



Introduction à l'Apprentissage Statistique

Théorie du SL

M2 Actuariat – ISFA – 2021/2022

Pierrick Piette
Actuaire à Seyna
pierrick.piette@gmail.com

Seyna.

In God we trust, all others bring data
- William Edwards Deming





Principes Généraux

Statistique classique / Apprentissage

Statistique classique

- Recherche le **modèle génératif** des données
- Construit l'estimateur sur un jeu de données unique
- Théorique asymptotique permet de juger sa qualité

Apprentissage statistique

- Recherche les **bonnes prévisions**
- On souhaite faire apprendre à l'algo la relation entre X et Y ,
- puis la généraliser à des occurrences de X pour lesquelles Y est inconnue.
- La qualité est jugée en fonction d'une mesure d'adéquation à l'échantillon test.

Statistique classique / Apprentissage

Statistique classique

- Approche privilégiant la **compréhension**
- Compréhension du mécanisme génératrice
- Modèle simple et interprétable

Apprentissage statistique

- Approche privilégiant la **prévision**
- pour de nouveaux individus : pouvoir de généralisation
- Les modèles sont en fait des algorithmes

Paramétrique / Non Paramétrique

Estimation paramétrique

- On cherche m parmi une famille indexée par un paramètre de dimension finie
- Exemple : la régression linéaire $m(x) = a + bx$
- Une fonction candidate s'identifie à 2 paramètres

Estimation non paramétrique

- On ne fait plus d'hypothèse
- On cherche $m(x)$ parmi toutes les fonctions possibles (dimension infinie)
- Exemple : les estimateurs à noyaux

Supervisé / Non-supervisé

Deux grands domaines de l'apprentissage statistique : présence ou non d'une variable à expliquer

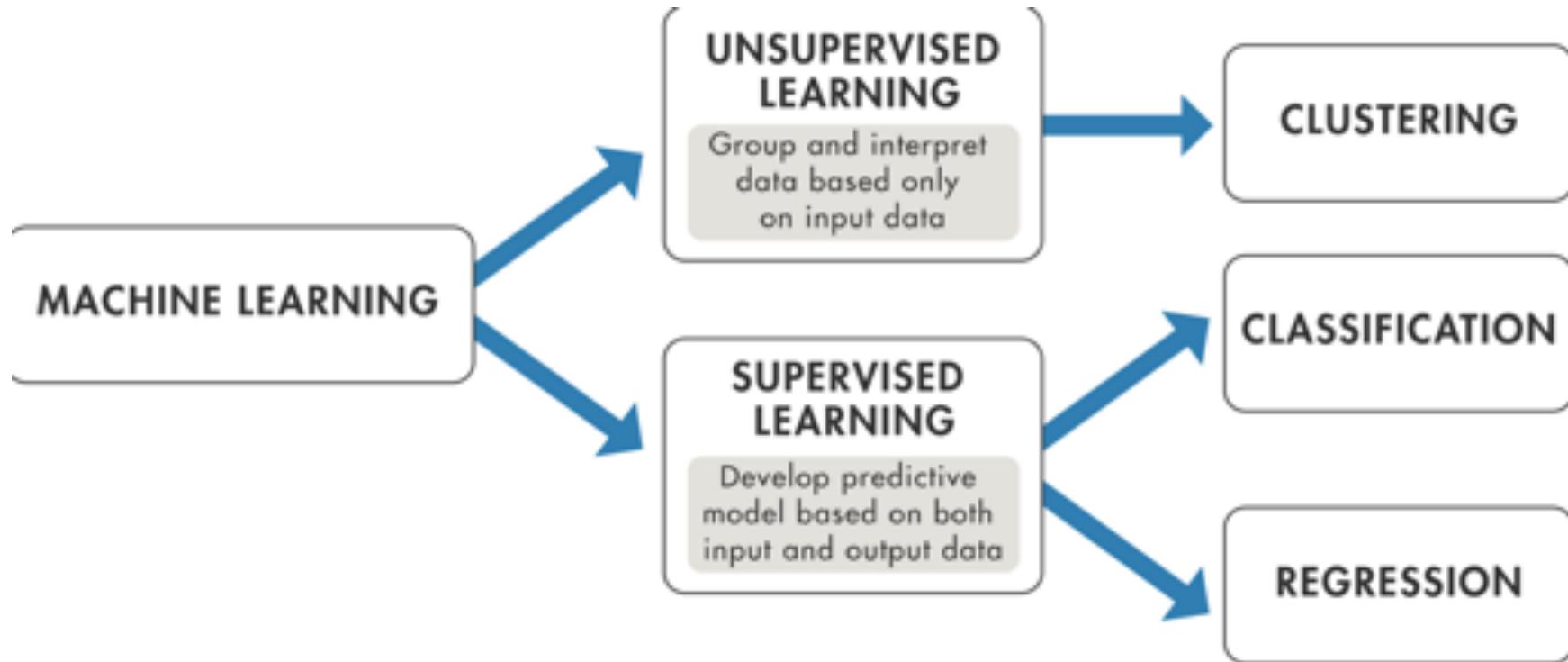
Supervisé

- Prédire une variable bien précise à partir d'autres variables
- $Y = f(X) + \epsilon$
- Optimisation du pouvoir de prédiction de f selon un critère de performance

Non-supervisé

- Découvrir des structures particulières dans les données
- Catégoriser les observations et connaître les variables les plus discriminantes
- Problème de clustering

Supervisé / Non-supervisé





Théorie de Vapnik

Qualité d'estimation

Théorème : Soit $X \in \mathbb{R}^d$, et m une fonction k fois dérivable à dérivées bornées. La vitesse optimale de convergence d'un estimateur non paramétrique \hat{m} est

$$\hat{m}(x) - m(x) = O(n^{\frac{-k}{2k+d}}) \text{ p.s.}$$

Si la fonction m est régulière (e.g. infiniment dérivable) à d fixé, la vitesse de convergence est en \sqrt{n}

Si d est « grand » par rapport à n , la performance d'estimation est considérablement dégradée.

Dimension Vapnik-Chervonenkis

Mesure de la capacité d'un algorithme de classification

Le nombre de points du plus grand ensemble que le modèle peut pulvériser (*shatter*)

- Exemple : la droite
- 3 points peuvent toujours être pulvériser
- Mais 4 pas forcément
- La dimension VC est de 3

Mesure de la complexité d'un modèle

Inégalité de Vapnik

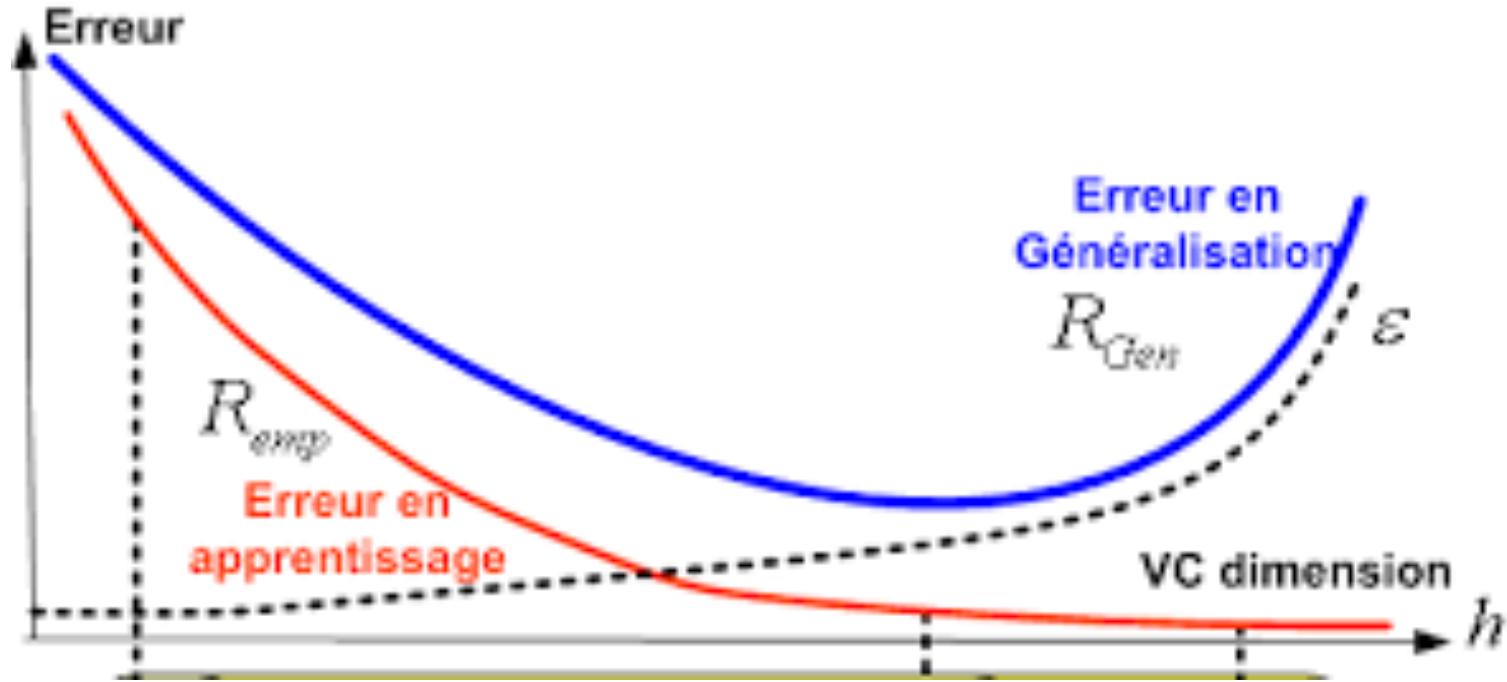
Soit un classificateur de dimension VC h , alors avec une grande probabilité $(1 - \eta)$, on a l'inégalité de Vapnik

$$\text{TestError} \leq \text{TrainError} + \text{GeneralizationError}$$

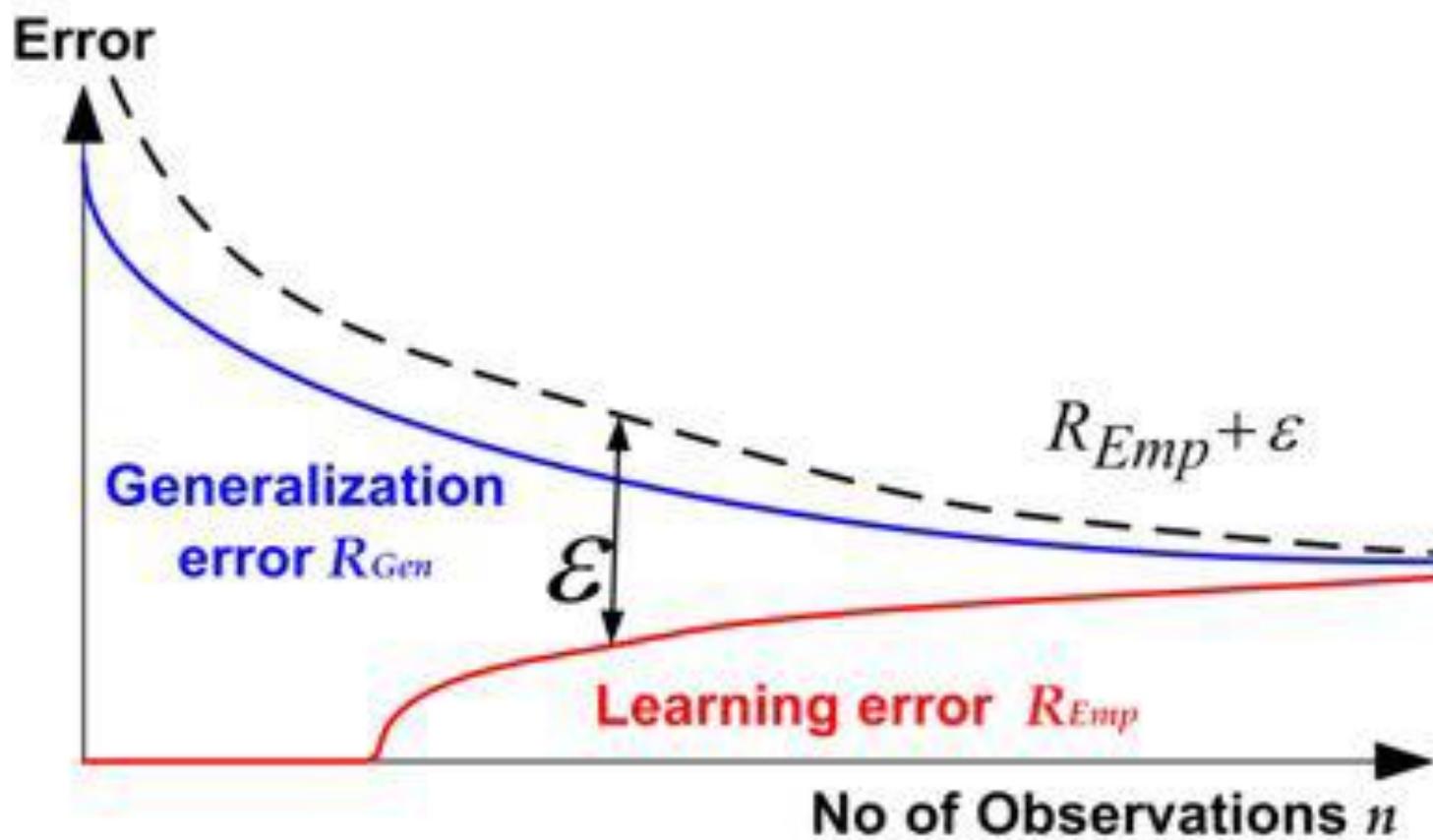
avec

$$\text{GeneralizationError} = \sqrt{\frac{h \ln\left(\frac{2n}{h}\right) + h - \ln\left(\frac{\eta}{4}\right)}{n}}$$

Théorie de Vapnik



Théorie de Vapnik



Remarques

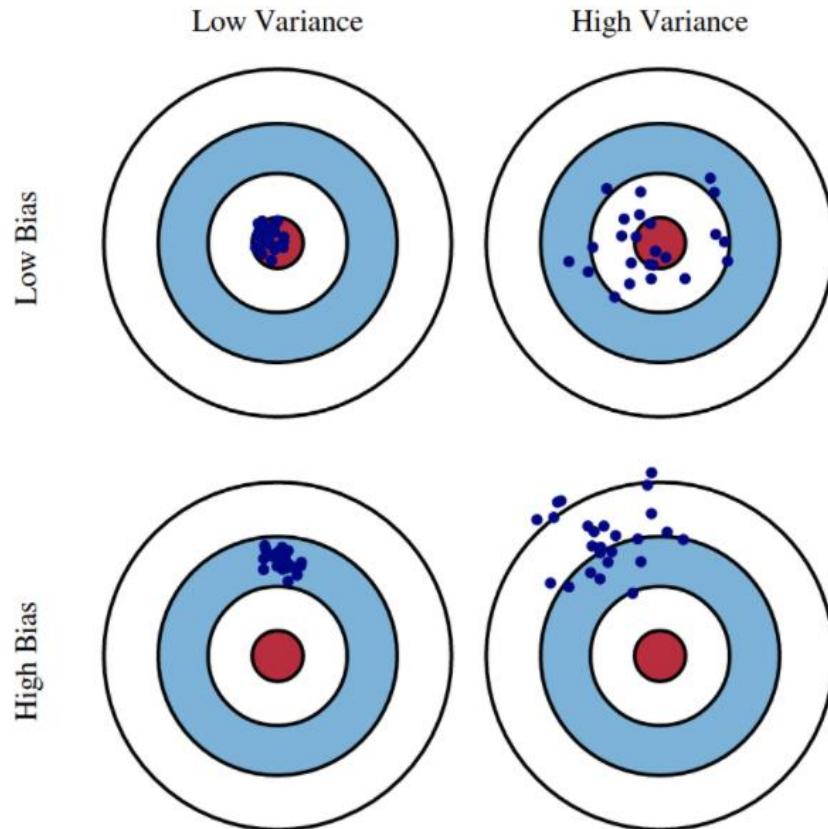
L'erreur de généralisation augmente quand la dimension VC augmente

- Les modèles de grande dimension ont un faible biais
- Mais une grande variance

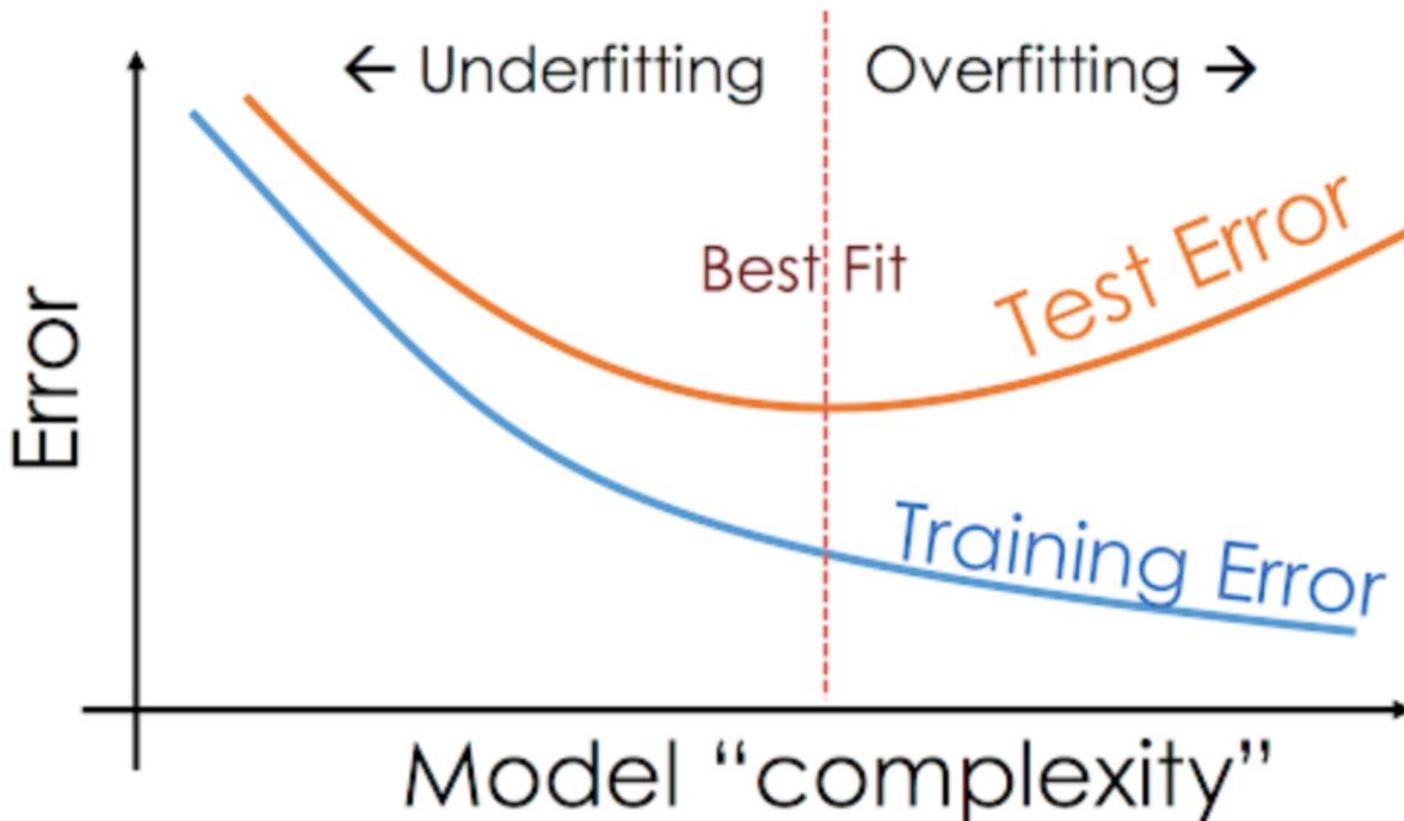
L'erreur est dépendante du rapport n/h , i.e. le rapport du nombre de données sur la complexité du modèle

- on augmente la capacité prédictive si h augmente mais moins vite que n
- On peut augmenter la complexité du modèle si on augmente aussi n

Biais - Variance

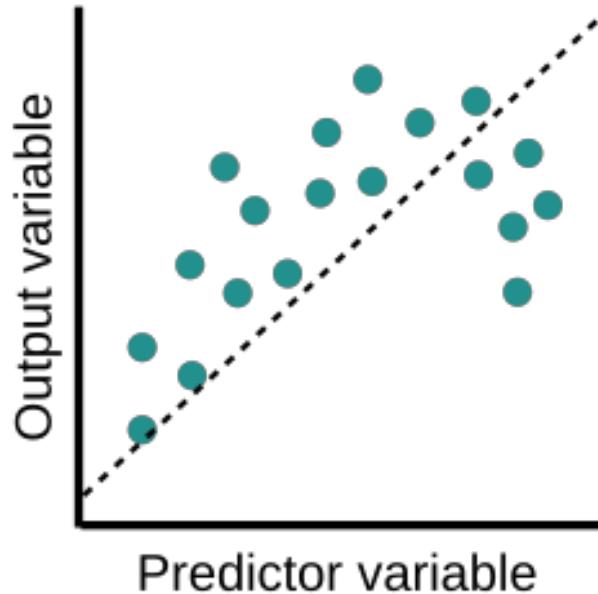


Underfitting / Overfitting

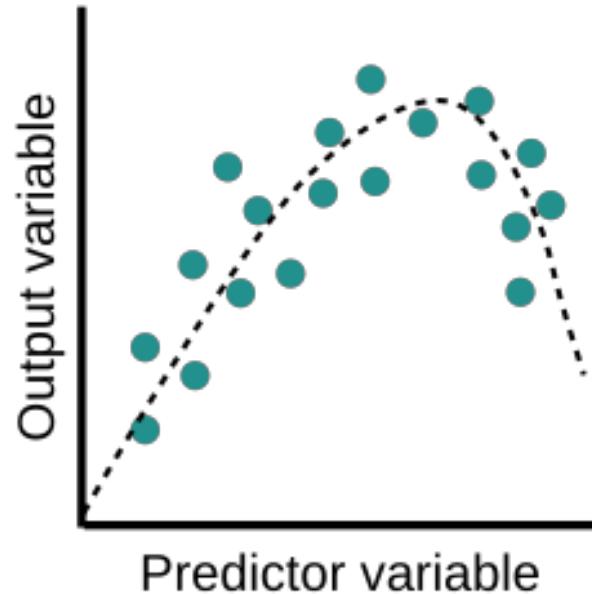


Underfitting / Overfitting

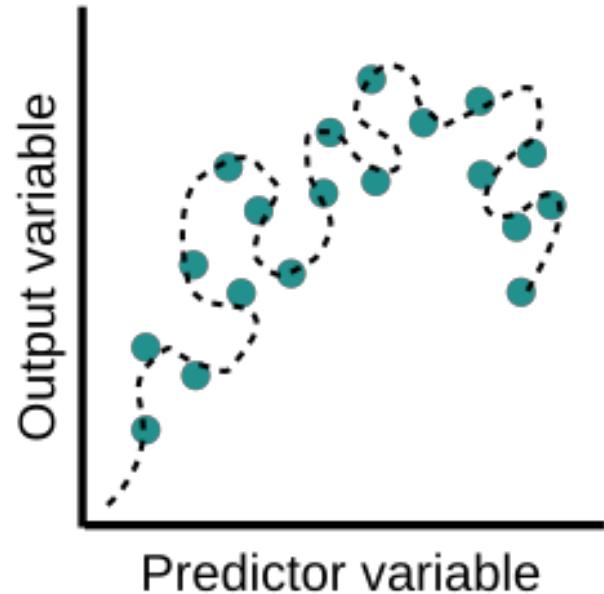
Underfit



Optimal



Overfit



Application en SL

Les méthodes d'apprentissage statistique introduisent le choix de paramètres de tuning ou hyper-paramètres

- Définissent la complexité du modèle
- et donc le pouvoir de généralisation

Il faut donc choisir leur valeur

- Validation croisée
- Ou création de plusieurs échantillons
- Apprentissage pour construire le modèle
- Validation pour optimiser les paramètres de tuning
- Test pour évaluer la performance du modèle

