

# Machine learning

## 1. Apprentissage statistique



# Plan

Introduction

Formalisation du problème

Pertes et risques

Biais et variance

Validation croisée

Critères d'évaluation de modèles

# Plan

Introduction

Formalisation du problème

Pertes et risques

Biais et variance

Validation croisée

Critères d'évaluation de modèles

## Apprentissages supervisé et non-supervisé

- ▶ **Apprentissage supervisé :**

Inférer (prédirer) une fonction ou une relation à partir de **données d'apprentissage labellisées** (ex : classification supervisée, régression).

- ▶ **Apprentissage non-supervisé :**

Trouver une « structure » dans des **données non-labellisées** (ex : clustering). Même s'il est plus « subjectif » que l'apprentissage supervisé, il peut être utile comme étape de pré-traitement pour l'apprentissage supervisé.

## Quelques méthodes d'apprentissage supervisé

- ▶ **Statistique « classique »** : régression linéaire, régression paramétrique non-linéaire, régression logistique, méthodes de régularisation (Ridge, Lasso, Lars), PLS.
- ▶ **Méthodes bayésiennes.**
- ▶ **Méthodes de moyennage local** : plus proches voisins, noyau de lissage, CART.
- ▶ **Méthodes à bases de splines** : régression spline, GAM.
- ▶ **Méthodes à directions révélatrices** : SIM, SIR, etc.
- ▶ **Agrégation** : bagging, boosting.
- ▶ **Méthodes à noyau** : SVM.
- ▶ **Réseaux de neurones.**

## Exemples d'apprentissage supervisé

- ▶ Régression :
  - ▶ Pollution.
  - ▶ Vente de produits.
  - ▶ Prix de marché.
- ▶ Classification supervisée :
  - ▶ Médecine.
  - ▶ Credit scoring.
  - ▶ Reconnaissance de texte.
  - ▶ Reconnaissance d'images.

## Modélisation et/ou prévision

- ▶ On peut distinguer **modélisation** et **prévision**, par exemple compression d'image vs reconnaissance d'images.
- ▶ Un modèle s'appuie sur la **régularité** des phénomènes sous-jacents.
- ▶ La **prévision** consiste à **généraliser** un modèle.

## Un exemple

- ▶ Source : <http://yann.lecun.com/exdb/mnist/>.

- ▶ Modéliser :

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3  
4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4  
5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5  
6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6  
7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7  
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8  
9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9

- ▶ Prévoir :

## Quelques enjeux en prévision

- ▶ Compromis entre la **qualité de la prévision** et l'**interprétabilité** (notion de « boîte noire »).
- ▶ Privilégier des **modèles parcimonieux** (« **sparse** ») qui éviteront le **sur-apprentissage** : *less is more.*

# Plan

Introduction

Formalisation du problème

Pertes et risques

Biais et variance

Validation croisée

Critères d'évaluation de modèles

# Données

- ▶ On dispose d'un échantillon de  $(X, Y)$  :

$$\mathcal{D}_n = (X_i, Y_i)_{i \in \{1, \dots, n\}}$$

où  $X \in \mathcal{X}$  et  $Y \in \mathcal{Y}$ .

- ▶ On note :

$$d_n = (x_i, y_i)_{i \in \{1, \dots, n\}}.$$

## Objectif

On se placera dans le cadre de la **prévision** : on souhaite prévoir  $y$  pour une nouvelle valeur  $x$ .

## Covariables

On considèrera très souvent dans la suite que :

$$X \in \mathbb{R}^p .$$

Par défaut les covariables seront considérées comme quantitatives mais on indiquera régulièrement comme traiter les variables qualitatives.

## Régression et classification supervisée

- ▶ **Régression** : la variable  $Y$  est quantitative.  
Dans la suite on considèrera que  $Y \in \mathbb{R}$ .  
Mais il est possible de considérer plus généralement  $Y \in \mathbb{R}^d$ .
- ▶ **Classification supervisée** : la variable  $Y$  est qualitative.  
Dans la suite on considèrera que  $Y \in \{-1, 1\}$ .  
Par défaut la classification supervisée sera considérée binaire mais on indiquera régulièrement comme traiter plus de 2 modalités.

## Prévision

- ▶ On suppose que  $(x_i, y_i)$  est la **réalisation** d'une v.a.r  $(X_i, Y_i)$  de loi de probabilité inconnue  $P_{X,Y}$  (modèle statistique non-paramétrique).
- ▶ La fonction de prévision de  $Y$  est une fonction  $f : \mathcal{X} \rightarrow \mathcal{Y}$ .
- ▶ On suppose que  $f \in \mathcal{F}$ .
- ▶ Dans la suite, de manière plus spécifique que  $f$ , on désignera la fonction de lien par :
  - ▶ Cas de la **classification supervisée** :  $g$  .
  - ▶ Cas de la **régression** :  $m$  .
- ▶ On cherche à **estimer**  $f$  par  $\hat{f}$ .

# Plan

Introduction

Formalisation du problème

Pertes et risques

Biais et variance

Validation croisée

Critères d'évaluation de modèles

## Qualité d'un prédicteur

- ▶ La qualité d'un prédicteur  $\hat{f}$  est évaluée par le **risque**  $R$  (ou encore **erreur de généralisation**) qui :
  - ▶ permet de sélectionner un modèle,
  - ▶ fournit un indice de la confiance qu'on peut avoir en une prévision.
- ▶ Le risque est définie à partir d'une **fonction de coût** (ou encore **fonction de perte**).

## Fonctions de perte

- ▶ On appelle **fonction de perte** une fonction  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$  telle que :
  - ▶  $\ell(y, y) = 0$ ,
  - ▶  $\forall y \neq y' : \ell(y, y') > 0$ .
- ▶ Exemples de fonctions de perte :
  - ▶ Cas de la **classification supervisée binaire** :

$$\ell(y, y') = \mathbb{1}_{y \neq y'} = \frac{|y - y'|}{2} = \frac{(y - y')^2}{4}.$$

- ▶ Cas de la **régression** :

$$\ell(y, y') = |y - y'|^q$$

avec  $q \in \mathbb{R}^+$ .

## Risque (erreur de généralisation)

Le **risque** (ou erreur de généralisation) d'un prédicteur  $\hat{f}$  est défini par :

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

## Oracle

Si on connaissait  $P_{X,Y}$ , on pourrait déterminer le prédicteur optimal, appelé **oracle** :

$$f^* = \arg \min_{f \in \mathcal{F}} R(f) .$$

## Exemples d'oracles

- ▶ Cas de la classification supervisée binaire :

Si  $\ell(y, y') = \mathbb{1}_{\{y \neq y'\}}$  alors :

$$g^*(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 / X = x) \geq \mathbb{P}(Y = -1 / X = x) \\ -1 & \text{sinon} \end{cases} .$$

- ▶ Cas de la régression :

- ▶ Si  $\ell(y, y') = |y - y'|$  alors :

$$m^*(x) = \text{Med}(Y / X = x) .$$

- ▶ Si  $\ell(y, y') = (y - y')^2$  alors :

$$m^*(x) = \mathbb{E}(Y / X = x) .$$

## Enjeu

- ▶ L'objectif du data scientist est de déterminer une estimation  $\hat{f}$  de  $f$ , à partir de l'échantillon, telle que :

$$R(\hat{f}) \approx R(f^*) .$$

- ▶ En pratique, pour estimer  $f \in \mathcal{F}$  :
  1. On restreint  $\mathcal{F}$  à  $\mathcal{S}$ .
  2. On considère le risque empirique  $R_n$  (et non le risque).

D'où :

$$\hat{f} = \arg \min_{f \in \mathcal{S}} R_n(f) .$$

## Risque empirique

Le risque empirique est défini par :

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

C'est un estimateur de  $R(\hat{f})$ .

# Plan

Introduction

Formalisation du problème

Pertes et risques

**Biais et variance**

Validation croisée

Critères d'évaluation de modèles

## Biais et variance d'un estimateur I

- ▶ Dans le cas général ( $\mathcal{F}$ ), on cherche :

$$f^* = \arg \min_{f \in \mathcal{F}} R(f) .$$

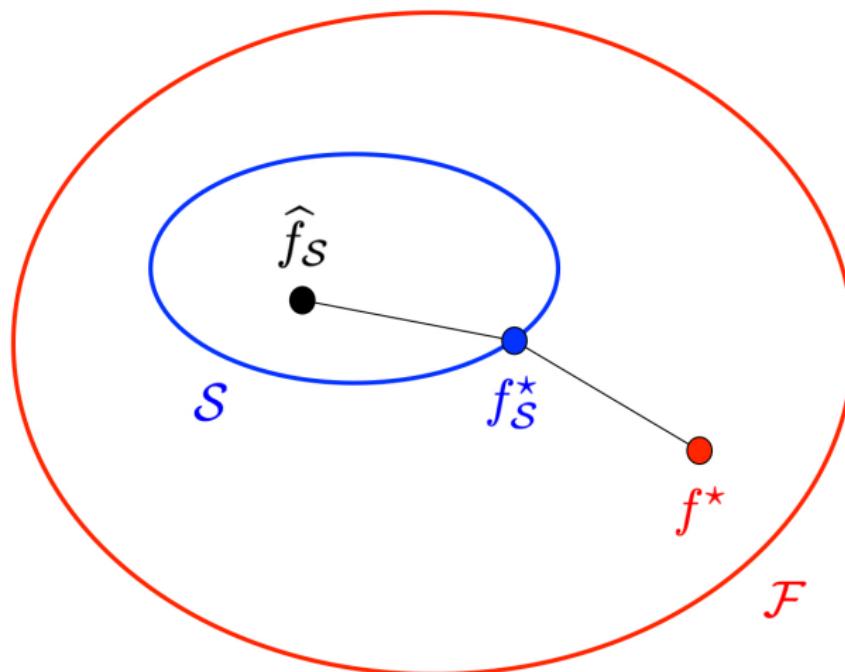
- ▶ Dans le cas restreint ( $\mathcal{S} \subset \mathcal{F}$ ), on cherche :

$$f_S^* = \arg \min_{f \in \mathcal{S}} R(f) .$$

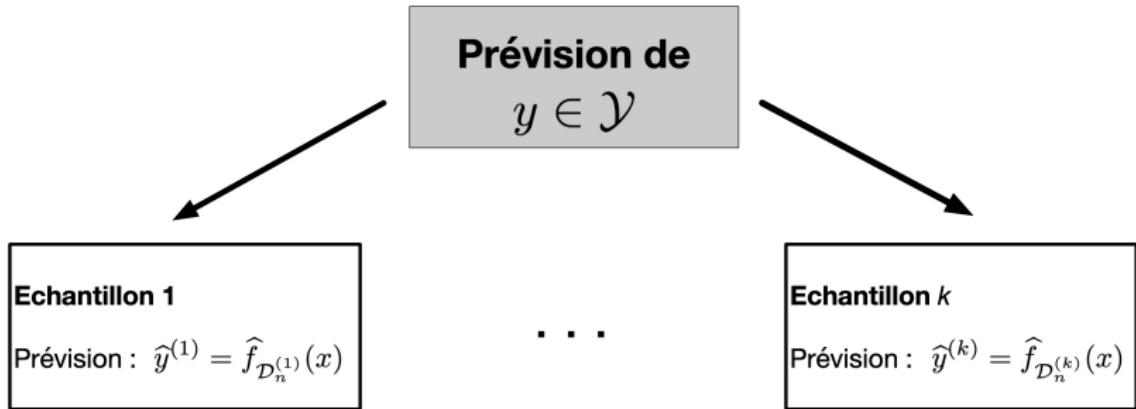
La décomposition **biais** (erreur d'approximation)-variance (erreur d'estimation) s'écrit :

$$R(\hat{f}_S) - R(f^*) = \underbrace{R(f_S^*) - R(f^*)}_{\text{erreur d'approximation}} + \underbrace{R(\hat{f}_S) - R(f_S^*)}_{\text{erreur d'estimation}} .$$

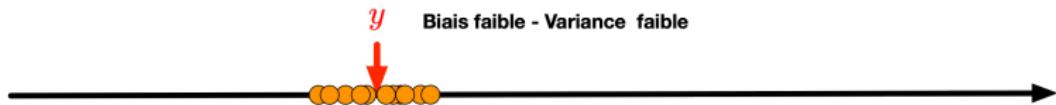
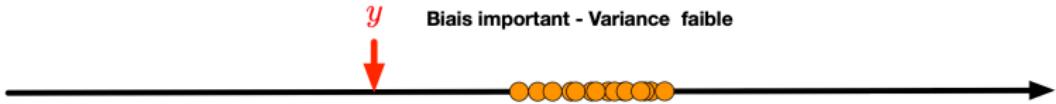
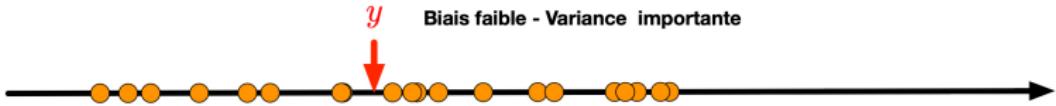
## Biais et variance d'un estimateur II



# Biais et variance d'un estimateur III



## Biais et variance d'un estimateur IV

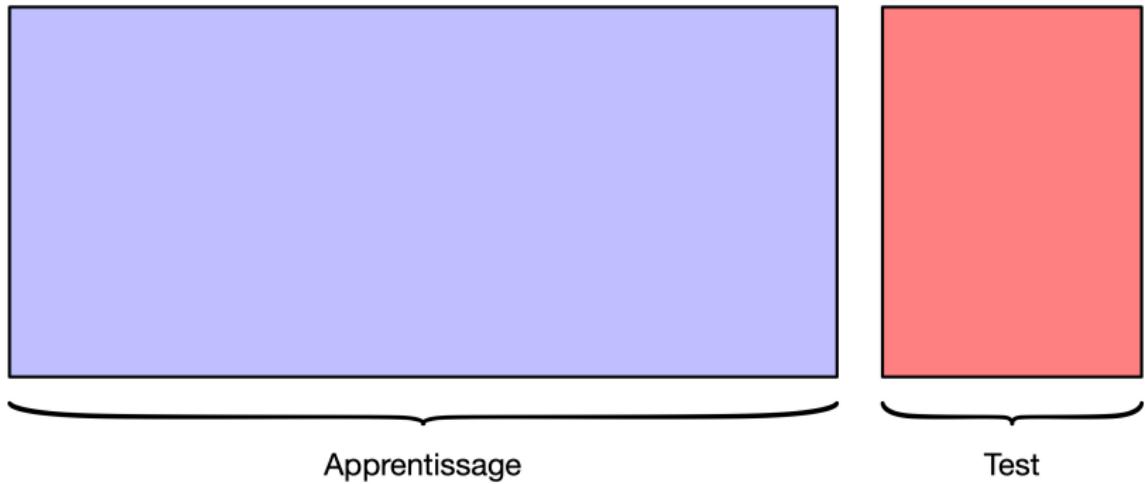


En orange :  $\hat{y}^{(1)}, \dots, \hat{y}^{(k)}$

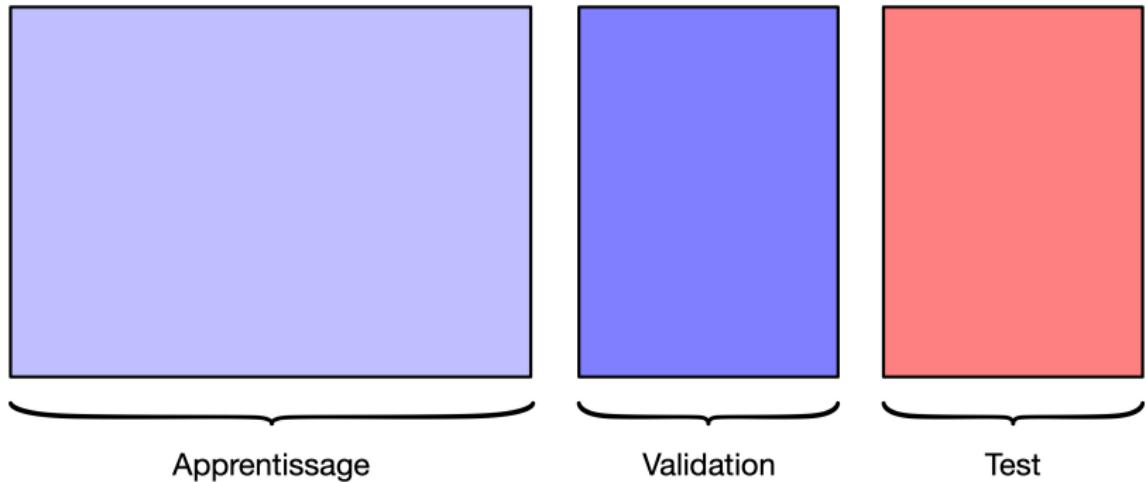
## Echantillons d'apprentissage et de test I



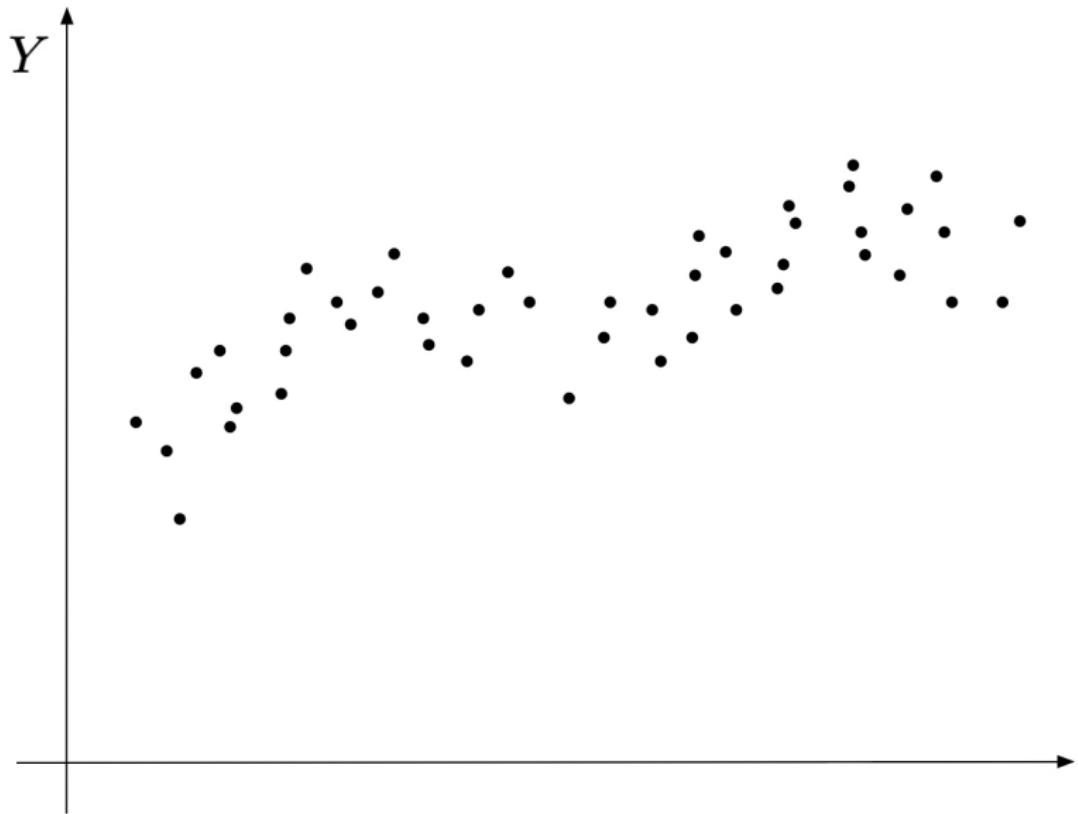
## Echantillons d'apprentissage et de test II



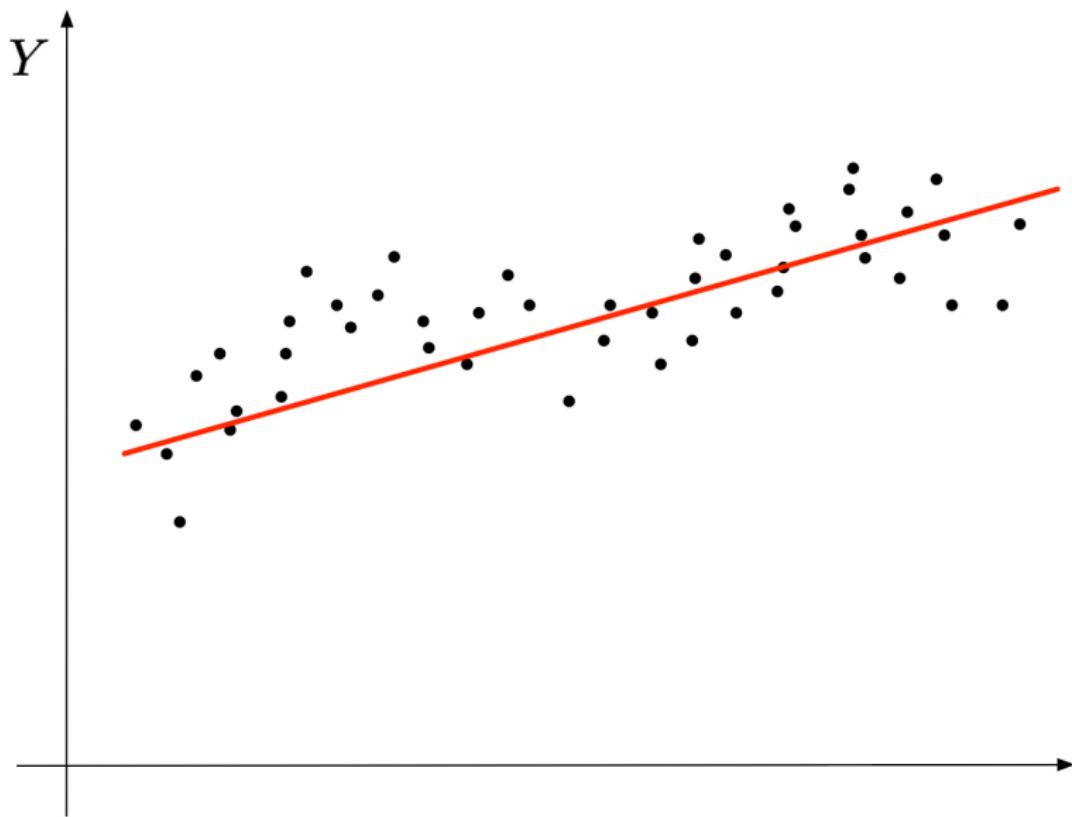
## Echantillons d'apprentissage et de test III



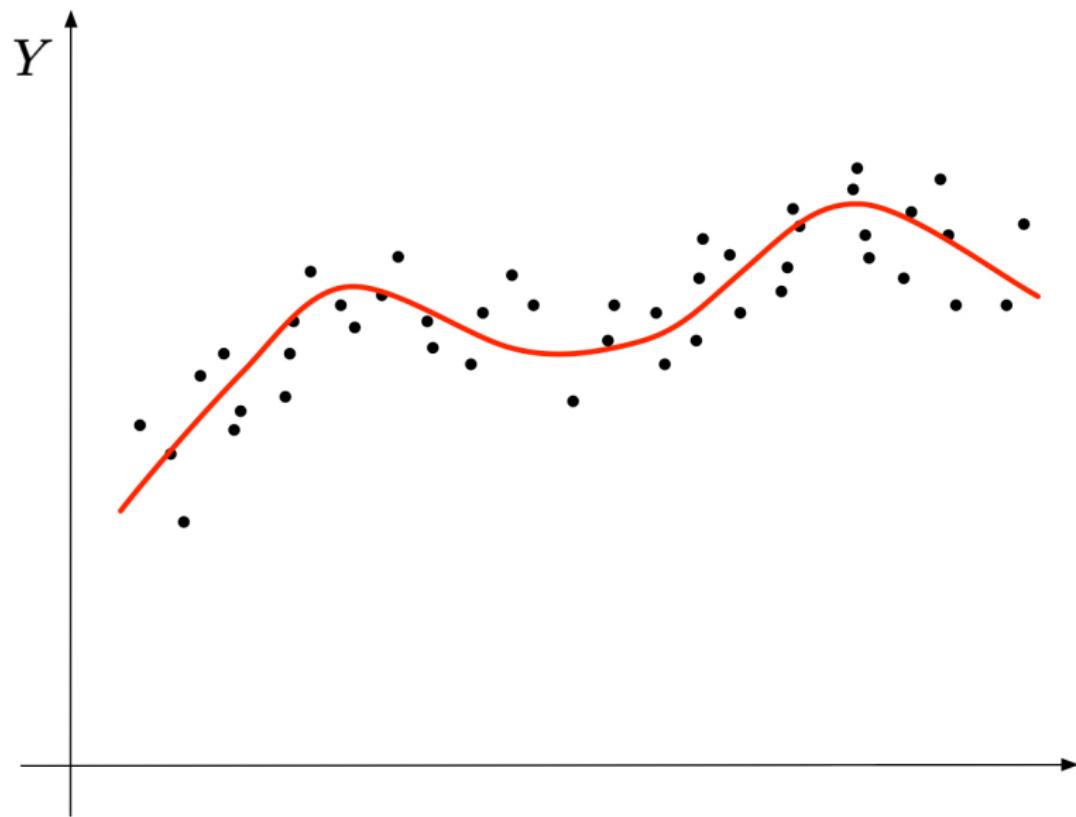
# Complexité I



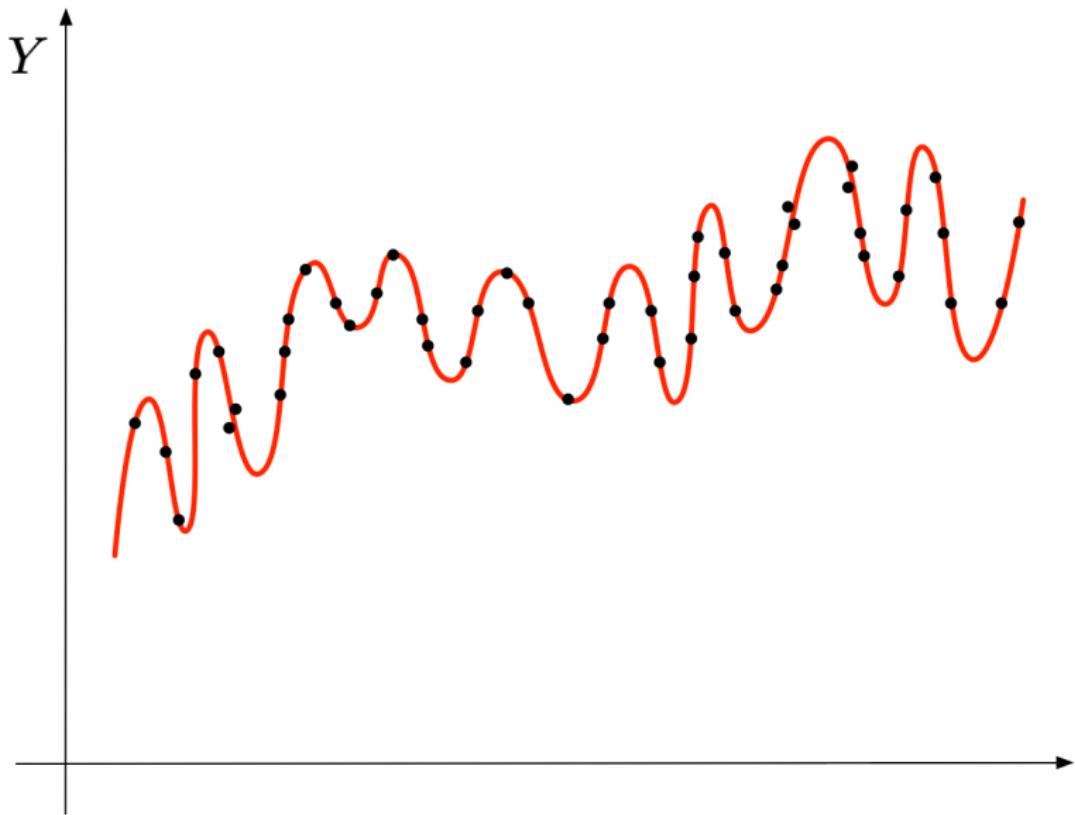
## Complexité II



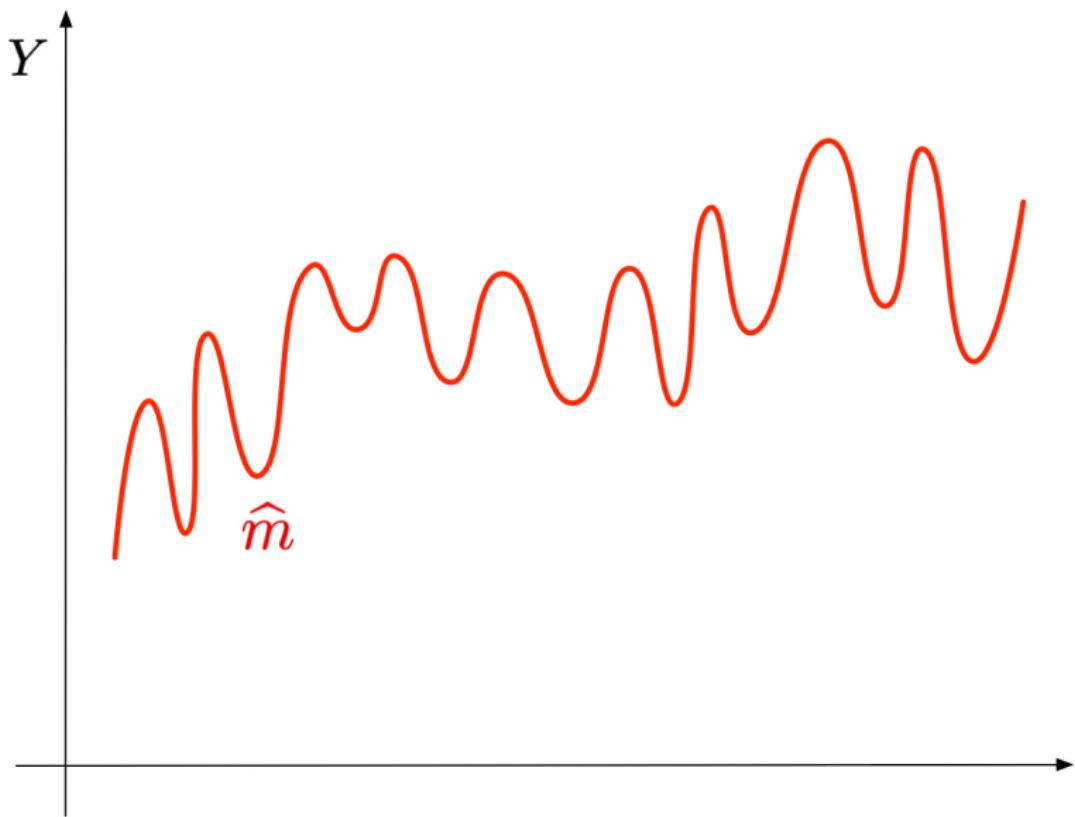
## Complexité III



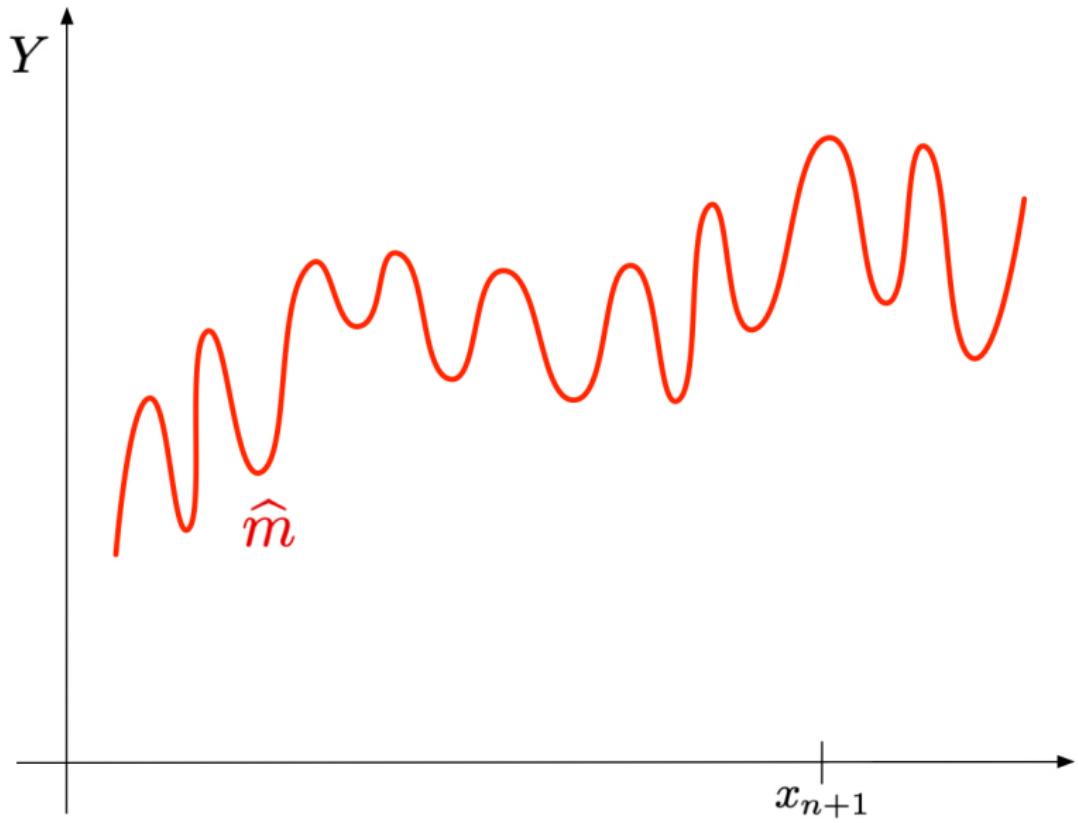
## Complexité IV



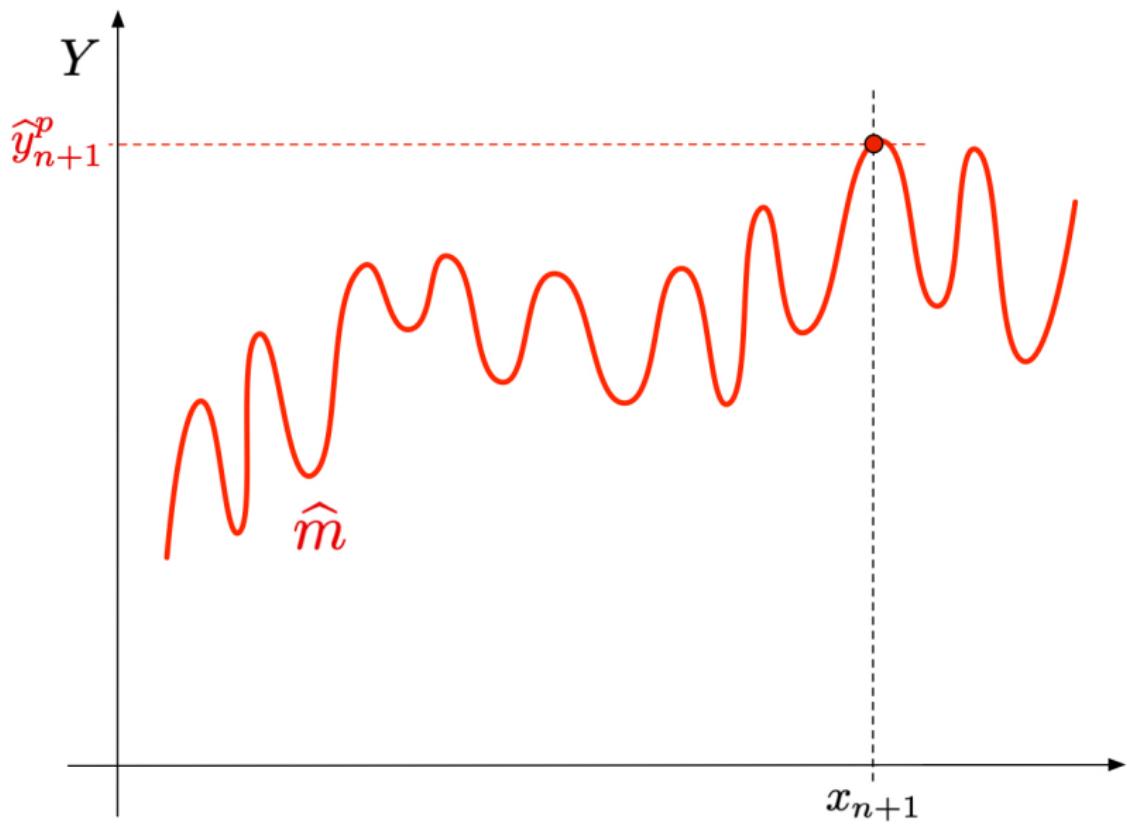
## Complexité V



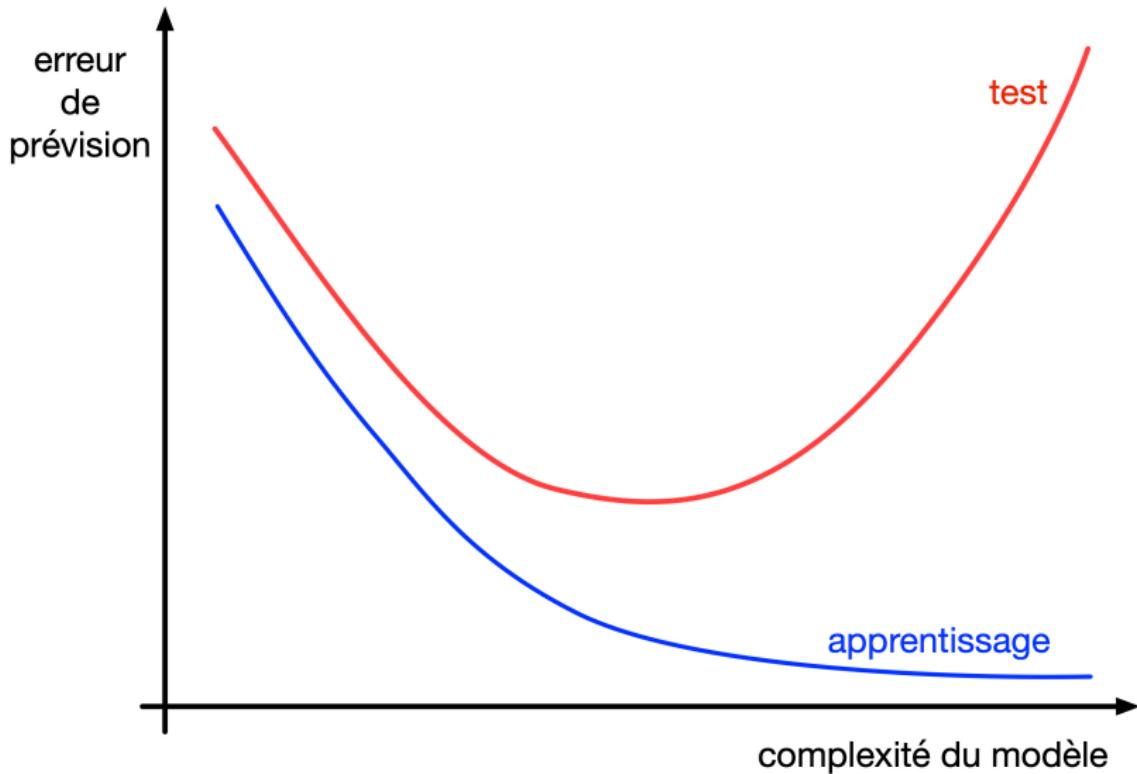
## Complexité VI



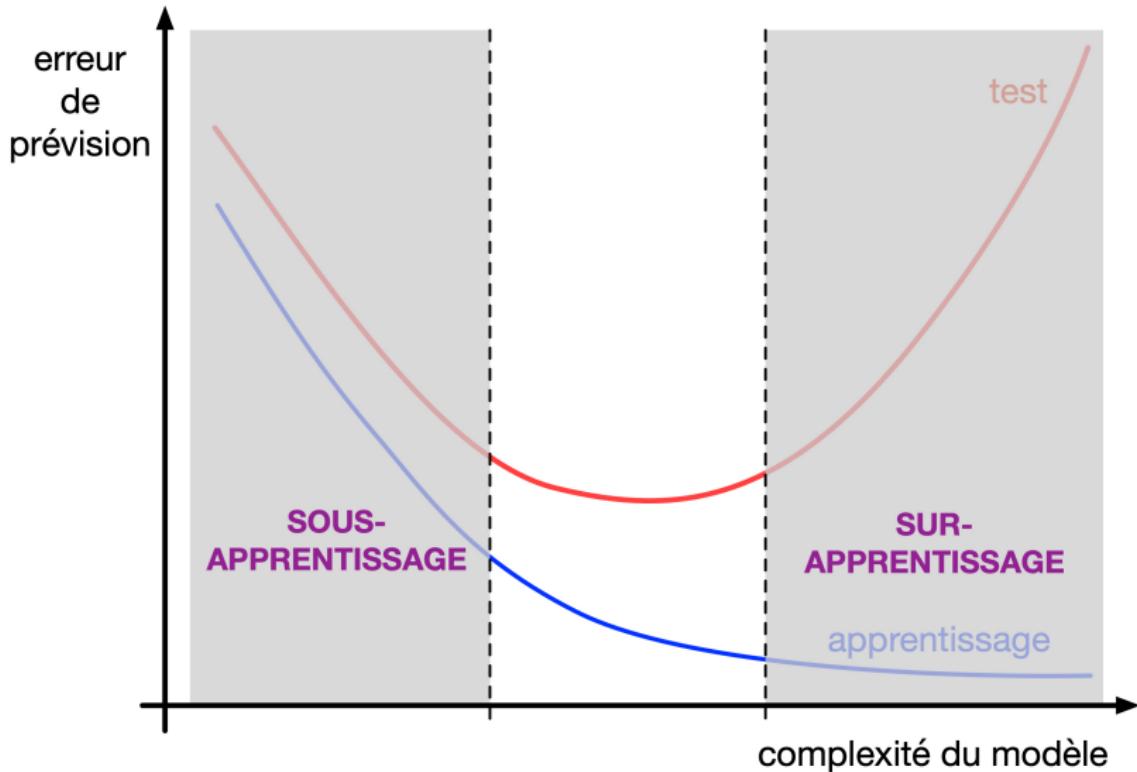
## Complexité VII



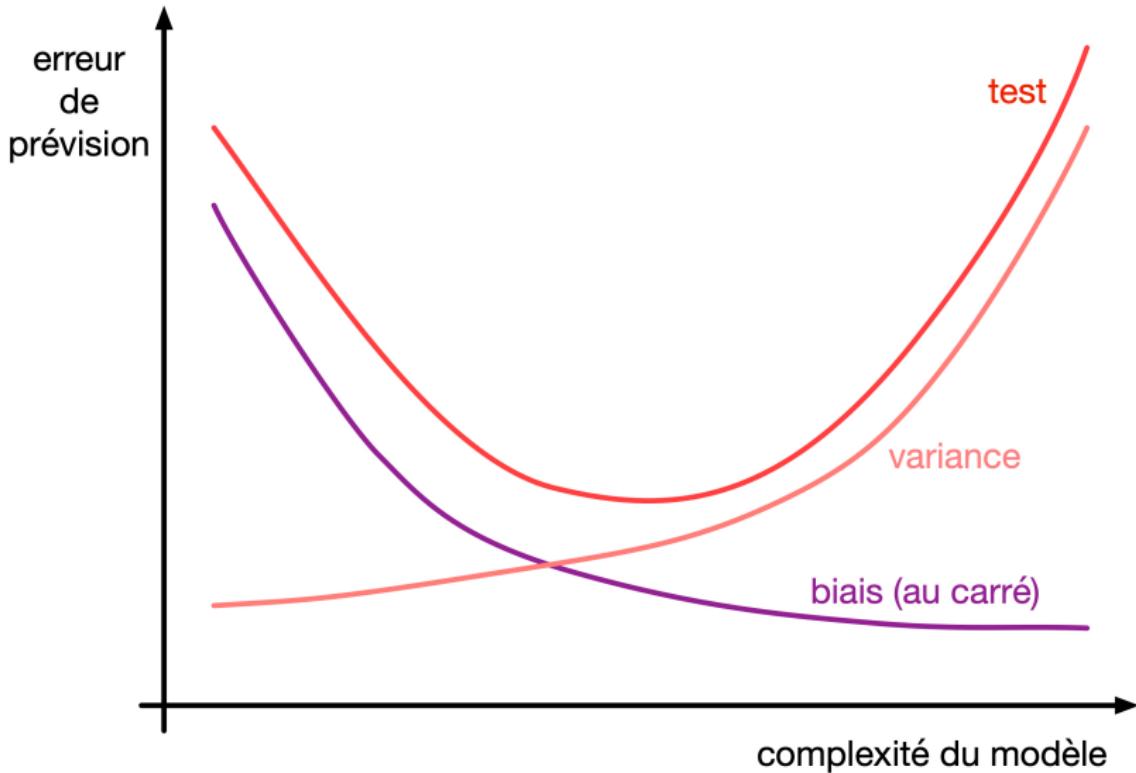
# Erreurs de prévision & complexité I



## Erreur de prévision & complexité II



## Erreur de prévision & complexité III



# Plan

Introduction

Formalisation du problème

Pertes et risques

Biais et variance

**Validation croisée**

Critères d'évaluation de modèles

## Retour sur le risque empirique

- ▶ Le risque empirique sous-estime le risque.
- ▶ Cela peut conduire à du **sur-apprentissage**.
- ▶ Il existe plusieurs parades pour obtenir un estimateur non-biaisé du risque :
  - ▶ Utilisation de critères tels que l'AIC, le BIC, le  $C_p$  de Mallows.
  - ▶ Méthodes de rééchantillonage : validation croisée ou bootstrap.

## Principe

1. Diviser aléatoirement les données en  $K$  blocs (égaux ou équivalents).

Le bloc  $k$  contient  $n_k$  observations :  $n_k = \frac{n}{K}$  si  $n$  est un multiple de  $K$ .

2. Pour  $k \in \{1, \dots, K\}$  :

- 2.1 Retirer le bloc  $k$  de la base d'apprentissage.
- 2.2 Estimer la fonction de prévision sur la base d'apprentissage.
- 2.3 Calculer un critère d'erreur de prévision sur le bloc  $k$  :  $CV_k$  (ex : MSE pour la régression).

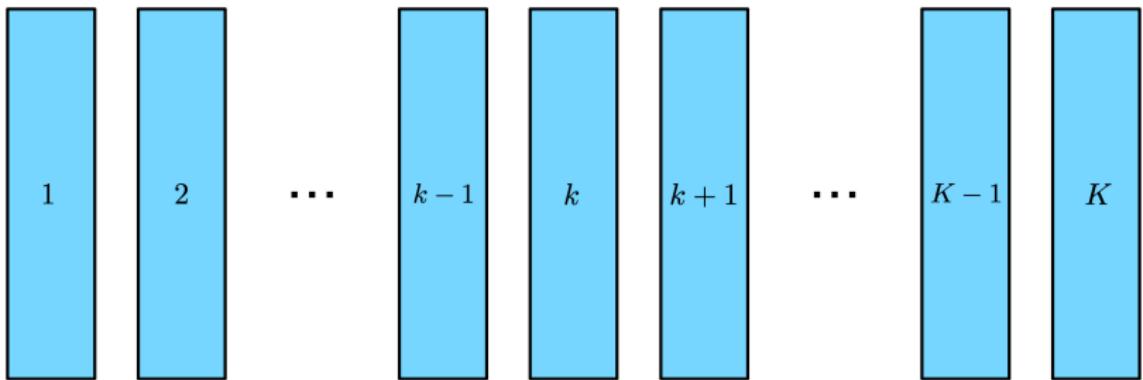
3. Calculer le critère de validation croisée :

$$CV = \sum_{k=1}^K \frac{n_k}{n} CV_k .$$

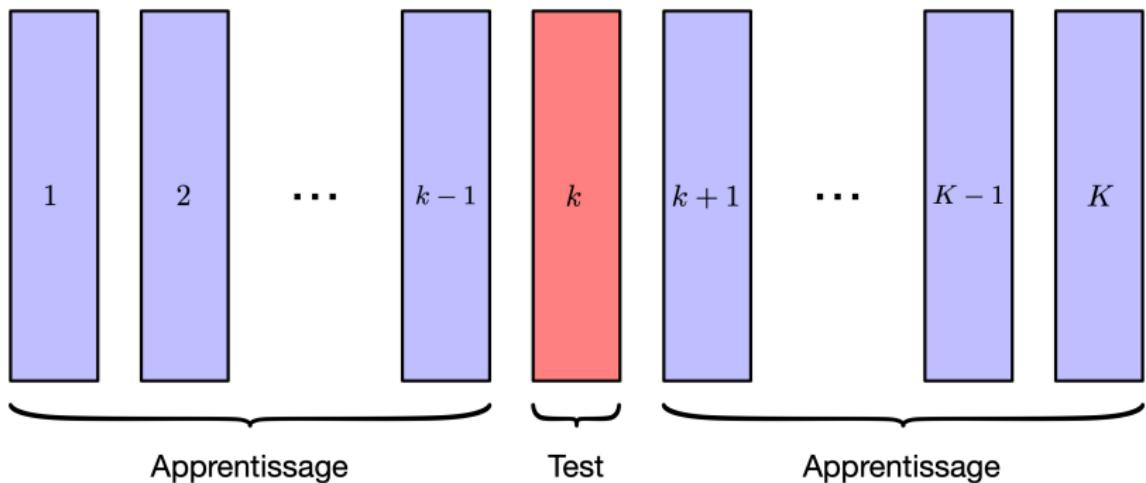
## Illustration I



## Illustration II



### Illustration III



## Remarques

- ▶ Usuellement :  $K = 5$  ou  $K = 10$ .
- ▶ Lorsque  $K = n$  : on parle d'estimateur « **leave one out** » (LOO)

# Plan

Introduction

Formalisation du problème

Pertes et risques

Biais et variance

Validation croisée

Critères d'évaluation de modèles

# Régression |

- Le RMSE (Root Mean Square Error) vaut :

$$\text{RMSE} = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\hat{y}_i - y_i)^2}.$$

- Le nRMSE (normalized RMSE) vaut :

$$\text{nRMSE} = \frac{\text{RMSE}}{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \hat{y}_i}.$$

## Régression II

- Le **MAE** (Mean Absolute Error) vaut :

$$\text{MAE} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} |\hat{y}_i - y_i| .$$

- Le **MAPE** (Mean Absolute Percent Error) vaut :

$$\text{MAPE} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times 100 .$$

## Classification supervisée I

- ▶ Dans le cas de la classification supervisée binaire, la **matrice de confusion** vaut :

		Prévision	
		1 (Positif)	0 (Négatif)
Vérité	1 (Positif)	Vrai positif (VP)	Faux négatif (FN)
	0 (Négatif)	Faux positif (FP)	Vrai négatif (VN)

- ▶ Dans le cas de la classification supervisée multi-classes, on peut établir la matrice de confusion, avec autant de lignes et de colonnes que de classes, et en déduire les nombres VP, FP, VN et FN.

## Classification supervisée II

Les indicateurs suivants prennent leurs valeurs sur  $[0, 1]$ , plus ils sont proches de 1, meilleur est le modèle.

- ▶ L'**exactitude** (*accuracy*) vaut :

$$\text{exactitude} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}} .$$

Notons que l'erreur de classification (*classification error*) vaut :

$$\text{erreur} = \frac{\text{FP} + \text{FN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}} = 1 - \text{exactitude} .$$

- ▶ La **spécificité** (*specificity*), le taux de négatifs classés négatifs (« vrais négatifs »), vaut :

$$\text{spécificité} = \frac{\text{VN}}{\text{FP} + \text{VN}} .$$

## Classification supervisée III

- ▶ La **précision** (*precision*), ou valeur prédictive positive, vaut :

$$\text{précision} = \frac{\text{VP}}{\text{VP} + \text{FP}}.$$

- ▶ La **sensibilité** (*sensitivity*), ou rappel (*recall*), est le taux de positifs classés positifs (« vrais positifs ») :

$$\text{sensibilité} = \frac{\text{VP}}{\text{VP} + \text{FN}}.$$

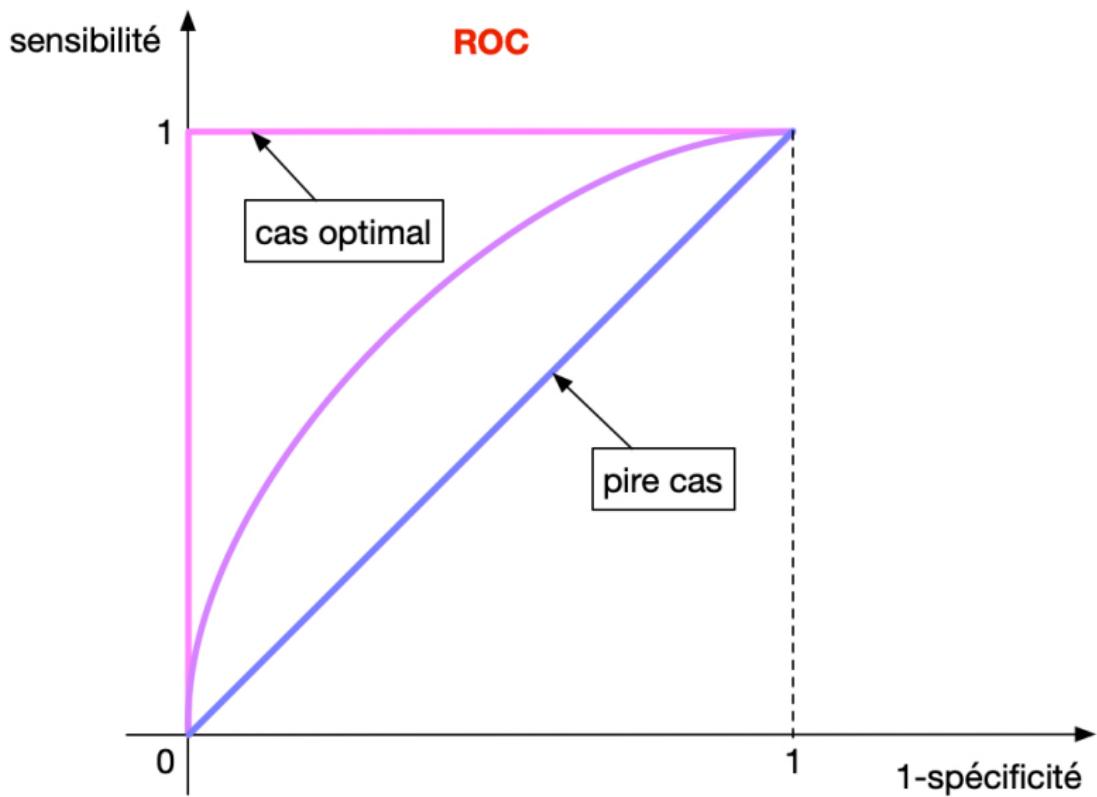
- ▶ Le score **F<sub>1</sub>** est la moyenne harmonique de la précision et de la sensibilité :

$$F_1 = 2 \frac{\text{précision} \times \text{sensibilité}}{\text{précision} + \text{sensibilité}}.$$

## Classification supervisée IV

- ▶ La courbe **ROC** (Receiver Operating Characteristic) représente la sensibilité (taux de vrais positifs) en fonction de l'anti-spécificité (taux de faux positifs) pour différents seuils de décision  $s$ .
- ▶ Plus le seuil  $s$  est important :
  - ▶ plus le taux de vrais positifs est important,
  - ▶ moins le taux de faux positifs est important.
- ▶ La courbe ROC est croissante et au-dessus de la première bissectrice (correspondant à une prédiction de type « tirage au sort »).
- ▶ La prédiction « optimale » fournirait une courbe ROC égale à 0 pour  $s = 0$  et égale à 1 pour  $s \in ]0, 1]$ .

## Classification supervisée V

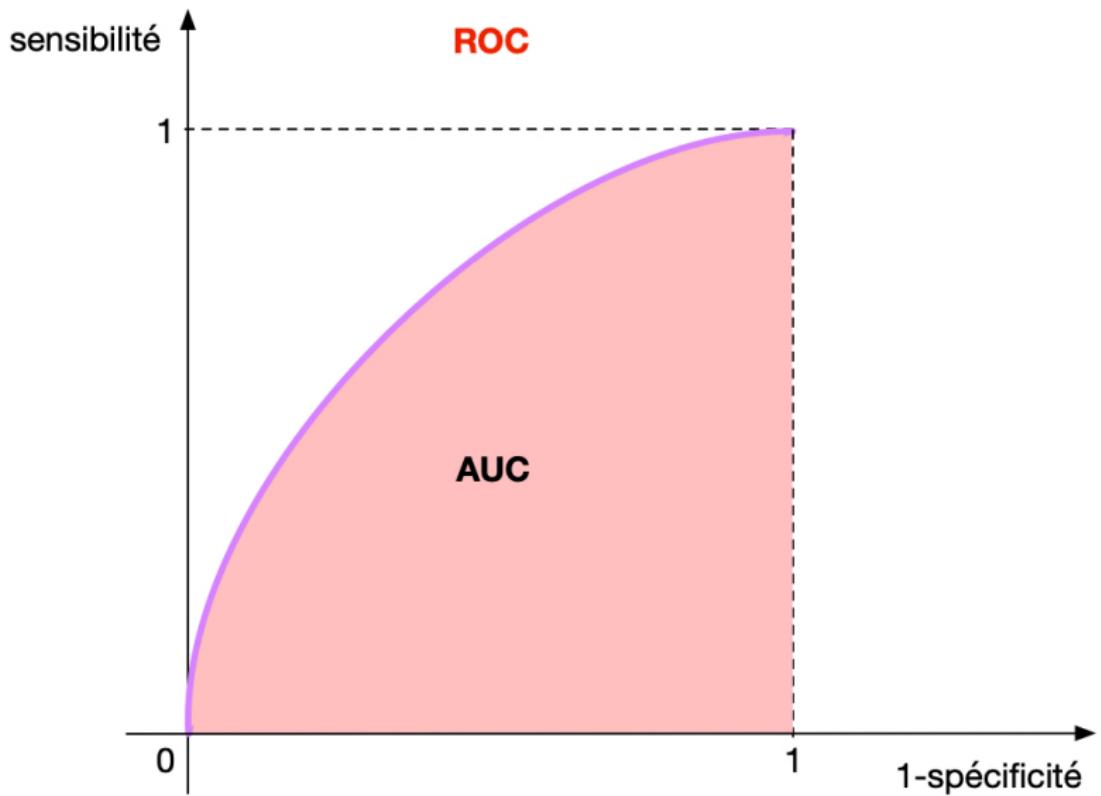


## Classification supervisée VI

L'aire sous la courbe ROC, l'**AUC** (Area Under the ROC), est une mesure de la qualité de la classification et varie entre :

- ▶  $\text{AUC} = \frac{1}{2}$  : le pire des cas (prédiction de type « tirage au sort »),
- ▶  $\text{AUC} = 1$  : le meilleur des cas (prédiction « optimale »).

## Classification supervisée VII



## Références

- Hastie, T., R. Tibshirani et J. H. Friedman. 2009, *The elements of statistical learning. Data Mining, inference, and prediction*, 2<sup>e</sup> éd., Springer Series in Statistics, Springer.
- James, G., D. Witten, T. Hastie et R. Tibshirani. 2015, *An introduction to statistical learning with applications in R*, Springer Texts in Statistics, Springer.