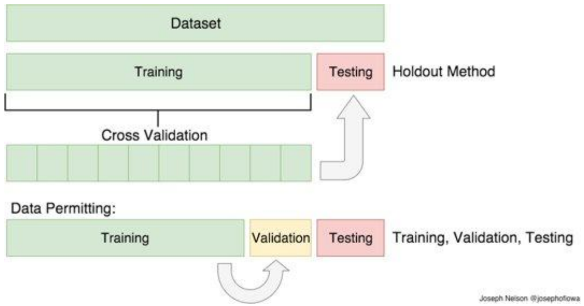
Cours 3: Biais, Variance, sur et sous-apprentissage

Gauthier Gidel
9 Septembre 2024

Qu'est-ce que cela signifie pour un modèle d'être performant ?

Performance:

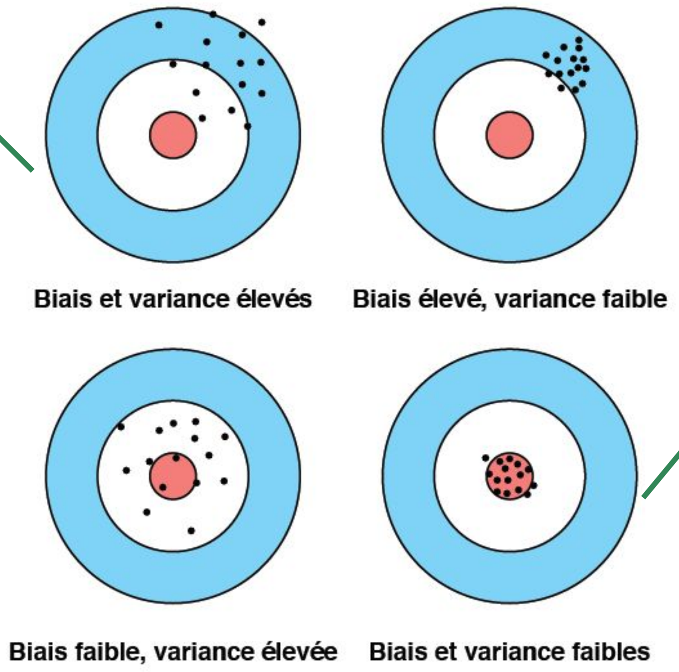
- On entraîne un modèle à prédire:
  - P(cat or dog | x, Θ)
- Il devrait **bien** modéliser cette proba.
- On évalue sur l'ensemble de **test**.
- Le concept important: **généralisation.**
    - Découvrez les meilleurs caractéristiques pour déterminer si une image contient un chat/chien
    - Bon exemples : A des moustaches, un iris vertical, des griffes ou non,
    - Mauvais exemple : le pixel 20,47 est gris, le blob est plus gros
- On étudie ce concept via le **biais et la variance.**



Exemples de distributions avec biais et variance

Le pire scénario

Performance d'un joueur en début d'entraînement.

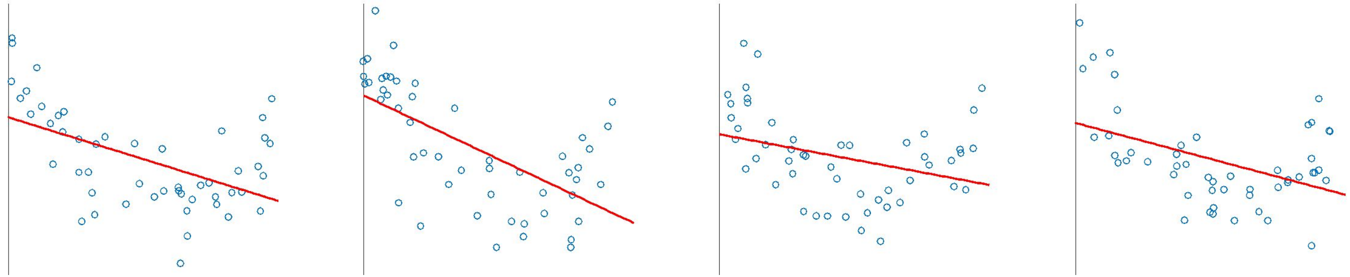


Le meilleur scénario

Performance d'un joueur bien entraîné.

Biais

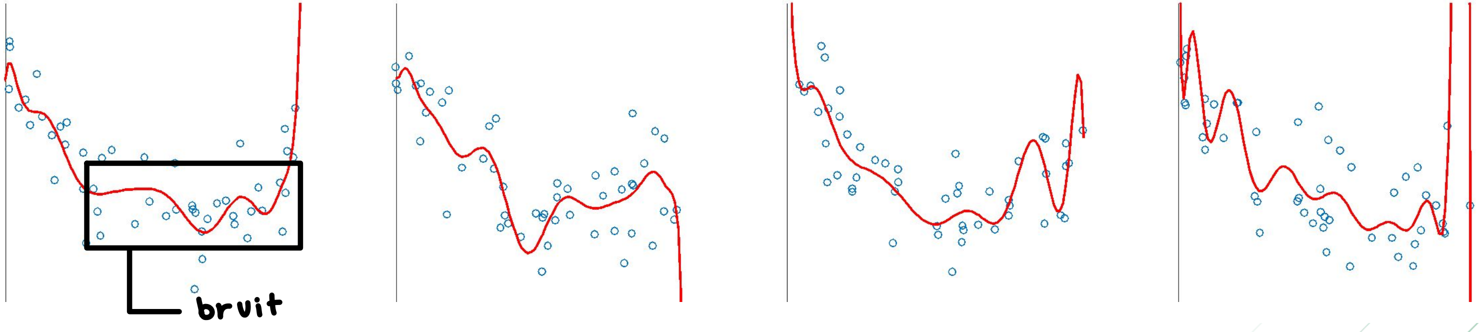
- On essaie d'entraîner un modèle simple h(x).
- Données bruitées provenant d'un signal f(x).
- Indépendamment de l'échantillon; le modèle h(x) produira des erreurs systématiques.



- La moyenne des courbes h(x) ne reproduit pas bien f(x): biais élevé.
- Il y a peu de variations entre chaque modèle: variance faible

Variance

- On essaie d'entraîner un modèle plus complexe h(x).
- Données bruitées provenant d'un signal f(x).
- Différents échantillons produisent différents ajustements de modèle avec de grandes oscillations peu réalistes.



- Il y a beaucoup de variations entre chaque modèle: variance élevée.

Bias and Variance (2019)

Le compromis biais-variance

nb données

Supposons que l'on utilise 2M différents jeux de N données chacun:

- $\{\mathbf{x}_i^t, y_i^t\}$, i=1,...,2M

générés à partir d'une fonction bruitée

- $y_i^t = f(\mathbf{x}^t) + e_i^t$, t=1,...,N,

ou f(x) est une fonction inconnue et *e* est un bruit de mesure.

N.B.: Les valeurs de $\mathbf{x}^t$ sont les mêmes pour chaques jeux.

Séparons les données en M ensembles d'entraînement et M ensembles de test.

Le compromis biais-variance

Pour chaque jeu de N données on entraîne un modèle $h_i(x)$.

La moyenne des Modèles est

$$m(x) = \frac{1}{M} \sum_{i=1}^{M} h_i(x)$$

Le bias est calculé à l'aide de la fonction:

$$b(x) = m(x) - f(x)$$

sur les données de test.

Idéalement il devrait être nul partout.

Le compromis biais-variance

On calcule les mesures suivantes avec le biais et les données de **test**.

- Biais^2(h): $$\frac{1}{M}\sum_{t=1}^{M}[b(\mathbf{x}^t)]^2$$
- Variance^2(h): $$\frac{1}{NM}\sum_{t=1}^{M}\sum_{i=1}^{M}[m(\mathbf{x}^t) - h_i(\mathbf{x}^t)]^2$$
- MSE(h): $$\frac{1}{MN}\sum_{t=1}^{M}\sum_{t=1}^{N}[y_i - h_i(\mathbf{x}^t)]^2$$

On peut montrer la relation suivante entre MSE (Moindre carrés moyen) et les autres mesures:

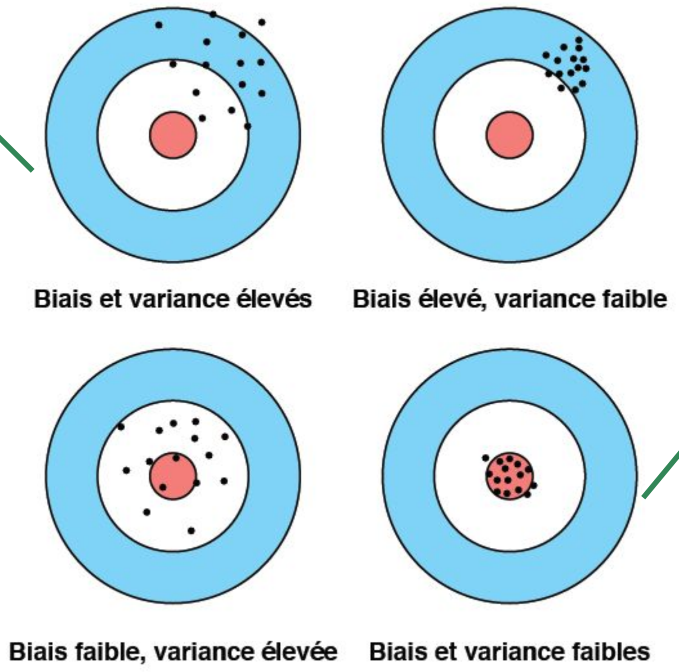
- **MSE(h) = Biais^2(h) + Variance^2(h) + σ^2**

Avec σ^2 étant la variance du bruit: Variance^2(e)

Exemples de distributions avec biais et variance

Le pire scénario

Performance d’un joueur en début d'entraînement.



Le meilleur scénario

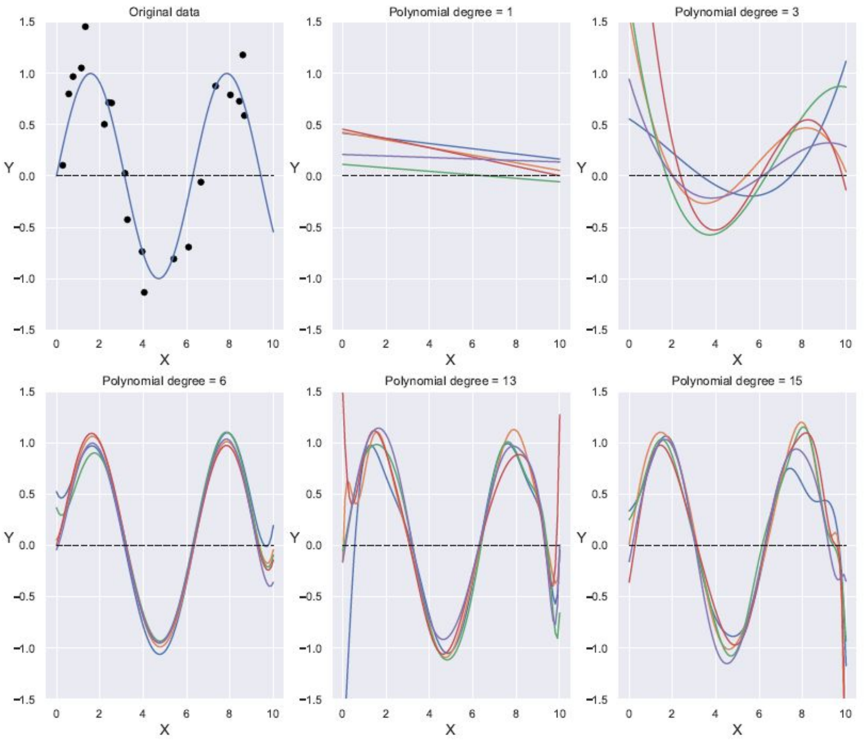
Performance d’un joueur bien entraîné.

Exemples avec polynômes $h_i(x)$ de degrés i divers

Exemple de signal bruité sinusoïdal.

- Cas degrés faibles :
  - Modèles $\mathbf{h}_i(\mathbf{x})$ très différents du signal idéal.
  - Biais important.
  - Variance faible.

- Cas degrés élevés :
  - Modèles $\mathbf{h}_i(\mathbf{x})$ très similaires au signal idéal.
  - Biais négligeable.
  - Variance importante aux extrémités.



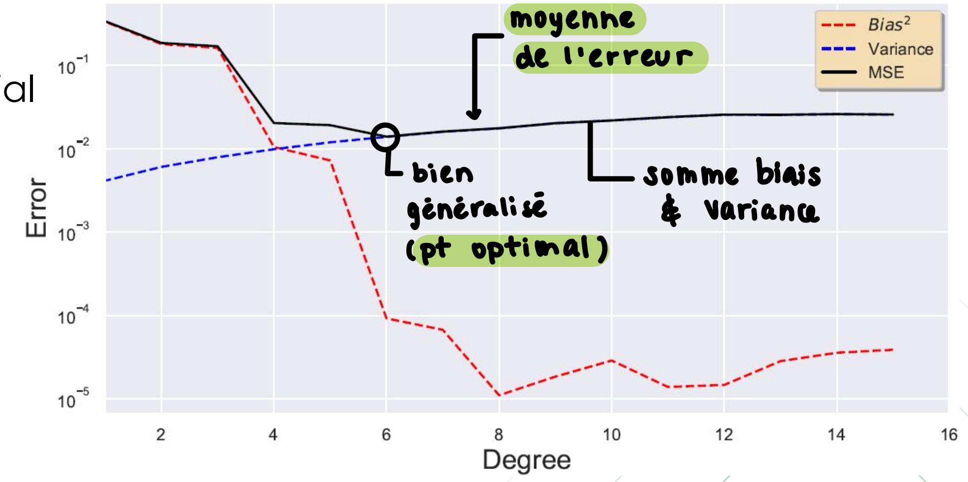
Exemples avec polynômes $h_i(x)$ de degrés *i* divers

- La figure montre comment varie l'erreur quadratique moyenne (MSE) en fonction du degré du polynôme d'approximation $\mathbf{h}_i(\mathbf{x})$.

- La MSE est dominée par le biais des modèles trop simples ou par la variance des modèles trop complexes.

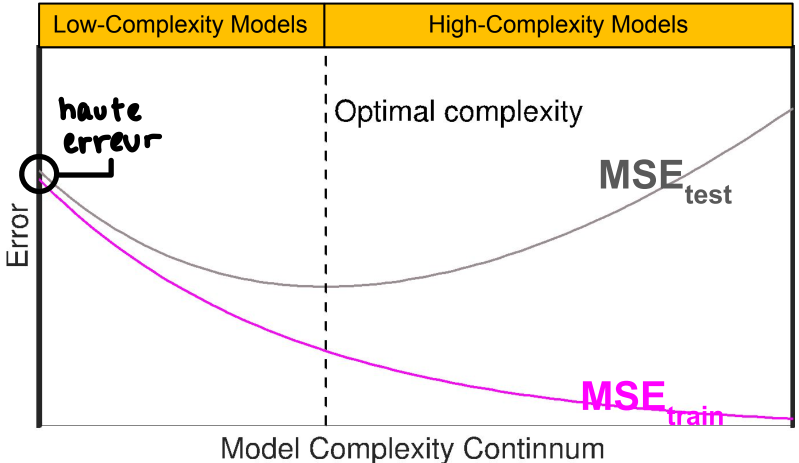
- La MSE est minimale pour le modèle polynomial de degré 6.

- On se sert de cette courbe pour trouver le meilleur modèle pour analyser nos données.



Erreur selon le compromis biais-variance

- On aurait pu utiliser un modèle approximatif $\mathrm{h}_i(\mathbf{x})$ différent, mais les résultats auraient été similaires.
- La figure suivante montre le comportement général.
- Le compromis biais-variance est révélé via un ensemble de test et non un ensemble



Problèmes de biais et de variance

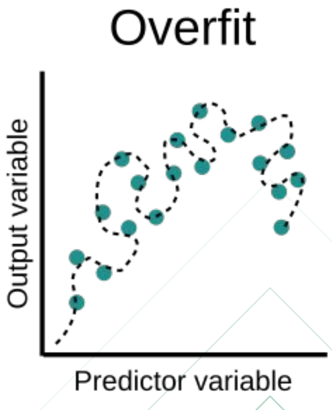
- **Biais**
  - Erreur systématique non nulle
  - Modèles trop simples
  - Révèle le sous-apprentissage
  - Mesure : moyenne des prédictions d'un modèle

- **Variance**
  - Résultats bons en moyenne mais trop variables en pratique
  - Modèles trop complexes
  - Révèle le surapprentissage
  - Mesure : variance des prédictions d'un modèle

Sur et sous-apprentissage

Qu’est ce que le sur apprentissage ?

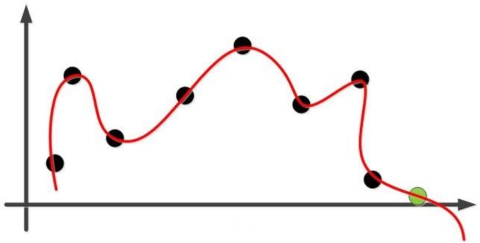
- Un modèle :
  - Apprend les détails et le bruit dans les données d’entraînement.
  - Produit de bons résultats sur les données d’entraînement, mais de mauvais résultats sur les données de validation et de test.
  - Mémorise les données au lieu d'apprendre et de comprendre la tendance sous-jacente des données.
  - N’est pas en mesure de généraliser avec de nouvelles données.
  - Biais faible mais variance élevée.

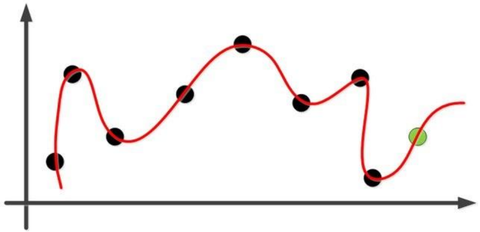


Le problème de la variance

- Le sur apprentissage est parfois appelé le problème de la variance.

- Un petit changement dans les données (point vert) affecte complètement le modèle ci-dessous.



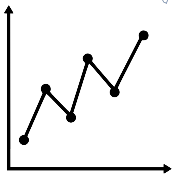


Causes

problème de surapprentis-sage

Solut° :
(1) ↓ variance
(2) ↑ taille données ($)
(↓ biais & ↓ variance)

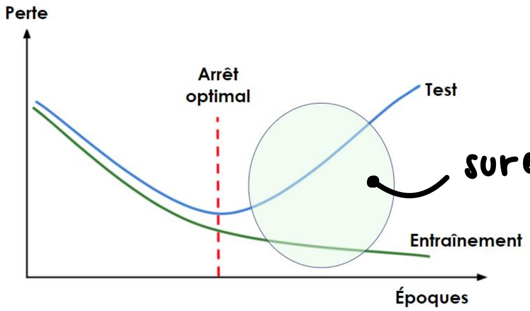
- Le modèle est trop complexe.



- Le modèle a une variance élevée.

- La taille du jeu de données d'entraînement n’est pas suffisante.

- Entraîné trop longtemps!

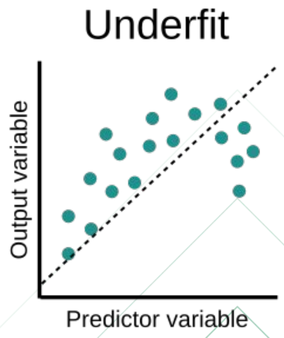


surentraînement

Qu’est ce que le sous apprentissage ?

- Un modèle trop simple
  - Ne performe pas bien avec les données d’entraînement et de test.
  - N'est pas en mesure de saisir la relation entre les exemples en entrée et les valeurs cibles.

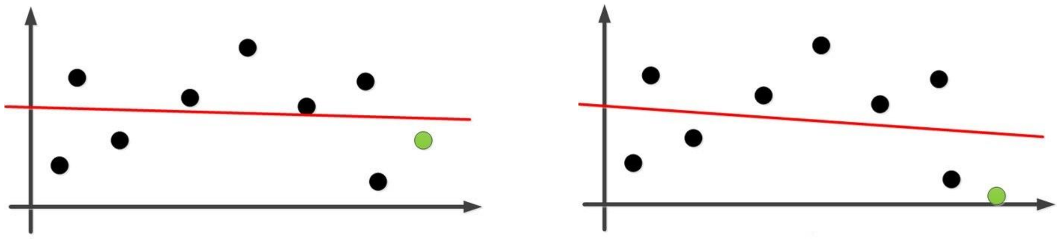
- Variance faible ✅ mais biais élevé ❌
  - Le modèle fait systématiquement des erreurs similaires aux mêmes endroits.



Variance faible, mais biais élevé

- Le modèle ci-dessous varie peu lorsque l’on déplace le point vert.

- Le modèle fait systématiquement des erreurs similaires aux mêmes endroits; il est donc biaisé.



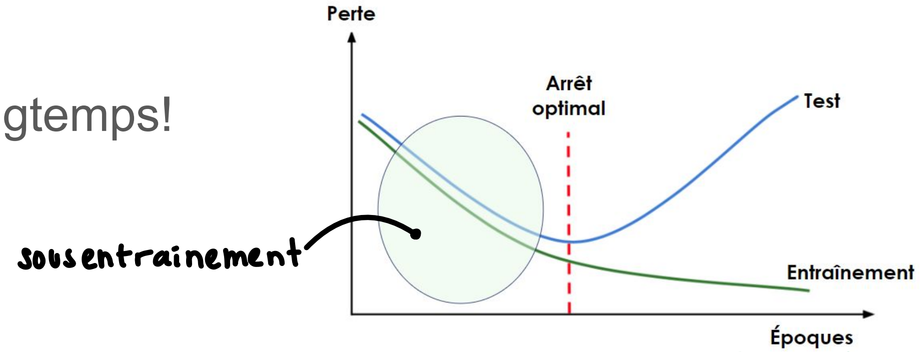
Causes

- Le modèle est trop simple.

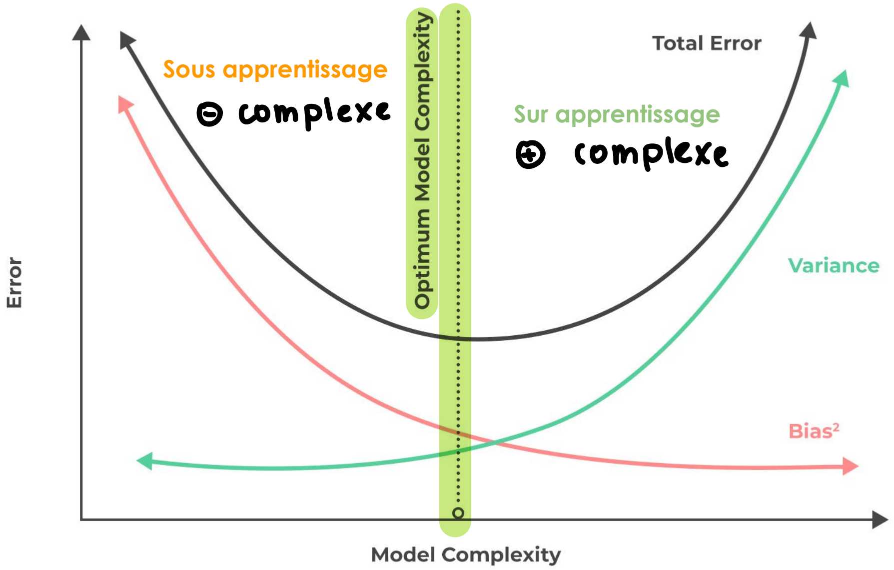
- Le modèle a un biais élevé.

- La taille du jeu de données d'entraînement n’est pas suffisante.

- Pas entraîné assez longtemps!



Compromis biais-variance

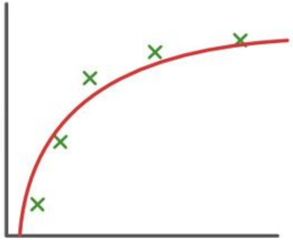


Un point optimal de complexité du modèle existe là où les courbes d'erreur de biais et de variance se croisent, et l'erreur minimisée.

Exemples en régression

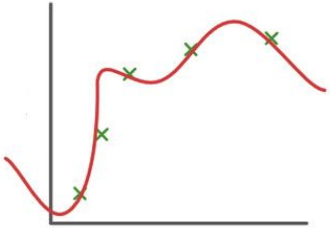
$$\hat{y}^{(t)} = \beta_0 + \beta_1 x^{(t)}$$

Sous apprentissage; modèle trop simple pour reproduire la distribution des données.



$$\hat{y}^{(t)} = \beta_0 + \beta_1 \overbrace{x^{(t)} + \beta_2 x^{(t)^2}}^{\text{entrées}}$$

Bon apprentissage: bon compromis biais-variance

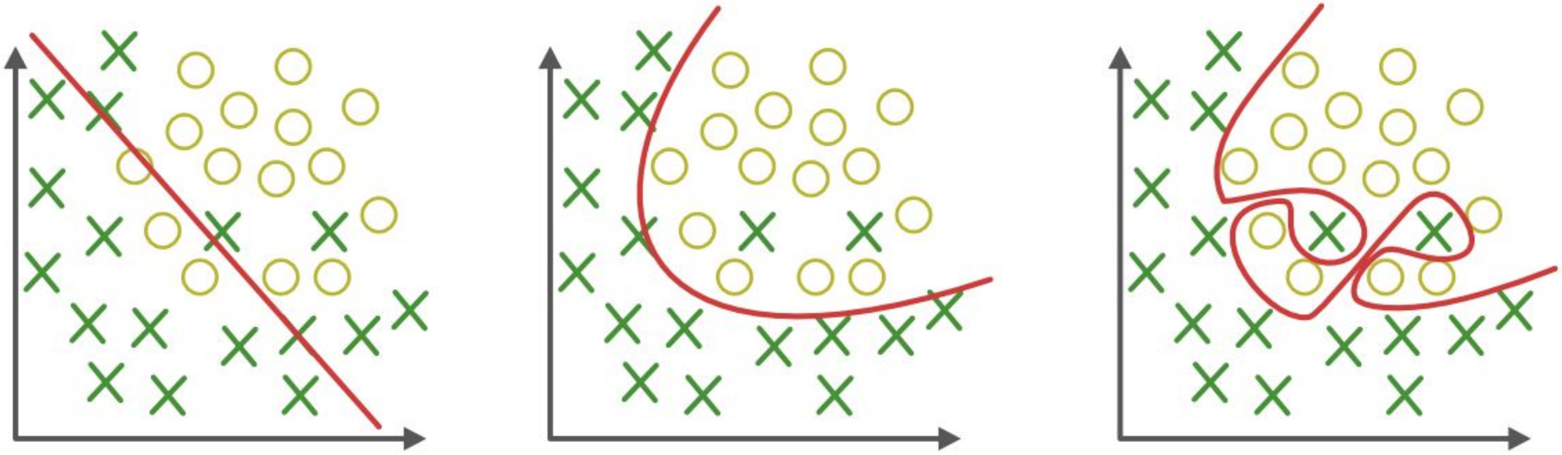


$$\hat{y}^{(t)} = \beta_0 + \beta_1 x^{(t)} + \beta_2 x^{(t)^2} + \beta_3 x^{(t)^3} + \beta_4 x^{(t)^4}$$

Sur apprentissage; modèle tellement flexible qu'il reproduit la distribution des données bruitées.

└ ⊕ caractéristiques → modèle ⊕ complexe

Exemples en classification



Sous apprentissage; modèle trop simple pour reproduire la distribution des données.

Bon apprentissage: bon compromis biais-variance

Sur apprentissage; modèle tellement flexible qu'il reproduit la distribution des données bruitées.

↳ ⊕ erreurs de généralisat°

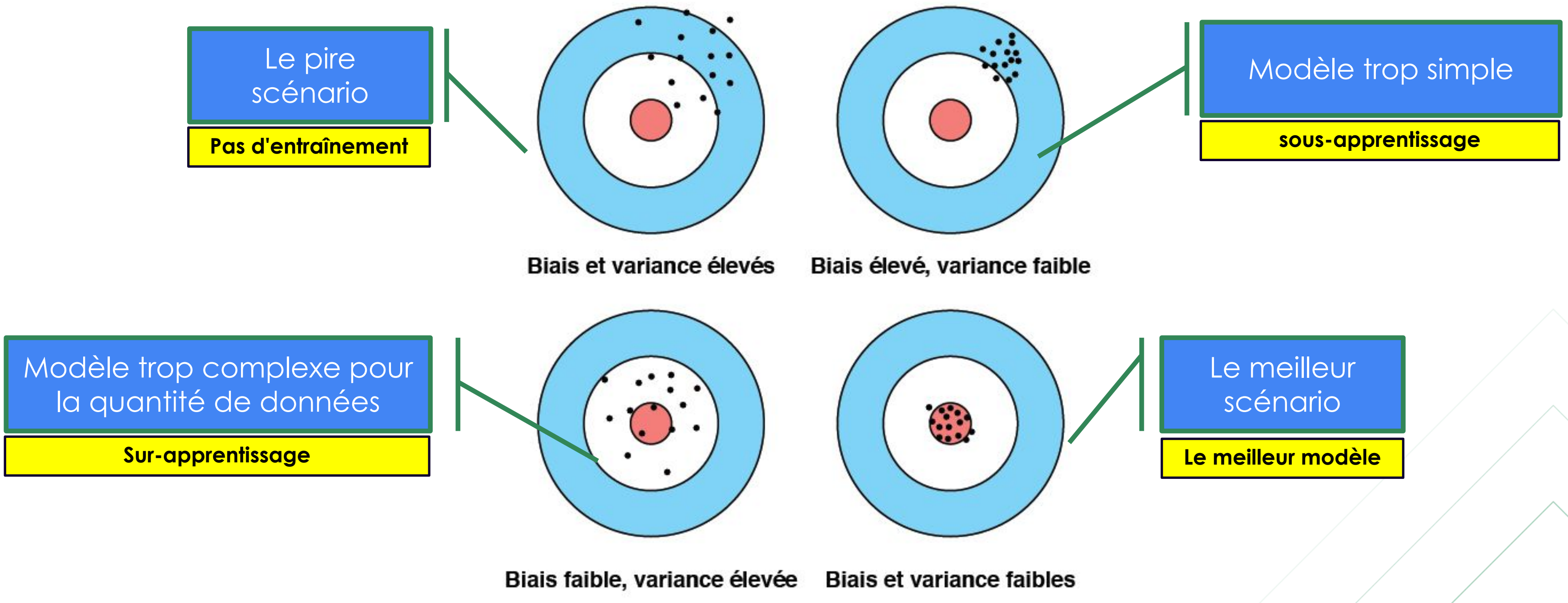
Solutions pour le sous apprentissage

- Ajouter des caractéristiques
- Ajouter des interactions entre les caractéristiques
- Complexifier le modèle
- Changer de modèle

Solutions pour le sur apprentissage

- Enlever des caractéristiques
- Réduction de dimensionnalité
- Régularisation
- Ajouter des données
- Nettoyer données (valeurs aberrantes)
- Simplifier le modèle
- Changer de modèle

Exemples de distributions avec biais et variance

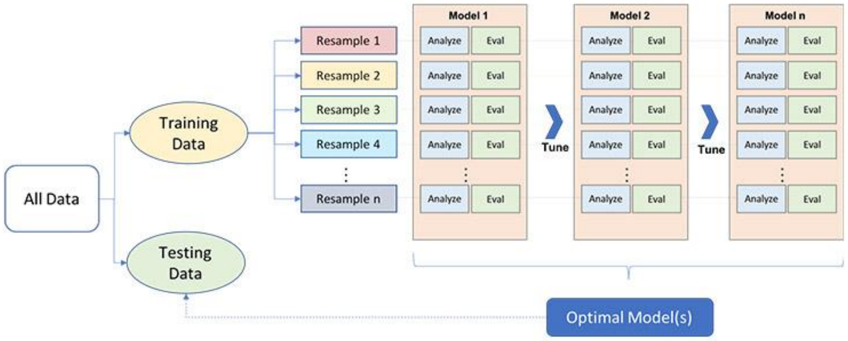


Stratégie utile pour trouver le modèle avec la complexité optimale

- On entraîne plusieurs modèles avec les **mêmes données d'entraînement.**

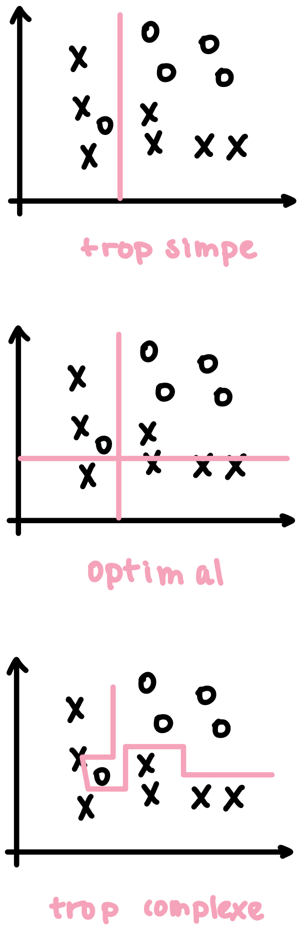
- Leurs performances en généralisation sont mesurées avec les **mêmes données de validation.**

- Le modèle avec les meilleures performances est celui avec la complexité optimale.



Sources

Arbre de décis°



MOOC d'IVADO