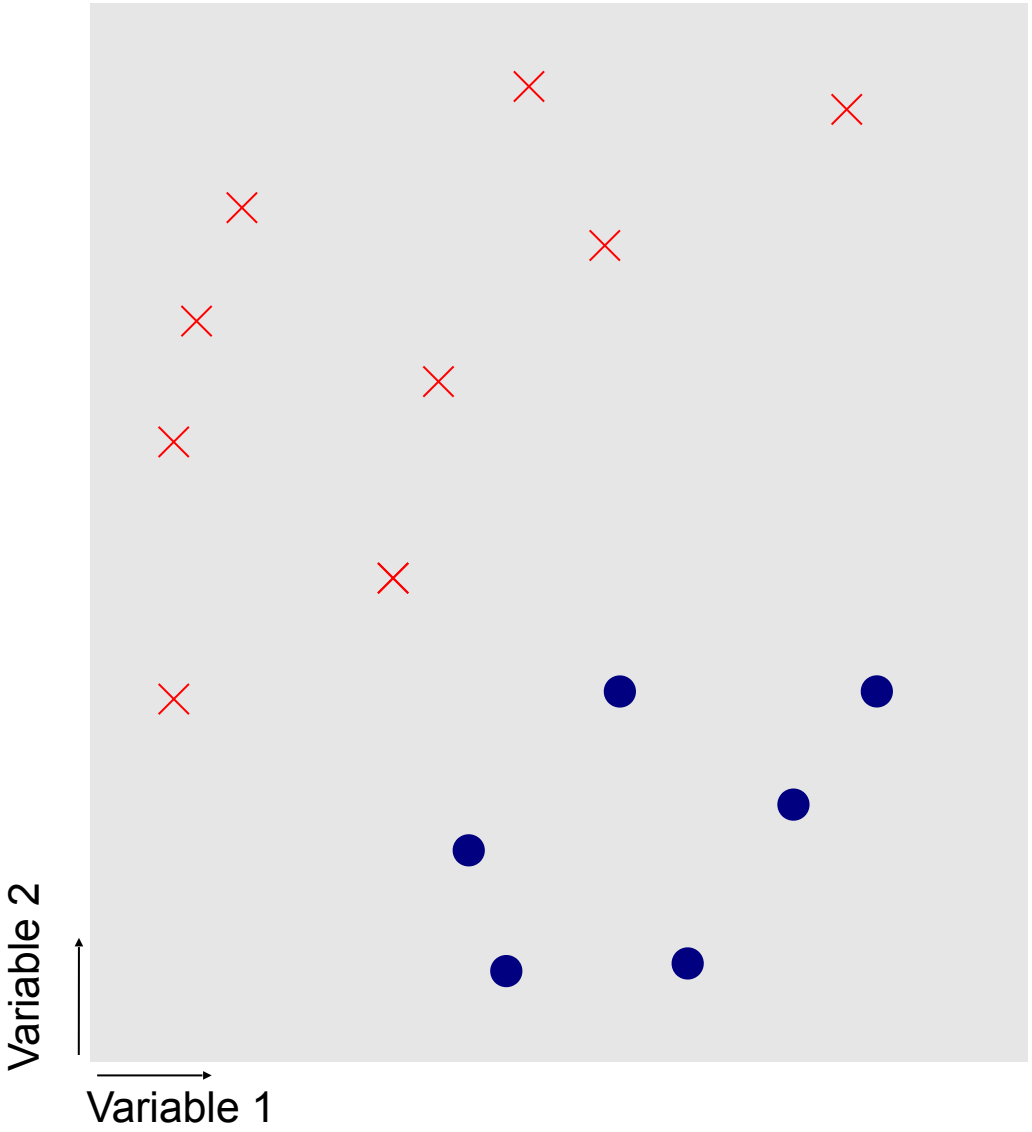
… pour chaque observation, on cherche à nouveau la graine la plus proche …

Algorithme des K-means (classification - non supervisé)



… et on recommence jusqu'a convergence.

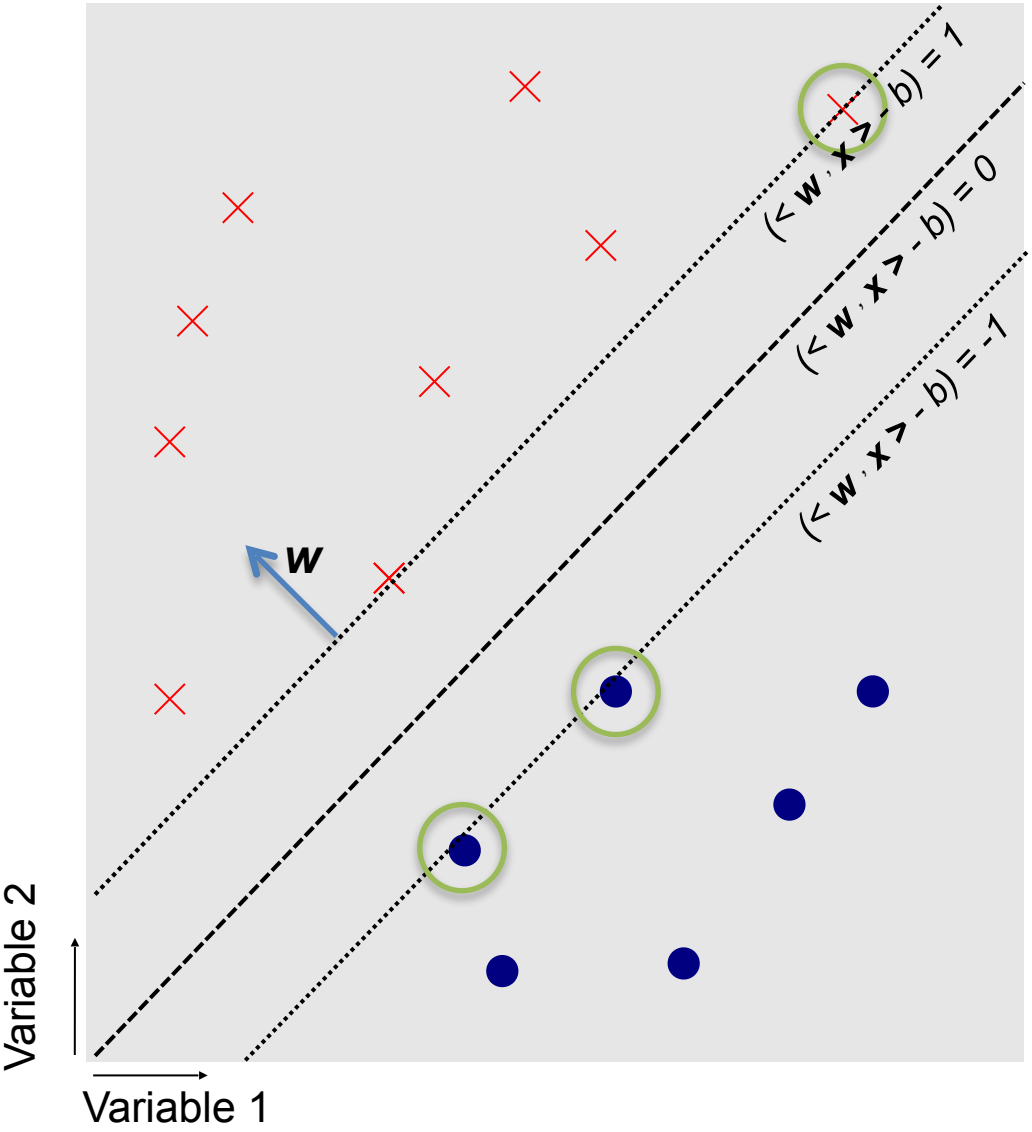
Les Support Vector Machine (classification - supervisé) — Principe



$\boldsymbol{x_1}$, $\boldsymbol{x_2}$, … , $\boldsymbol{x_N}$ les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × ➔ 1 et ● ➔ -1)

Les Support Vector Machine (classification - supervisé) — Principe



$\boldsymbol{x_1}$, $\boldsymbol{x_2}$, … , $\boldsymbol{x_N}$ les observations
(ici : x_i est la coordonnée du point en 2D)

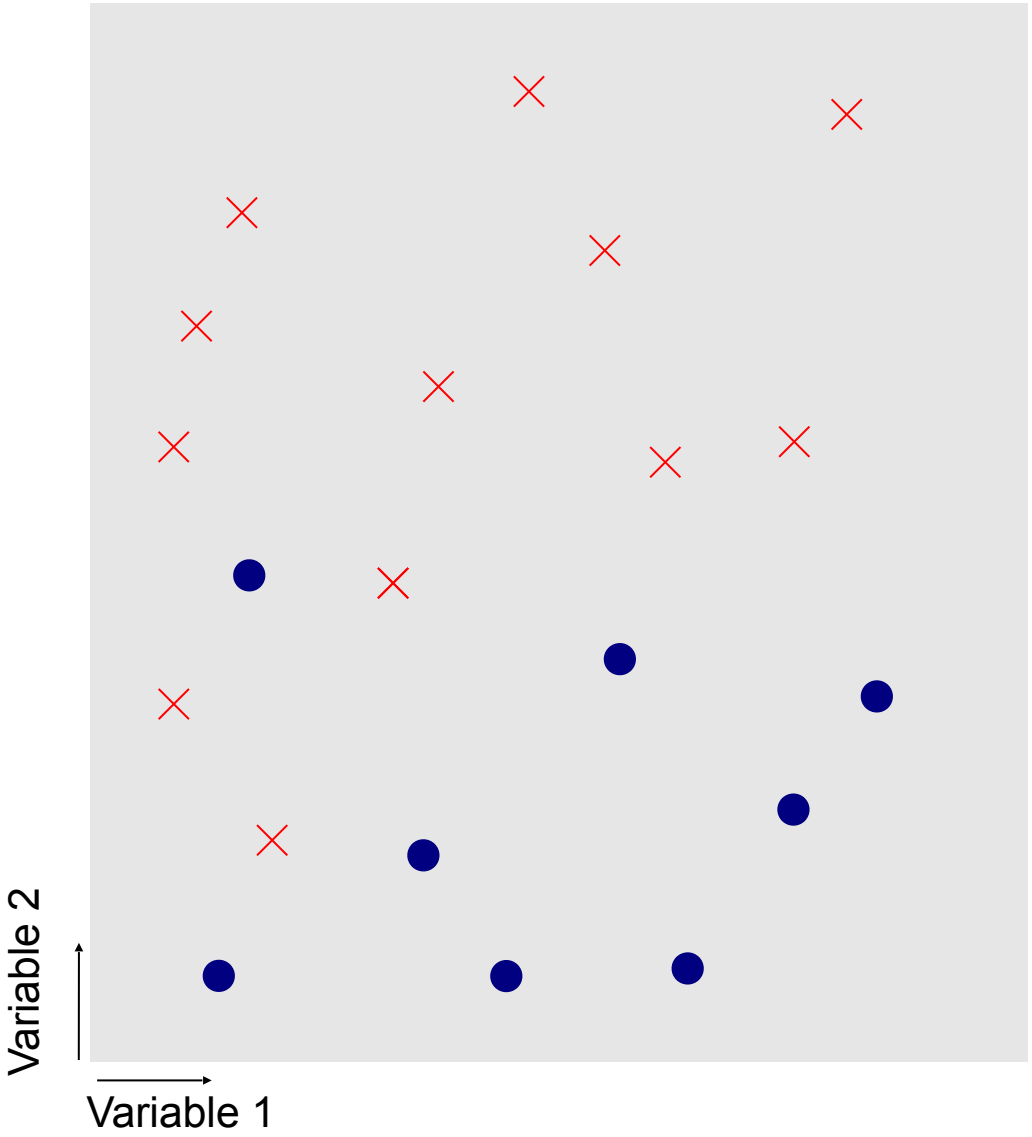
y_1 , y_2 , … , y_N les labels
(ici : × ➔ 1 et ● ➔ -1)

On optimise $\boldsymbol{w}$ et b tel que :

$$y_i (< w, x_i > - b) \geq 1 \text{ pour tout } 1 \leq i \leq n$$

$$< w, x_i > = \sum_{j=1}^{p} w^j x_i^j$$

Les Support Vector Machine (classification - supervisé) — Formulation simplifiée

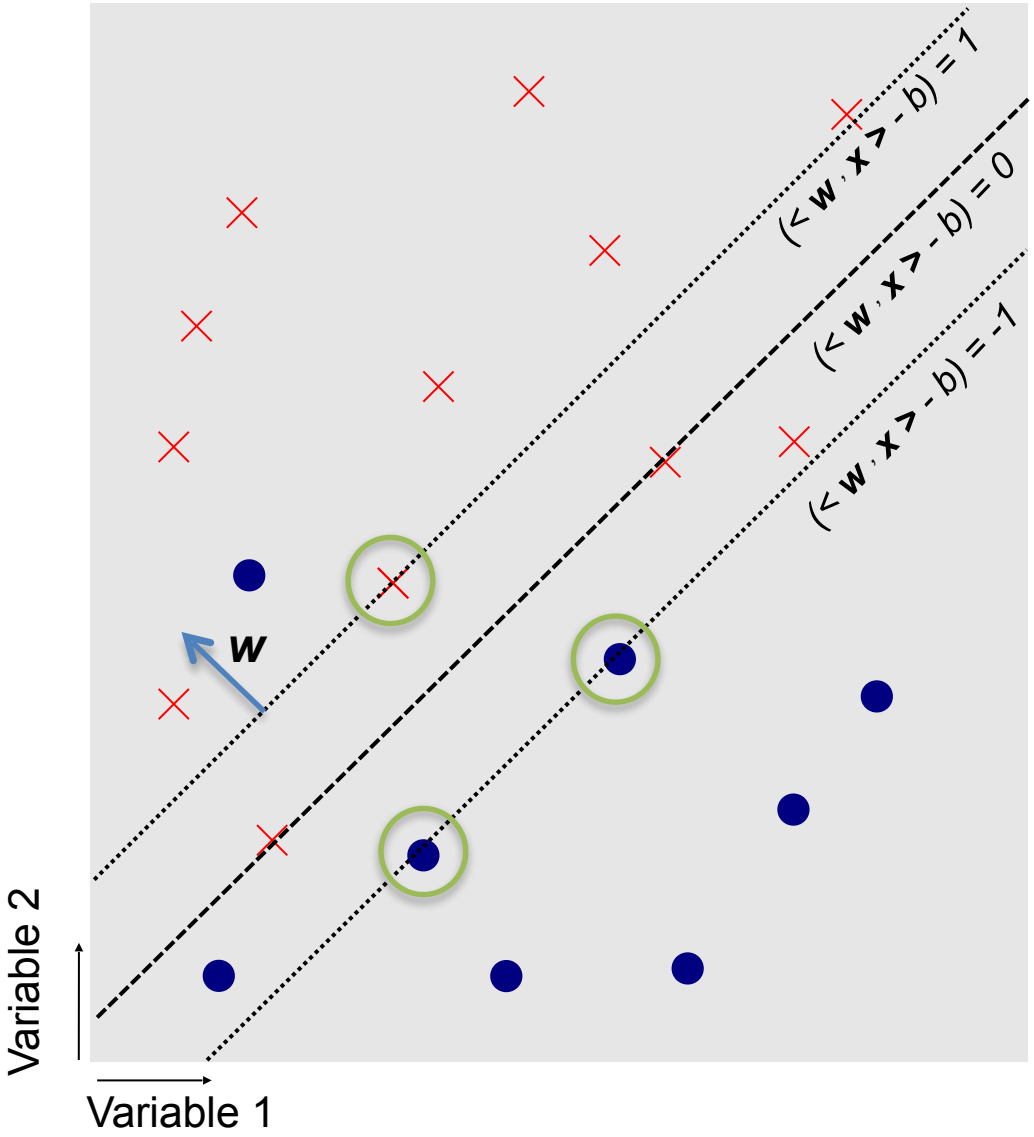


$\boldsymbol{x_1}$, $\boldsymbol{x_2}$, … , $\boldsymbol{x_N}$ les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × ➔ 1 et ● ➔ -1)

Que faire si il est impossible de séparer tous les points?

Les Support Vector Machine (classification - supervisé) — Formulation simplifiée



x_1 , x_2 , … , x_N les observations
(ici : x_i est la coordonnée du point en 2D)

y_1 , y_2 , … , y_N les labels
(ici : × ➔ 1 et ● ➔ -1)

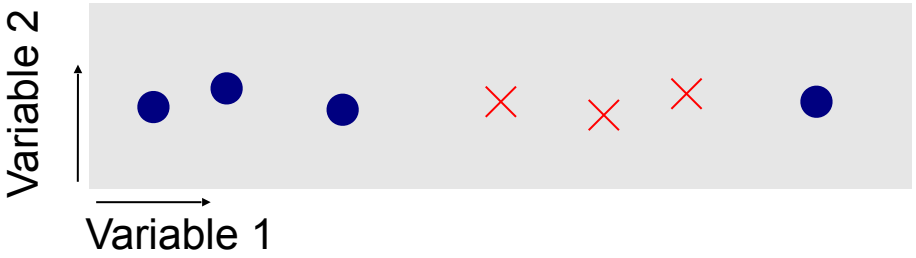
Que faire si il est impossible de séparer tous les points?

On cherche $\boldsymbol{w}$ et b qui minimisent :

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(< w, x_i > - b)\right)\right] + \lambda\,||w||_2^2$$

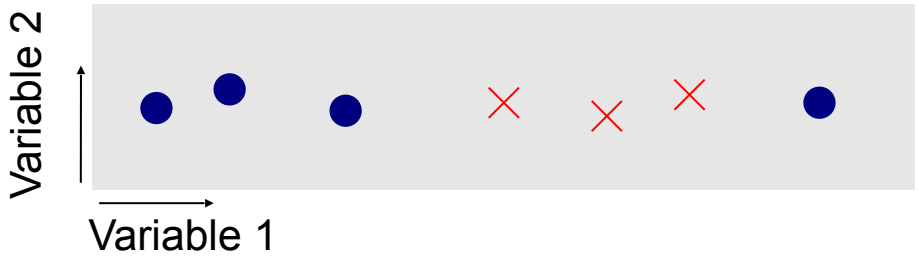
> 0 si $\boldsymbol{x_i}$ est mal classé

Les Support Vector Machine (classification - supervisé) — Méthodes à noyaux



Que faire dans ce cas là ?

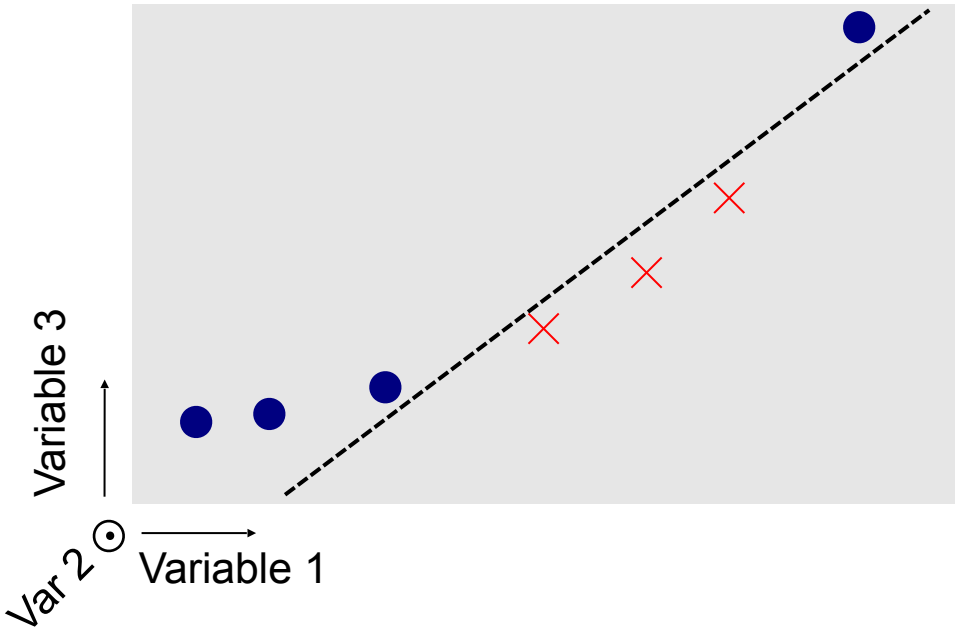
Les Support Vector Machine (classification - supervisé) — Méthodes à noyaux



Que faire dans ce cas là ?

On note $x_i = (x_i^1, x_i^2)$ une observation

On va séparer les $\Phi(x_i) = \left(x_i^1, x_i^2, (x_i^1)^2\right)$ plutôt que les x_i



En pratique le passage à des dimensions supérieures sera traité à l'aide du « Kernel trick »

Réseaux de neurones (classification/régression - supervisée… variantes non supervisées)

Image de chien ou chat

Réseau de neurones

0.99	(chien vs chat)
0.03	(gentil vs agressif)

Réseaux de neurones (classification/régression - supervisée… variantes non supervisées)

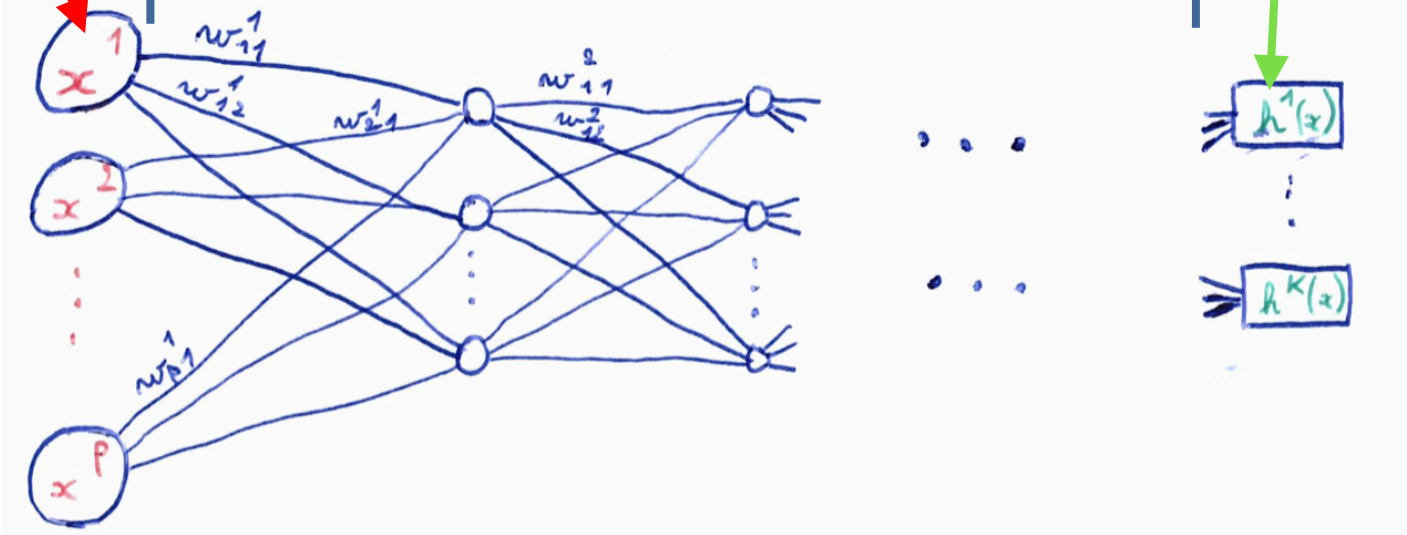
Image de chien ou chat

Réseau de neurones

(perceptron multi-couches)

0.99 (chien vs chat)

0.03 (gentil vs agressif)



Réseaux de neurones (classification/régression - supervisée… variantes non supervisées)

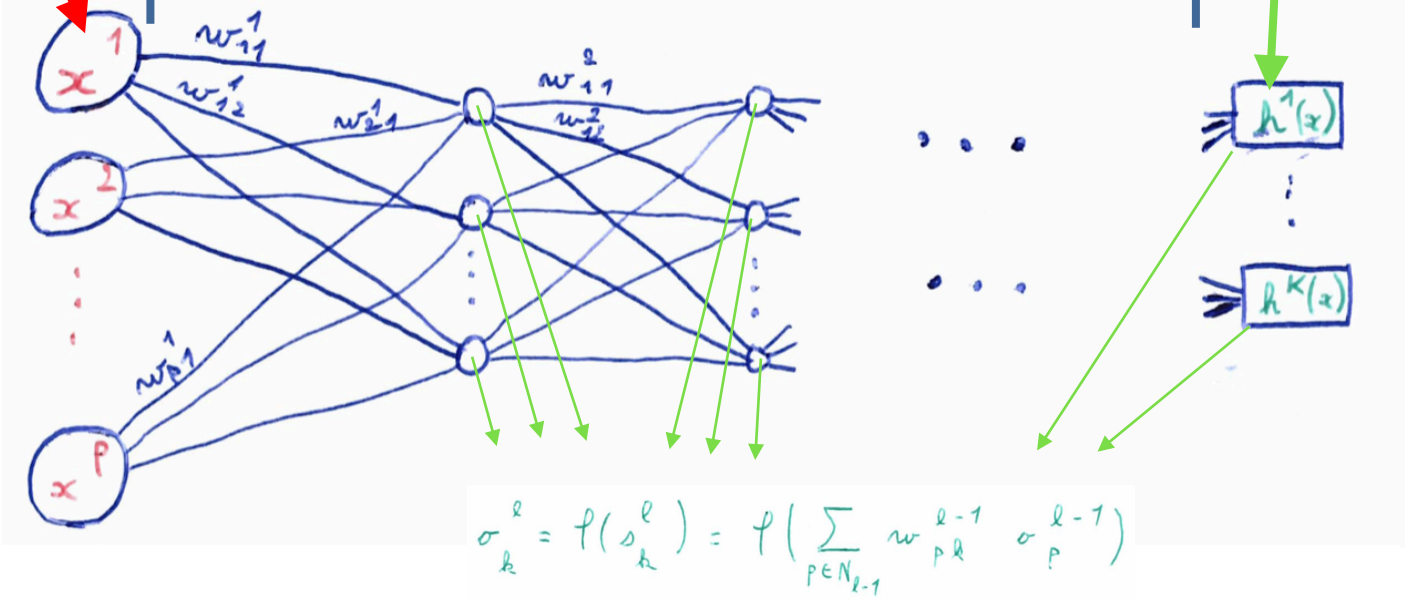
Image de chien ou chat

Réseau de neurones

(perceptron multi-couches)

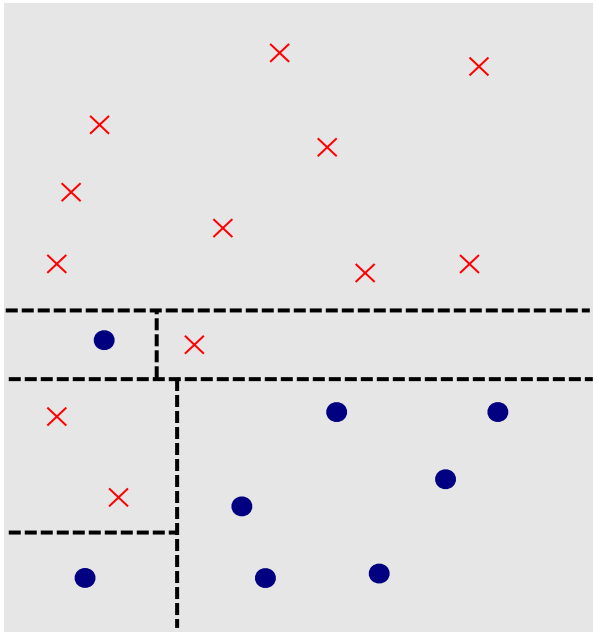
0.99 (chien vs chat)

0.03 (gentil vs agressif)

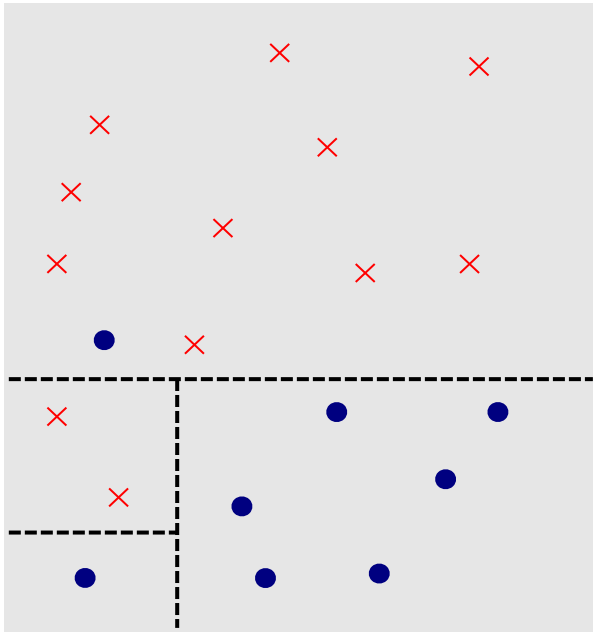


Sur-apprentissage et validation croisée

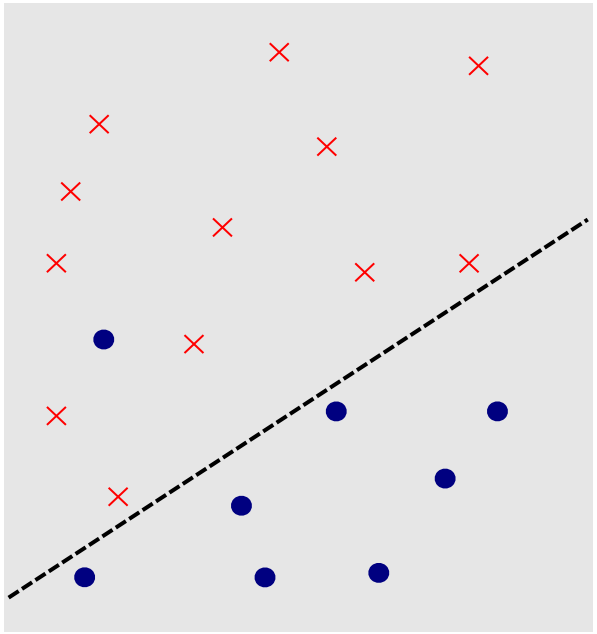
Sur-apprentissage



Arbre de décision complet

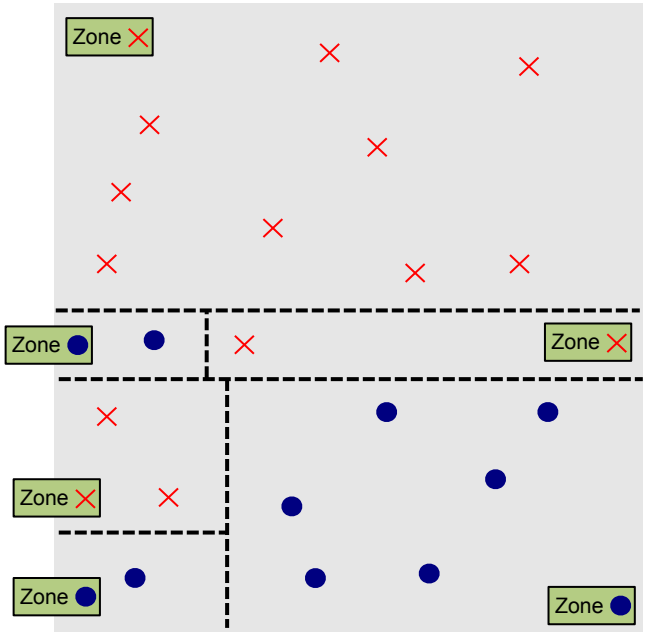


Arbre de décision incomplet

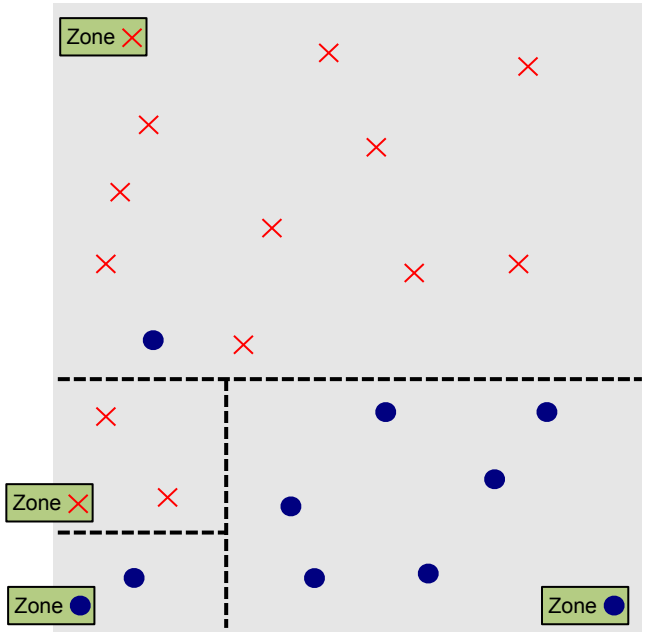


SVM linéaire

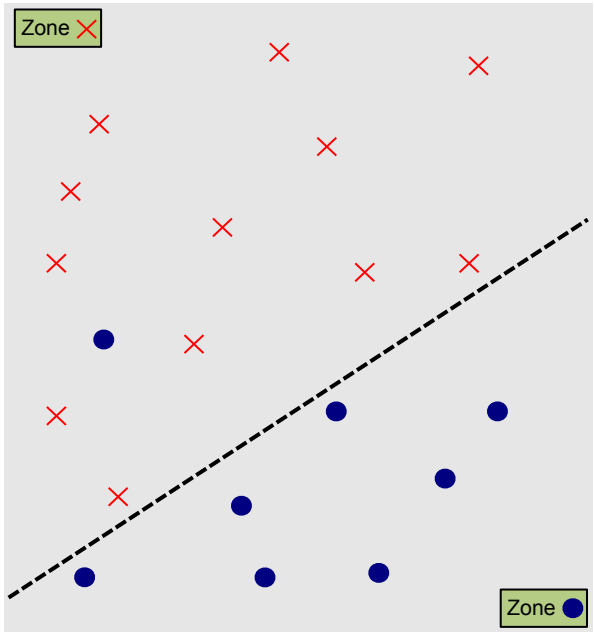
Sur-apprentissage



Arbre de décision complet



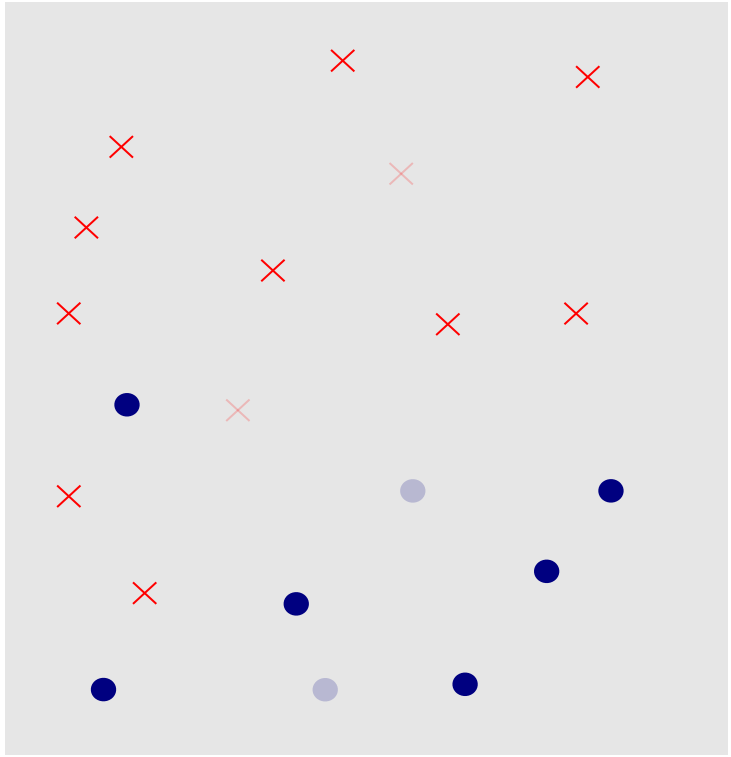
Arbre de décision incomplet



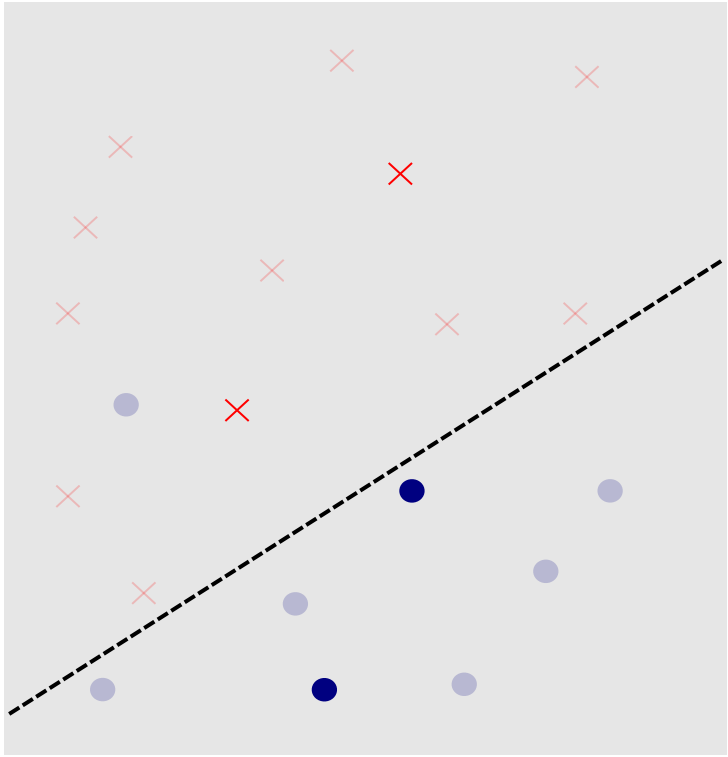
SVM linéaire

A quelle stratégie feriez-vous le plus confiance pour prédire le label d'une nouvelle observation ?

Séparation des données d'apprentissage et données test

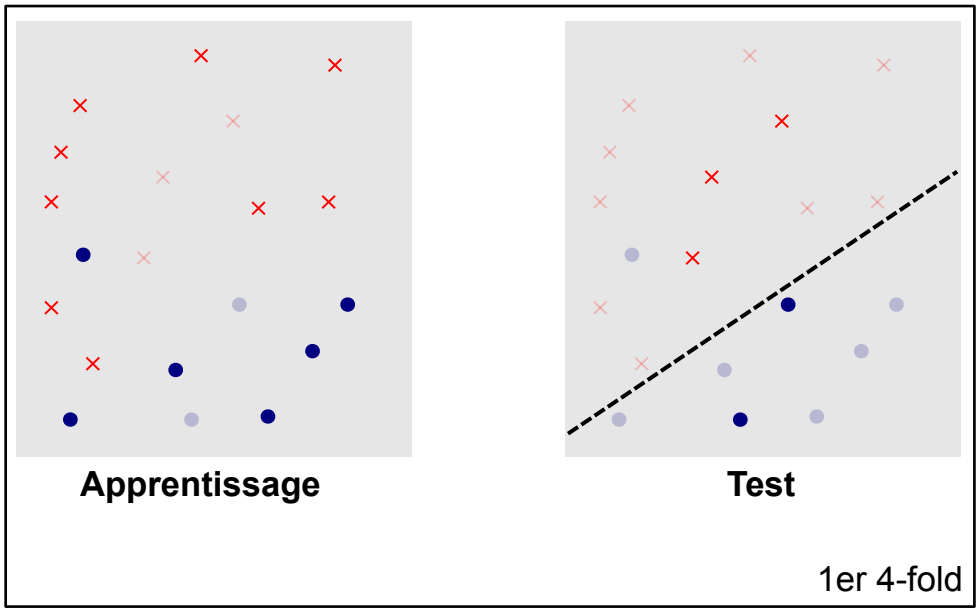


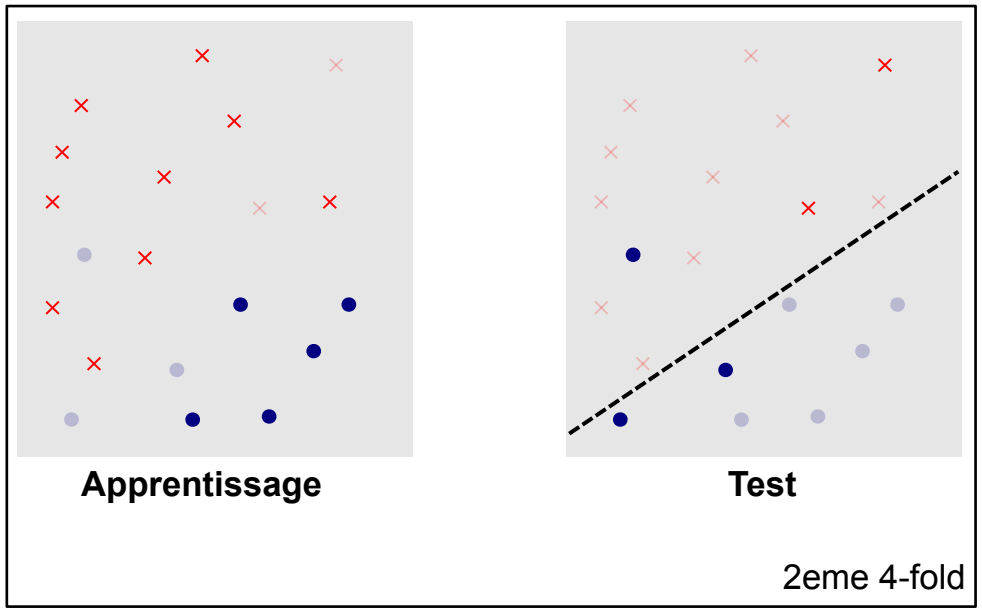
Apprentissage sur toutes les observations sauf 4

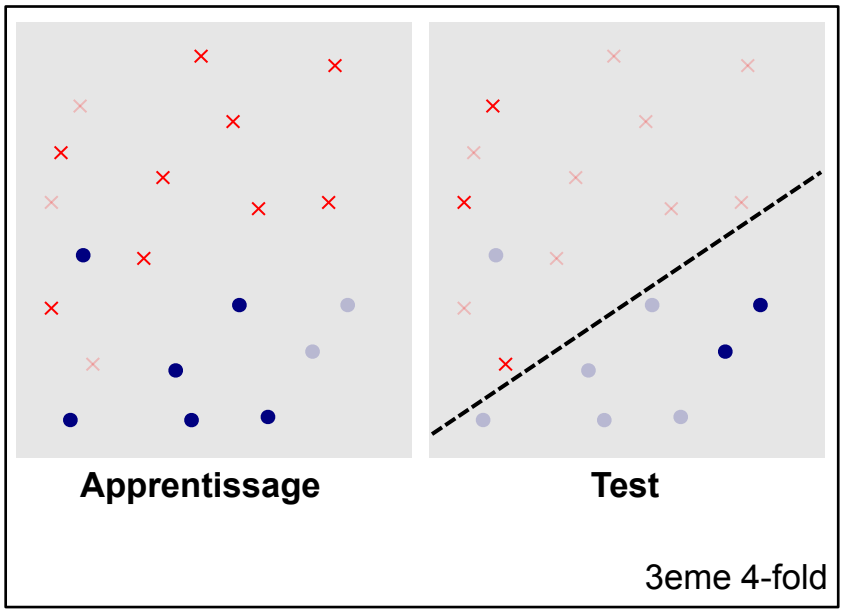


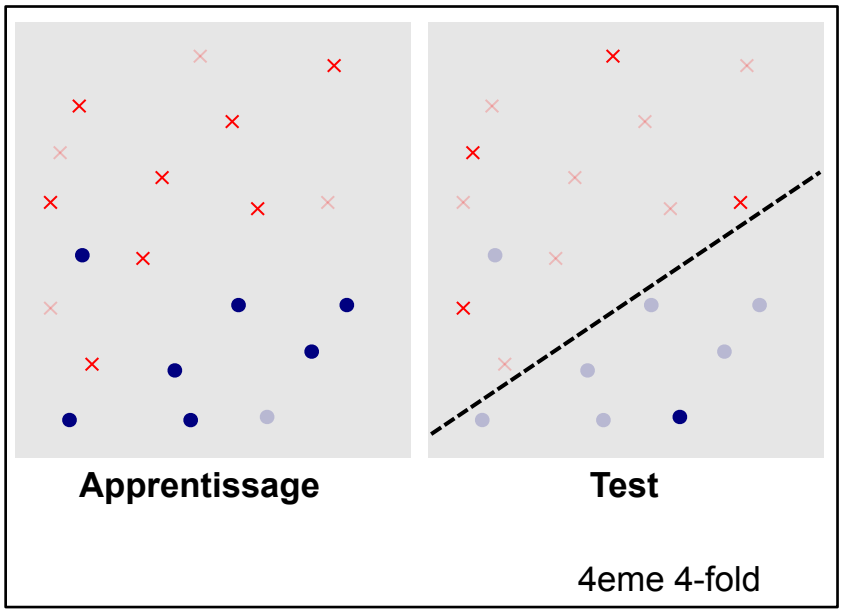
Test sur les 4 données enlevées

K-folds

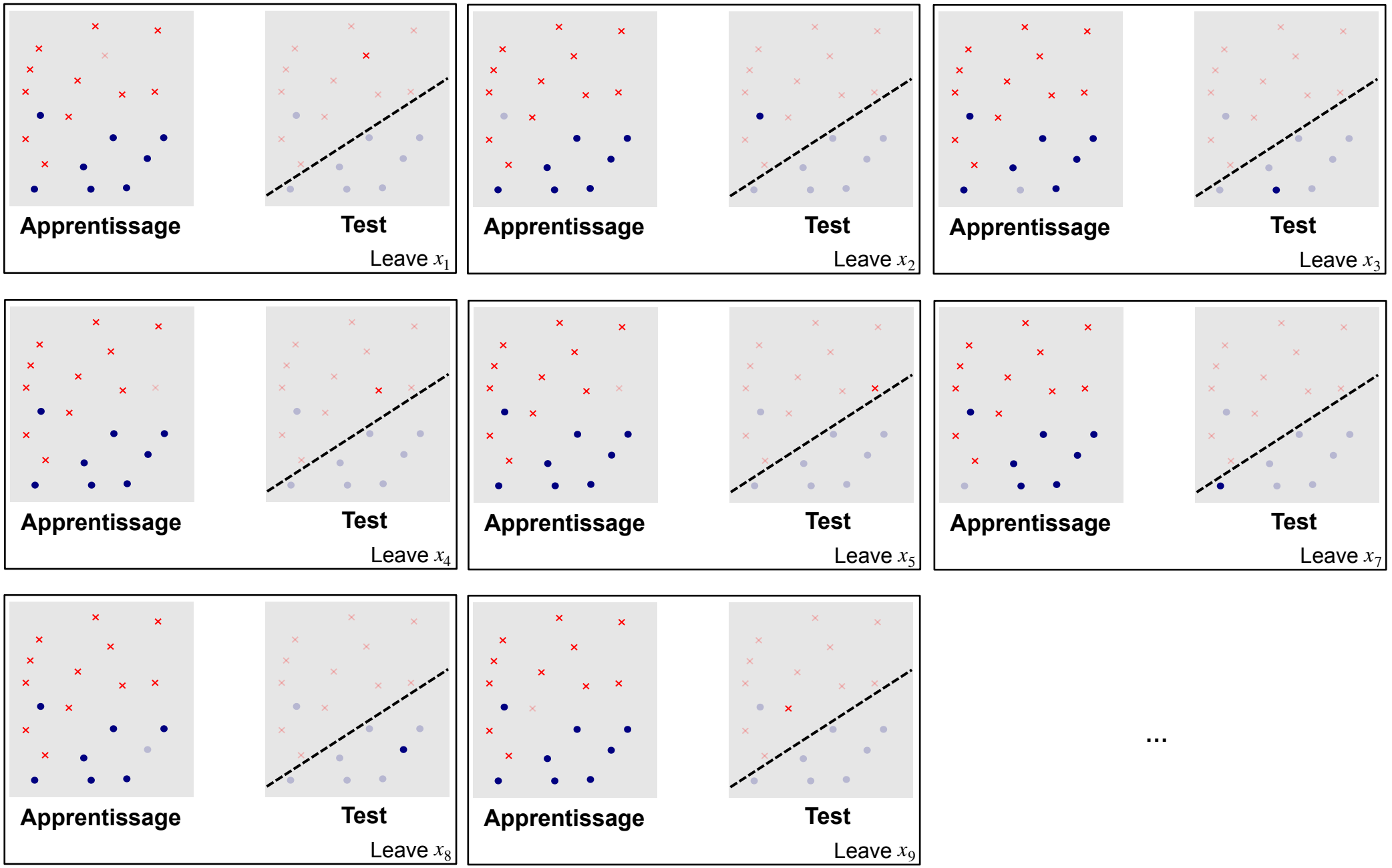

Apprentissage
Test
1er 4-fold


Apprentissage
Test
2eme 4-fold


Apprentissage
Test
3eme 4-fold


Apprentissage
Test
4eme 4-fold

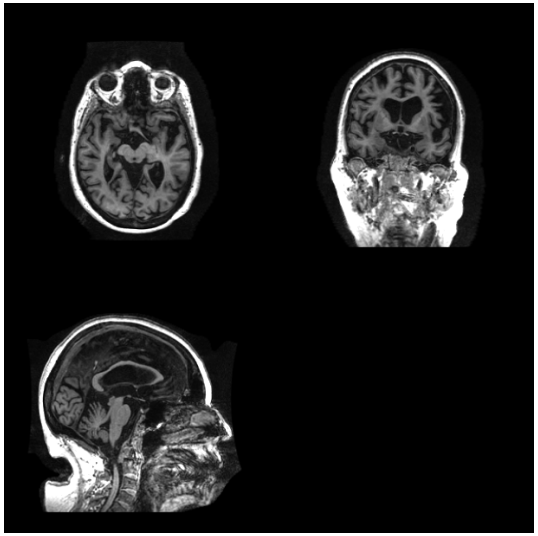
Leave-1-out


Apprentissage
Test
Leave x_1
Apprentissage
Test
Leave x_2
Apprentissage
Test
Leave x_3
Apprentissage
Test
Leave x_4
Apprentissage
Test
Leave x_5
Apprentissage
Test
Leave x_7
Apprentissage
Test
Leave x_8
Apprentissage
Test
Leave x_9
...

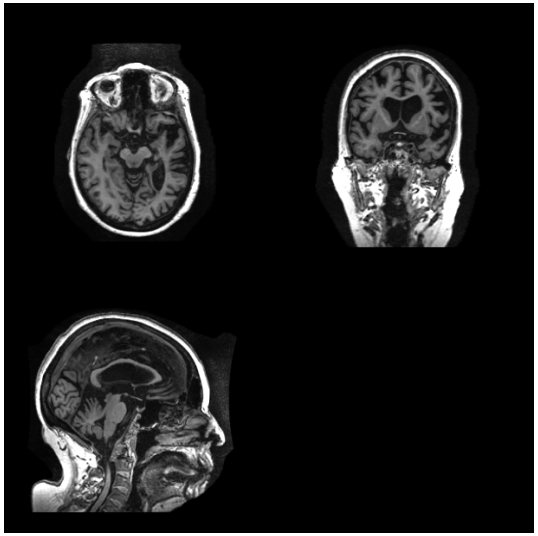
Grande dimension… régularisation et sélection de modèle

Context du projet :

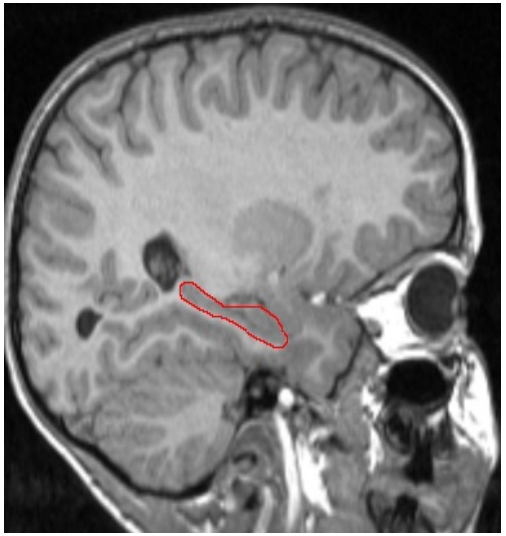
- Observations = IRM du cerveau à différent temps d'acquisition (ADNI*)
- Labels = Etat du patient (MCI/AD)
- Prédiction maladie d'Alheimer en fonction de l'évolution morphométrique de l'hippocampe ?



[Baseline]



[Baseline + 12 mois]



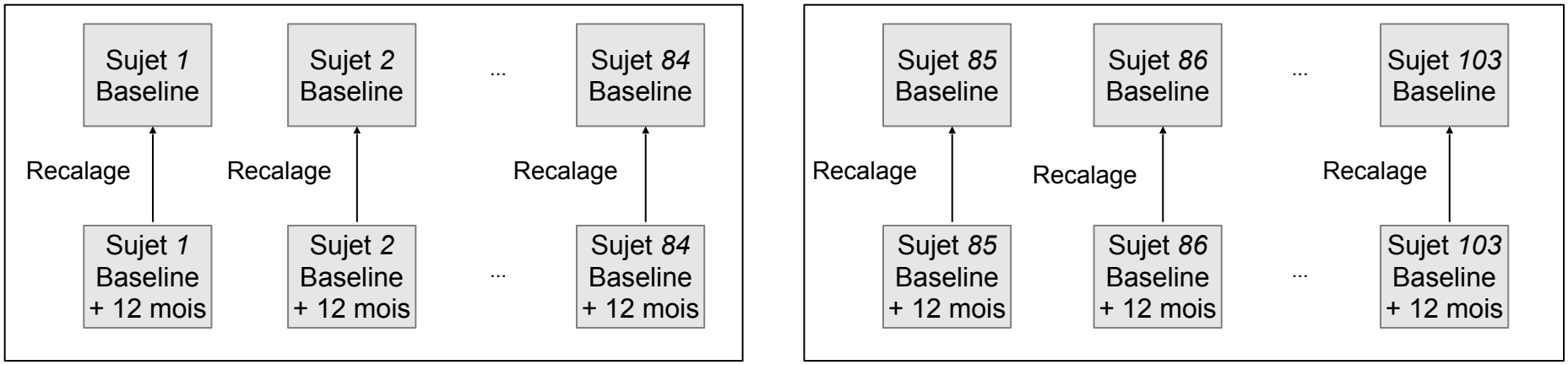
Hippocampe

Echantillon d'apprentissage :

- *[Baseline]* : 103 patients sont MCI
- *[Baseline + 12 mois]* : 84 patients sont MCI / 19 patients sont AD

* http://adni.bmap.ucla.edu/

Pré-traitement des données d'apprentissage :



Groupe des MCI

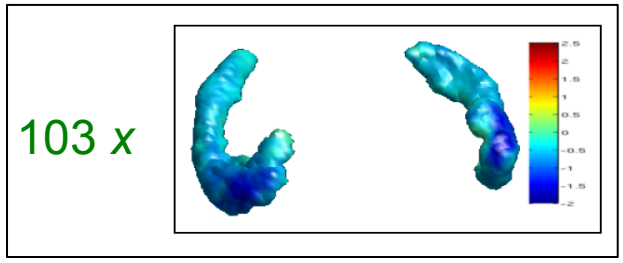
Groupe des AD

Suivi des déformations et transport des marqueurs d'évolution sur une forme *moyenne* [Vialard et al. IJCV, 2012]

Données d'apprentissage :

Pour chacun des n = 103 sujets :

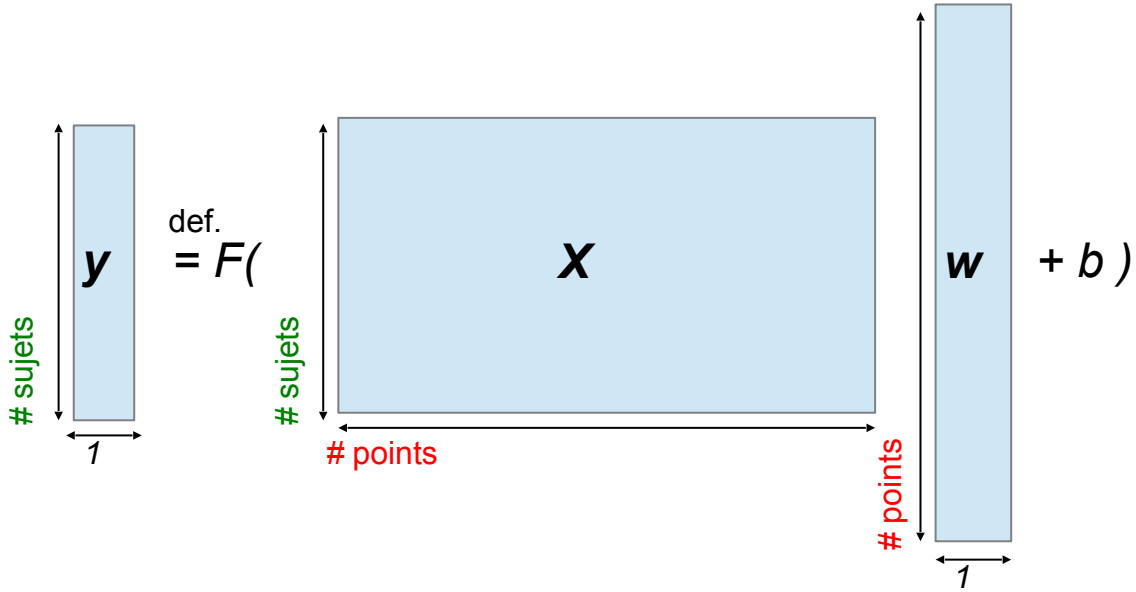
- $\boldsymbol{x_i}$: Observation de l'évolution de la forme sur environ p = 20000 points
- y_i : Etat AD ou MCI



Questions : Classification possible ? Points les plus discriminants ???

Modèle prédictif de régression logistique qui défini la probabilité des y_i en fonctions des $\boldsymbol{x_i}$:

$$p\left(y_i \mid \mathbf{x}_i, \mathbf{w}, b\right) \overset{\text{def.}}{=} \frac{1}{1+\exp\left(-y_i \ \left(\mathbf{x}_i^T \mathbf{w}+b\right)\right)}$$



Où :

$\boldsymbol{X} \in \boldsymbol{R^{n*p}}$: matrice des $n = 103$ observations de dimension $p = 20000$

$\boldsymbol{y} \in \boldsymbol{\{\mp 1\}^n}$: Etat ($AD = -1$ / $MCI = 1$)

$\boldsymbol{(w, b)} \in \boldsymbol{R^p * R}$: parametres à estimer

Optimisation de la log-vraisemblance :

Régularisation obligatoire car p>n

$$\text{Find } (\hat{\mathbf{w}}, \hat{b}) \text{ in } \underset{\mathbf{w},b}{\operatorname{argmin}} \ \mathcal{L}(\mathbf{w}, b) + \lambda J(\mathbf{w}) \quad \text{où} \quad \mathcal{L}(\mathbf{w}, b) \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-y_i(\mathbf{x}_i^T\mathbf{w}+b)\right)\right)$$

Modèle prédictif de régression logistique qui défini la probabilité des y_i en fonctions des $\mathbf{x}_i$:

$$p\left(y_i \mid \mathbf{x}_i, \mathbf{w}, b\right) \stackrel{\text{def.}}{=} \frac{1}{1+\exp\left(-y_i \ \left(\mathbf{x}_i^T \mathbf{w}+b\right)\right)}$$

$2x = 3$	*n = 1* et *p = 1* — *OK*
$2x_1 + 3x_2 = 3$	*n = 1* et *p = 2* — *KO*
$2x_1 + 3x_2 = 3$ $3x_1 + 1x_2 = 1$	*n = 2* et *p = 2* — *OK*
$2x_1 + 3x_2 + 1x_3 - x_4 = 1$ $5x_1 - x_2 + 2x_3 + x_4 = 1$	*n = 2* et *p = 4* — *KO*

Où :

$\boldsymbol{X} \in \boldsymbol{R}^{n*p}$: matric

$\boldsymbol{y} \in \{\mp 1\}^n$: Etat

$(\boldsymbol{w}, \boldsymbol{b}) \in \boldsymbol{R}^p * \boldsymbol{R}$: parametres à estimer

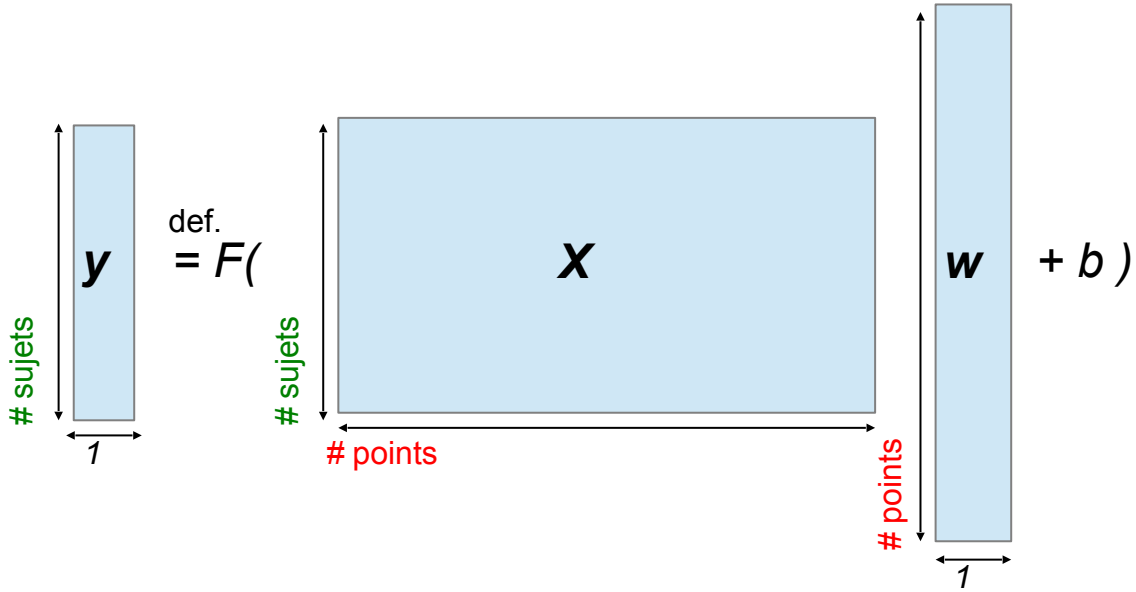
Optimisation de la log-vraisemblance :

Régularisation obligatoire car p>n

$$\text{Find } (\hat{\mathbf{w}}, \hat{b}) \text{ in } \underset{\mathbf{w},b}{\operatorname{argmin}} \ \mathcal{L}(\mathbf{w}, b) + \lambda J(\mathbf{w}) \quad \text{où} \quad \mathcal{L}(\mathbf{w}, b) \stackrel{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-y_i(\mathbf{x}_i^T \mathbf{w}+b)\right)\right)$$

Modèle prédictif de régression logistique qui défini la probabilité des y_i en fonctions des $\boldsymbol{x_i}$:

$$p\left(y_i \mid \mathbf{x}_i, \mathbf{w}, b\right) \overset{\text{def.}}{=} \frac{1}{1+\exp\left(-y_i \ \left(\mathbf{x}_i^T \mathbf{w}+b\right)\right)}$$



Où :

$\boldsymbol{X} \in \boldsymbol{R^{n*p}}$: matrice des $n = 103$ observations de dimension $p = 20000$

$\boldsymbol{y} \in \boldsymbol{\{\mp 1\}^n}$: Etat ($AD = -1$ / $MCI = 1$)

$\boldsymbol{(w, b)} \in \boldsymbol{R^p * R}$: parametres à estimer

Optimisation de la log-vraisemblance :

Régularisation obligatoire car p>n

$$\text{Find } (\hat{\mathbf{w}}, \hat{b}) \text{ in } \underset{\mathbf{w},b}{\operatorname{argmin}} \ \mathcal{L}(\mathbf{w}, b) + \lambda J(\mathbf{w}) \quad \text{où} \quad \mathcal{L}(\mathbf{w}, b) \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-y_i(\mathbf{x}_i^T\mathbf{w}+b)\right)\right)$$

$$\text{Find } (\hat{\mathbf{w}}, \hat{b}) \text{ in } \underset{\mathbf{w},b}{\operatorname{argmin}} \ \mathcal{L}(\mathbf{w}, b) + \lambda J(\mathbf{w}) \qquad \text{où :} \qquad \mathcal{L}(\mathbf{w}, b) \overset{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i(\mathbf{x}_i^T \mathbf{w} + b)\right)\right)$$

Exploration de différents modèles de régularisation :

(1) Ridge : $J(\mathbf{w}) = ||\mathbf{w}||_2$

(2) LASSO : $J(\mathbf{w}) = ||\mathbf{w}||_1$

(3) Elastic net : $J(\mathbf{w}) = \lambda_1 ||\mathbf{w}||_1 + \lambda_2 ||\mathbf{w}||_2$

(4) Sobolev semi-norm: $J(\mathbf{w}) = \sum_{\omega \in \Omega} |\nabla_\Omega \mathbf{w}(\omega)|$

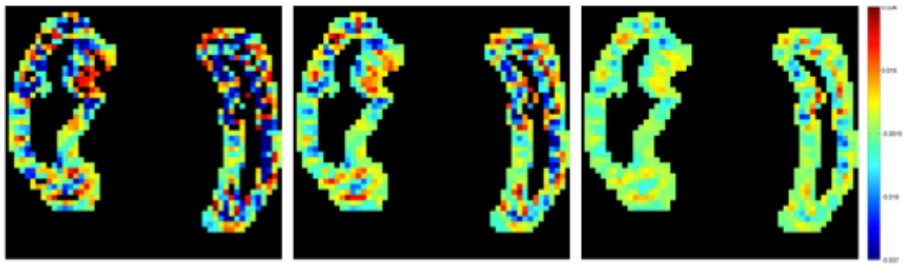
(5) Total Variation : $J(\mathbf{w}) = \left(\sum_{\omega \in \Omega} |\nabla_\Omega \mathbf{w}(\omega)|^2\right)^{1/2}$

(6) Fused LASSO : $J(\mathbf{w}) = \lambda_1 ||\mathbf{w}||_1 + \lambda_2 \sum_{\omega \in \Omega} |\nabla_\Omega \mathbf{w}(\omega)|$

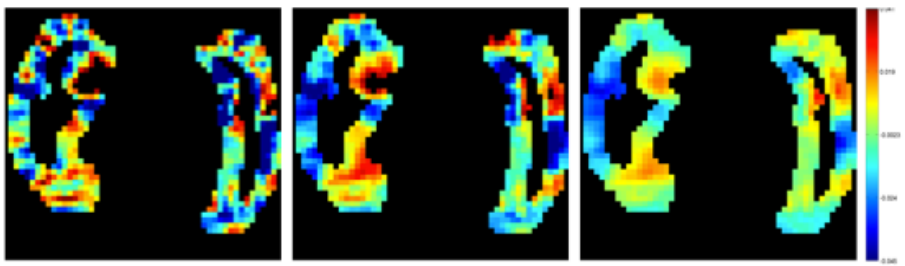
Minimisation de la log-vraisemblance en fonction de ***w***

[Lewis & Overton. Math. Programming 2012]

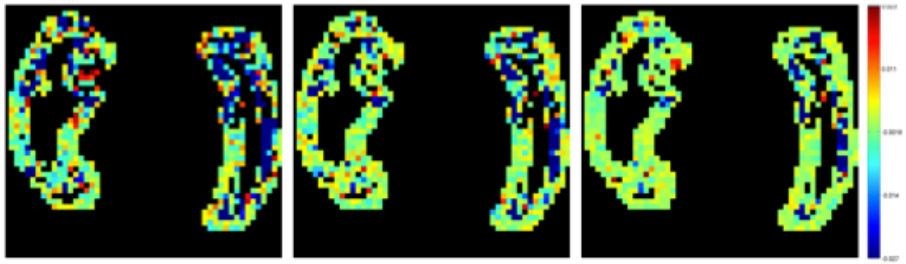
5.c) Grande dimension — Effet de la régularisation



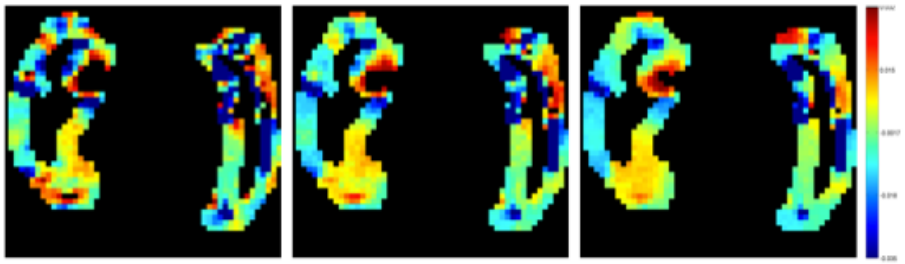
(1) Ridge



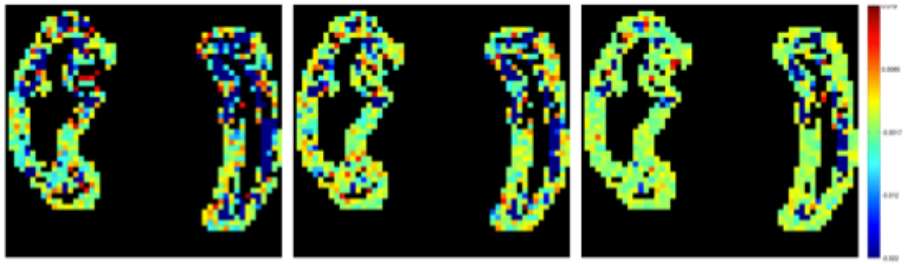
(4) Sobolev semi-norm



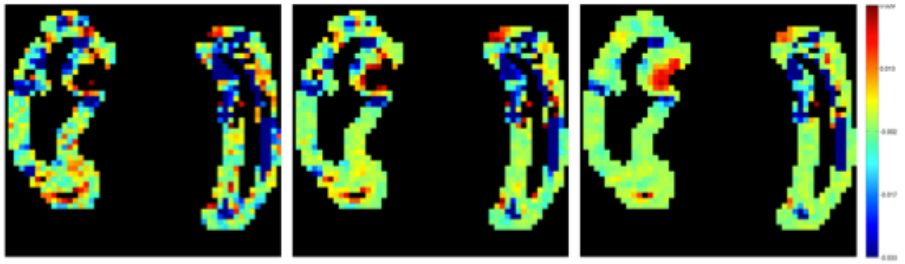
(2) LASSO



(5) Total Variation



(3) Elastic net



(6) Fused LASSO

Représentation de $\boldsymbol{w}$ pour trois λ sur un plan de l'hippocampe :

- Bleu et rouge : forte influence locale
- Vert : peu ou pas d'influence locale

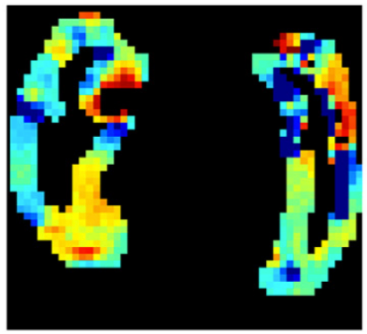
Résultats obtenus avec une méthode de cross validation (ici leave-10%-out) :

- Spec+Sens = 2 ⟶ bonne prédiction dans 100% des cas
- Spec+Sens = 1 ⟶ pile ou face aurait le même pouvoir prédictif

Regularization		λ **range**	$\hat{\lambda}$ **(optimal** λ**)**	**Spec+ Sens**
None		0	0	1.00
Standard	LASSO	$[10^{-9},10^{0}]$	0.01	1.04
	Ridge	$[10^{-9},10^{0}]$	0.001	1.06
	Elastic Net	$[10^{-9},10^{0}]^2$	$\begin{cases}\hat{\lambda}_1 = 0.01\\ \hat{\lambda}_2 = 1\end{cases}$	1.13
Spatial	Sobolev	$[10^{-9},10^{7}]$	10^4	1.17
	Total Variation	$[10^{-9},10^{0}]$	0.01	**1.31**
	Fused LASSO	$[10^{-9},10^{0}]^2$	$\begin{cases}\hat{\lambda}_1 = 0.01\\ \hat{\lambda}_2 = 10^{-4}\end{cases}$	**1.32**

Meilleurs résultats avec une régularisation en pertinente avec les données :

- Tient compte de la distribution spatial
- Permet quelques transitions franches



[Fiot J.B. et al., NeuroImage: Clinical, 2012]

Réduction de dimension par Analyse en Composantes Principales (ACP)

Records nationaux (en secondes) de p = 9 épreuves d'athlétisme pour n = 26 pays

	100m	200m	400m	800m	1500m	5000m	10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	1638.14	3609	7633
Canada	9.84	20.17	44.44	103.68	211.71	793.96	1656.01	3650	7809
Chine	10.17	20.54	45.25	106.44	216.49	805.14	1670.00	3635	7695
Croatie	10.25	20.76	45.64	104.07	213.30	817.76	1704.32	3827	8225
Ethiopie	10.50	21.08	45.89	106.08	211.13	757.35	1577.53	3535	7439
France	9.99	20.16	44.46	103.15	208.98	778.83	1642.78	3658	7596
Allemagne	10.06	20.20	44.33	103.65	211.58	774.70	1641.53	3634	7727
Inde	10.30	20.73	45.48	105.77	218.00	809.70	1682.89	3672	7920
Iran	10.29	21.11	46.37	104.74	218.80	833.40	1762.65	4103	8903
Italie	10.01	19.72	45.19	103.17	212.78	785.59	1636.50	3620	7642
Jamaïque	9.58	19.19	44.49	105.21	219.19	813.10	1712.44	3816	8199
Japon	10.00	20.03	44.78	106.18	217.42	793.20	1655.09	3625	7576
Kenya	10.26	20.43	44.18	102.01	206.34	759.74	1587.85	3513	7467
Lituanie	10.33	20.88	45.73	106.64	220.90	797.90	1651.50	3851	7955
NouvelleZélande	10.11	20.42	46.09	104.30	212.17	790.19	1661.95	3732	7815
Portugal	9.86	20.01	46.11	104.91	210.07	782.86	1632.47	3665	7596
Russie	10.10	20.23	44.60	102.47	212.28	791.99	1673.12	3675	7747
AfriqueduSud	10.06	20.11	44.59	102.69	213.56	794.16	1649.94	3678	7593
Espagne	10.14	20.59	44.96	103.83	208.95	782.54	1634.44	3592	7562
Suède	10.18	20.30	44.56	105.54	216.49	797.59	1675.74	3655	7838
Suisse	10.16	20.41	44.99	102.55	211.75	787.54	1673.16	3686	7643
Ukraine	10.07	20.00	45.11	105.08	210.33	790.78	1679.80	3711	7635
USA	9.69	19.32	43.18	102.60	209.30	776.27	1633.98	3583	7538

Comment faire un classement général entre les pays ???

→ Somme pondérée des scores, puis classement en fonction du rang.

Exemple issu d'un cours de Sébastien Déjean (IR Univ Toulouse 3 / IMT)

Records nationaux (en secondes) de p = 9 épreuves d'athlétisme pour n = 26 pays

	100m	200m	400m	800m	1500m	5000m	10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	1638.14	3609	7633
Canada	9.84	20.17	44.44	103.68	211.71	793.96	1656.01	3650	7809
Chine	10.17	20.54	45.25	106.44	216.49	805.14	1670.00	3635	7695
Croatie	10.25	20.76	45.64	104.07	213.30	817.76	1704.32	3827	8225
Ethiopie	10.50	21.08	45.89	106.08	211.13	757.35	1577.53	3535	7439
France	9.99	20.16	44.46	103.15	208.98	778.83	1642.78	3658	7596
Allemagne	10.06	20.20	44.33	103.65	211.58	774.70	1641.53	3634	7727
Inde	10.30	20.73	45.48	105.77	218.00	809.70	1682.89	3672	7920
Iran	10.29	21.11	46.37	104.74	218.80	833.40	1762.65	4103	8903
Italie	10.01	19.72	45.19	103.17	212.78	785.59	1636.50	3620	7642
Jamaïque	9.58	19.19	44.49	105.21	219.19	813.10	1712.44	3816	8199
Japon	10.00	20.03	44.78	106.18	217.42	793.20	1655.09	3625	7576
Kenya	10.26	20.43	44.18	102.01	206.34	759.74	1587.85	3513	7467
Lituanie	10.33	20.88	45.73	106.64	220.90	797.90	1651.50	3851	7955
NouvelleZélande	10.11	20.42	46.09	104.30	212.17	790.19	1661.95	3732	7815
Portugal	9.86	20.01	46.11	104.91	210.07	782.86	1632.47	3665	7596
Russie	10.10	20.23	44.60	102.47	212.28	791.99	1673.12	3675	7747
AfriqueduSud	10.06	20.11	44.59	102.69	213.56	794.16	1649.94	3678	7593
Espagne	10.14	20.59	44.96	103.83	208.95	782.54	1634.44	3592	7562
Suède	10.18	20.30	44.56	105.54	216.49	797.59	1675.74	3655	7838
Suisse	10.16	20.41	44.99	102.55	211.75	787.54	1673.16	3686	7643
Ukraine	10.07	20.00	45.11	105.08	210.33	790.78	1679.80	3711	7635
USA	9.69	19.32	43.18	102.60	209.30	776.27	1633.98	3583	7538

→ Matrice ***M***

Comment faire un classement général entre les pays ???

→ Somme pondérée des scores, puis classement en fonction du rang.

Somme pondérée des scores = produit [*matrice*] . [*vecteur*] :

→ Vecteur contenant les scores = $\boldsymbol{M} \cdot \boldsymbol{w}$

Records nationaux (en secondes) de p = 9 épreuves d'athlétisme pour n = 26 pays

	100m	200m	400m	800m	1500m	5000m	10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	1638.14	3609	7633
Canada	9.84	20.17	44.44	103.68	211.71	793.96	1656.01	3650	7809
Chine	10.17	20.54	45.25	106.44	216.49	805.14	1670.00	3635	7695
Croatie	10.25	20.76	45.64	104.07	213.30	817.76	1704.32	3827	8225
Ethiopie	10.50	21.08	45.89	106.08	211.13	757.35	1577.53	3535	7439
France	9.99	20.16	44.46	103.15	208.98	778.83	1642.78	3658	7596
Allemagne	10.06	20.20	44.33	103.65	211.58	774.70	1641.53	3634	7727
Inde	10.30	20.73	45.48	105.77	218.00	809.70	1682.89	3672	7920
Iran	10.29	21.11	46.37	104.74	218.80	833.40	1762.65	4103	8903
Italie	10.01	19.72	45.19	103.17	212.78	785.59	1636.50	3620	7642
Jamaïque	9.58	19.19	44.49	105.21	219.19	813.10	1712.44	3816	8199
Japon	10.00	20.03	44.78	106.18	217.42	793.20	1655.09	3625	7576
Kenya	10.26	20.43	44.18	102.01	206.34	759.74	1587.85	3513	7467
Lituanie	10.33	20.88	45.73	106.64	220.90	797.90	1651.50	3851	7955
NouvelleZélande	10.11	20.42	46.09	104.30	212.17	790.19	1661.95	3732	7815
Portugal	9.86	20.01	46.11	104.91	210.07	782.86	1632.47	3665	7596
Russie	10.10	20.23	44.60	102.47	212.28	791.99	1673.12	3675	7747
AfriqueduSud	10.06	20.11	44.59	102.69	213.56	794.16	1649.94	3678	7593
Espagne	10.14	20.59	44.96	103.83	208.95	782.54	1634.44	3592	7562
Suède	10.18	20.30	44.56	105.54	216.49	797.59	1675.74	3655	7838
Suisse	10.16	20.41	44.99	102.55	211.75	787.54	1673.16	3686	7643
Ukraine	10.07	20.00	45.11	105.08	210.33	790.78	1679.80	3711	7635
USA	9.69	19.32	43.18	102.60	209.30	776.27	1633.98	3583	7538

→ Matrice $\boldsymbol{M}$

On peut aussi chercher le vecteur (de norme 1) qui maximise la variabilité entre les scores

→ Vecteur optimal = 1er vecteur propre ($\boldsymbol{v_1}$) de l'ACP

→ Niveau de variabilité = 1ere valeur propre (λ_1) de l'ACP

→ Vecteur de scores avec la plus grande variabilité possible = $\boldsymbol{M} \cdot \boldsymbol{v_1}$

Records nationaux (en secondes) de p = 9 épreuves d'athlétisme pour n = 26 pays

	100m	200m	400m	800m	1500m	5000m	10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	1638.14	3609	7633
Canada	9.84	20.17	44.44	103.68	211.71	793.96	1656.01	3650	7809
Chine	10.17	20.54	45.25	106.44	216.49	805.14	1670.00	3635	7695
Croatie	10.25	20.76	45.64	104.07	213.30	817.76	1704.32	3827	8225
Ethiopie	10.50	21.08	45.89	106.08	211.13	757.35	1577.53	3535	7439
France	9.99	20.16	44.46	103.15	208.98	778.83	1642.78	3658	7596
Allemagne	10.06	20.20	44.33	103.65	211.58	774.70	1641.53	3634	7727
Inde	10.30	20.73	45.48	105.77	218.00	809.70	1682.89	3672	7920
Iran	10.29	21.11	46.37	104.74	218.80	833.40	1762.65	4103	8903
Italie	10.01	19.72	45.19	103.17	212.78	785.59	1636.50	3620	7642
Jamaïque	9.58	19.19	44.49	105.21	219.19	813.10	1712.44	3816	8199
Japon	10.00	20.03	44.78	106.18	217.42	793.20	1655.09	3625	7576
Kenya	10.26	20.43	44.18	102.01	206.34	759.74	1587.85	3513	7467
Lituanie	10.33	20.88	45.73	106.64	220.90	797.90	1651.50	3851	7955
NouvelleZélande	10.11	20.42	46.09	104.30	212.17	790.19	1661.95	3732	7815
Portugal	9.86	20.01	46.11	104.91	210.07	782.86	1632.47	3665	7596
Russie	10.10	20.23	44.60	102.47	212.28	791.99	1673.12	3675	7747
AfriqueduSud	10.06	20.11	44.59	102.69	213.56	794.16	1649.94	3678	7593
Espagne	10.14	20.59	44.96	103.83	208.95	782.54	1634.44	3592	7562
Suède	10.18	20.30	44.56	105.54	216.49	797.59	1675.74	3655	7838
Suisse	10.16	20.41	44.99	102.55	211.75	787.54	1673.16	3686	7643
Ukraine	10.07	20.00	45.11	105.08	210.33	790.78	1679.80	3711	7635
USA	9.69	19.32	43.18	102.60	209.30	776.27	1633.98	3583	7538

→ Matrice ***M***

Une fois enlevée l'influence de v_1, on cherche le vecteur (de norme 1) qui maximise la variabilité

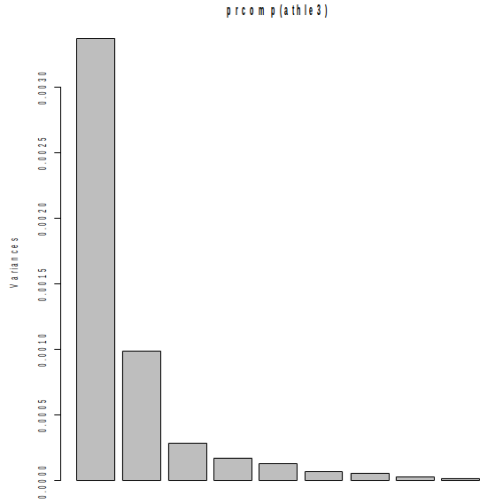
→ Vecteur optimal = 2er vecteur propre ($\boldsymbol{v_2}$) de l'ACP

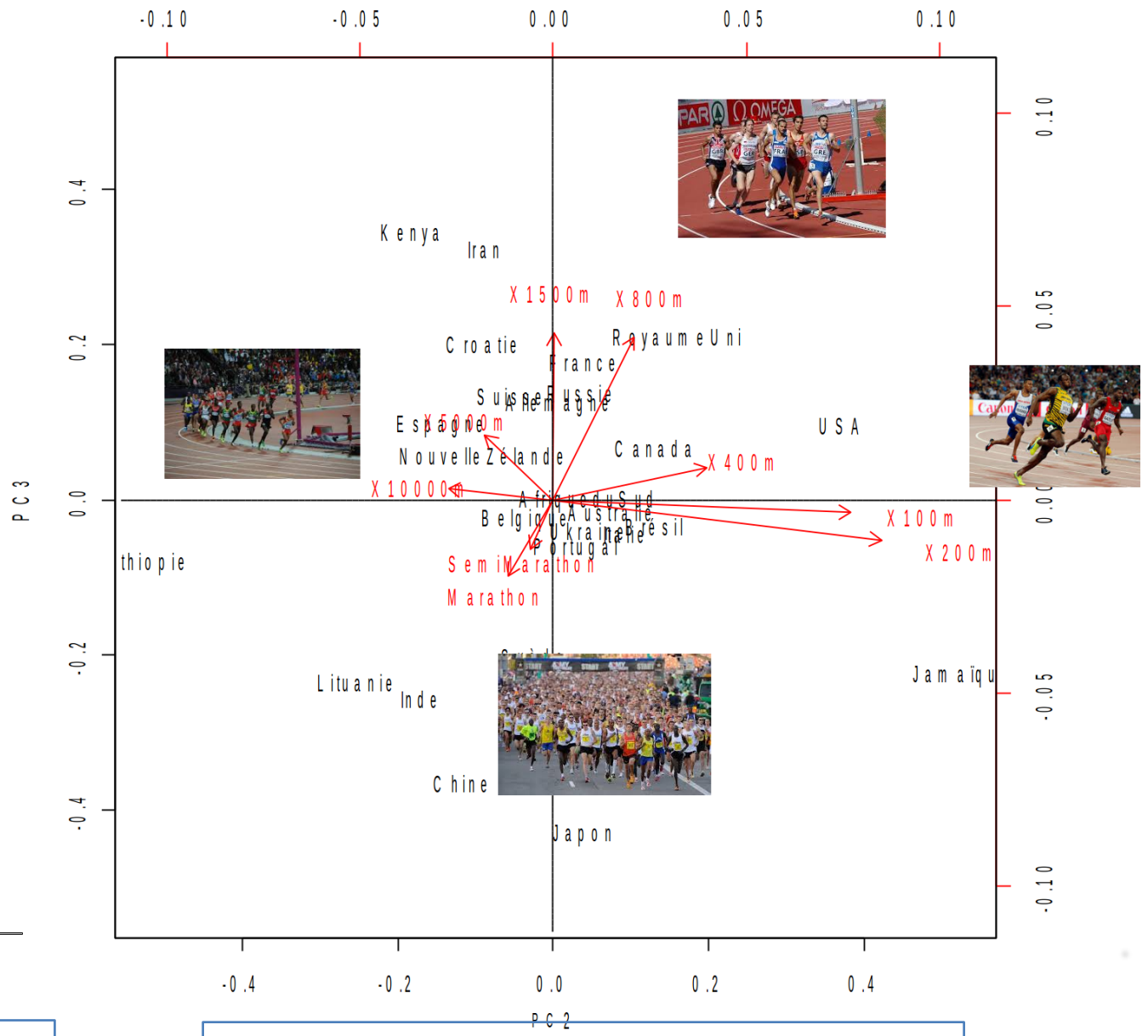
→ Niveau de variabilité = 2ere valeur propre (λ_2) de l'ACP

→ …

Calculable de manière analytique

6) Réduction de dimension par ACP





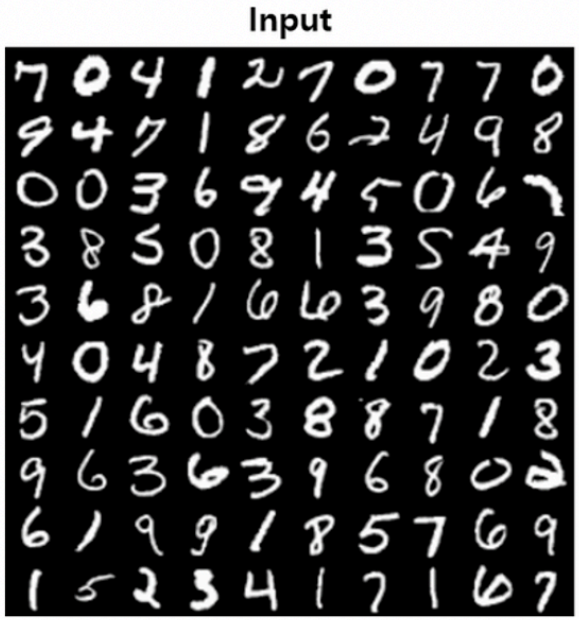
Éboulis des valeurs propres
(= variabilité capturée sur chaque direction principale)

En noir : Projection des données sur PC2 et PC3

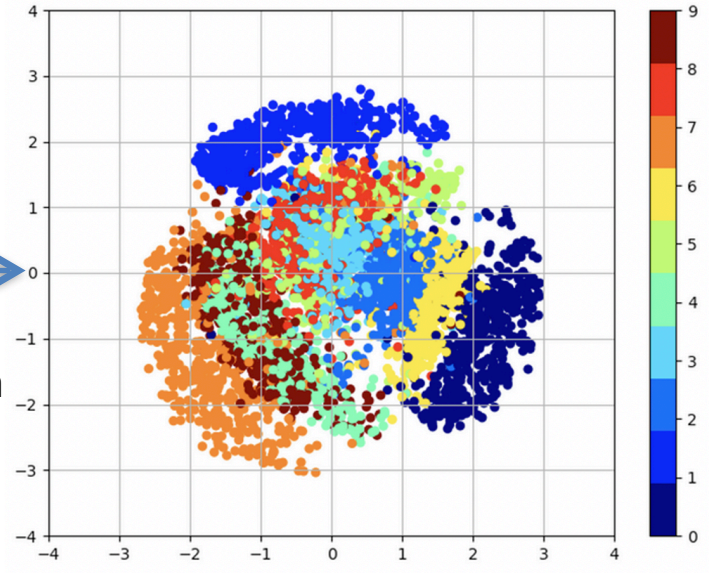
En rouge : Influence des variables dans PC2 et PC3

Pour aller plus loin

1. La réduction de dimension est importante pour l'apprentissage sur données en grande dimension
2. Au delà de l'ACP, un outil comme la PLS permet de réduire la dimension des observations de manière supervisée.
3. Les outils récents de réseaux de neurones profonds ont révolutionné l'apprentissage automatique en apprenant à classifier/régresser et en réduisant la dimension des données simultanément



Réduction de dimension avec un autoencodeur $\mathbb{R}^{1024} \to \mathbb{R}^2$



- Champ de recherche très actif et applications nombreuses
- Formalisme de description des données bien défini
- Beaucoup de perspectives actuellement liées à l'acceptabilité et la loyauté de l'I.A.

MERCI !!!