Master IA2S (Intelligence Artificielle, Science des données et Systèmes Cyber-Physiques)

Apprentissage Automatique

Ferhat ATTAL
ferhat.attal@u-pec.fr
Université Paris Est Créteil (UPEC)

Novembre 2020

Cours 2 : Généralités sur l'apprentissage automatique

Sommaire

Généralités sur l’apprentissage

Dans un cadre général, un apprentissage est l’acquisition de nouveaux savoirs (souvent sous forme de connaissances), c’est-à-dire le processus d’acquisition de compétences, par l’expérience ou l’enseignement. L’apprentissage est un terme utilisé dans plusieurs domaines (philosophie, linguistique, psychologie, biologie, . . .), mais chacun a sa propre définition de cette notion. Dans le domaine des sciences de l’ingénieur (informatique, automatique, robotique) et également dans de le domaine de l’IA Intelligence Artificielle) nous parlerons d’apprentissage automatique plutôt que d’apprentissage tout court.

- Croissance des quantités de données disponibles créées, stockées, manipulées,
- Importance de apprentissage automatique dans le progrès technologique.

Quelques domaines d’applications

Biologie

- Reconnaissance de séquences ADN.
- Aide au diagnostique des électrocardiogrammes (ECG).

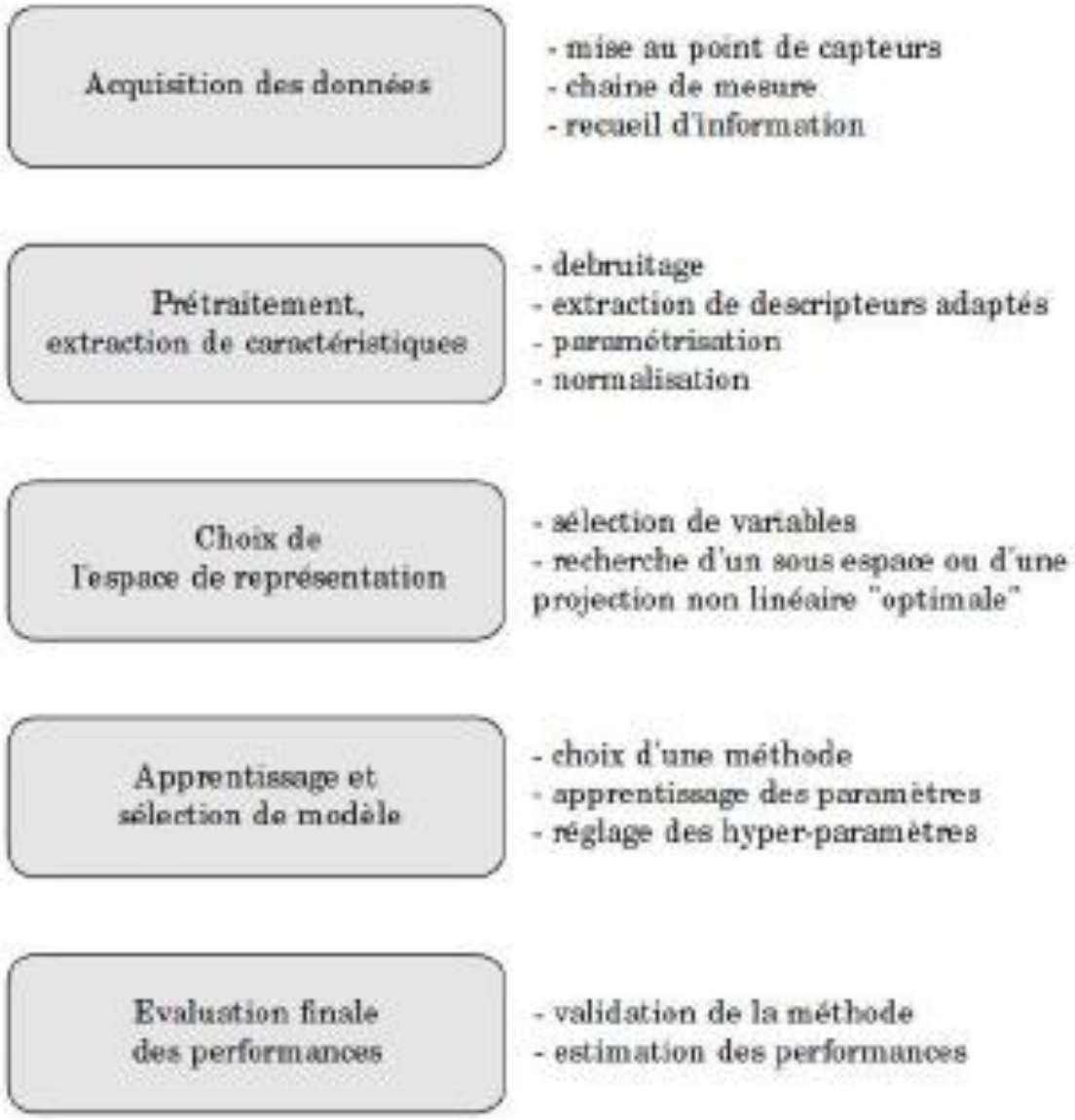
Finance

- Prédire si un investissement et bon ou pas en se basant sur les observations actuelles.
- Prédire avec précision le mouvement du cours des actions boursières.

Industrie

- Reconnaissance automatique de la parole.
- Reconnaissance de l’activité humaine.
- Reconnaissance de l’activité de conduite.

Différentes étapes du processus de classification automatique



Apprentissage supervisé

Définition: Dans la littérature, l'apprentissage supervisé est défini de plusieurs façons. Parmi ces définitions on trouve celle donnée par *Tom Mitchell* qui postule que l'apprentissage supervisé peut être caractérisé par trois éléments :

- La base d'apprentissage,
- La tache,
- La mesure de performance.

Apprentissage supervisé: éléments de l'apprentissage supervisé

L'apprentissage supervisé fait intervenir les éléments suivant :

- Des variables d'entrées ($X \in \mathcal{X}$) appelées aussi covariables appartenant à l'ensemble des réel de dimension d , $\mathbb{R}^d$.

- Des variables de sorties ($Y \in \mathcal{Y}$) appelées aussi variables d'intérêts appartenant à l'ensemble des réels $\mathbb{R}$. Les variables d'intérêts peuvent être continues (problème de régression) ou discrètes dites aussi catégorielles (problème de classification).

- Une fonction de prédiction notée $f_\theta(x)$ à paramètre θ appartenant à l'ensemble des paramètres Φ. Cette fonction a pour rôle d'associer des variables de l'espace d'entrée X avec des variables de l'espace de sortie Y.

- Un ensemble d'apprentissage $S = \{(x_i, y_i)\}_{i \leq n}$ constitué par l'ensemble de réalisation du couple (X, Y) = $\{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$ supposées indépendantes et identiquement distribuées (i.i.d) issues de la loi de probabilité jointe inconnue $p(x; y) = p(x)p(y|x)$.

- Un ensemble de test constitué par l'ensemble de nouvelles réalisation de X, et l'objectif est de prédire la variable $y = f_\theta(x)$. Une fonction de coût de l'erreur notée $\mathcal{L}(f_\theta(x), y)$ permettant de traduire l'erreur d'attribution de la variable d'entrée à sa sortie. La forme de cette fonction dépend essentiellement du problème étudié.

Apprentissage supervisé : Risque total et Risque empirique

Risque total moyen (théorique, réel)

$$R(f_\theta) = E[\mathcal{L}(y_i, f_\theta(x_i))] = \iint_{XY} \mathcal{L}(y, f_\theta(x))p(x,y)\,dy\,dx$$
$$= \iint_{XY} \mathcal{L}(y, f_\theta(x))p(x)p(y/x)\,dy\,dx$$

Risque empirique

$$R_e(F) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, f_\theta(x_i))$$

Apprentissage supervisé: Problème de classification

- y appartient à un ensemble fini $Y = \{c_1, c_2, \ldots, c_k\}$.
- Prédire l'appartenance d'un objet x à une classe donnée $\{c_l\}_{l \leq k}$

Dans ce cas de figure, la fonction de coût $\mathcal{L}$ prend souvent la forme suivante:

$$\mathcal{L}(y, f_\theta(x)) \triangleq \begin{cases} 1 \; si \; y \neq f_\theta(x) \\ \\ 0 \; si \; y = f_\theta(x) \end{cases}$$

Apprentissage supervisé : Problème de régression

- y est une variable continue qui appartient à l'ensemble des réels $\mathbb{R}$.
- Associer une valeur numérique à une nouvelle réalisation de la variable d'entrée x.

La fonction de coût $\mathcal{L}$ la plus utilisée est de la forme suivante :

$$\mathcal{L}(y, f_\theta(x)) = (y_i - f_\theta(x))^2$$

Apprentissage supervisé : Approches génératives

Dans l'approche générative, l'algorithme apprend un modèle à partir de la distribution conjointe $p(x, y)$.

- Modélisation des densités conditionnelles $p(x|y)$ pour tout $Y = \{c_1, c_2, \ldots, c_k\}$.

- Modélisation de la probabilité a priori $p(y)$.

- Calcul des probabilités a postériori de chaque classe en utilisant le théorème de Bayes :

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} = \frac{p(y)p(x|y)}{\sum p(y')p(x|y')}$$

Apprentissage supervisé: Approches discriminatives

Contrairement à l'approche générative, l'approche discriminative modélise directement la distribution conditionnelle $p(y|x)$.

A partir de cette distribution conditionnelle, nous pouvons faire des prédictions de y, pour toute nouvelle valeur de x, en utilisant la règle du maximum a posteriori (MAP) :

$$\hat{y} = \underset{y \in Y}{\operatorname{argmax}}\, p(y|x)$$

Apprentissage supervisé : Dilemme biais variance

L'objectif de l'apprentissage supervisé est de trouver un fonction de prédiction $\hat{f} \in \mathcal{H}$ avec les exemples d'apprentissage $S = \{(x_i, y_i)\}_{i \leq n}$ tel que la prédiction $\hat{y} = \hat{f}(x)$ soit proche de y

Risque empirique (données d'apprentissage)

$$\hat{R}_e(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{f}(x_i))$$

Risque moyen ou risque de généralisation (est défini en moyenne sur la distribution jointe de (X,Y))

$$R(\hat{f}) = E[\mathcal{L}(y, \hat{f}(x)]$$

L'objectif est de **minimiser le risque moyen**.

Apprentissage supervisé : Dilemme biais variance

Majoration du risque de généralisation

$$R(f_o) \leq R(\hat{f}) \leq R(f_o) + 2 \max_{h \in H} |R(h) - \hat{R}_e(h)|$$

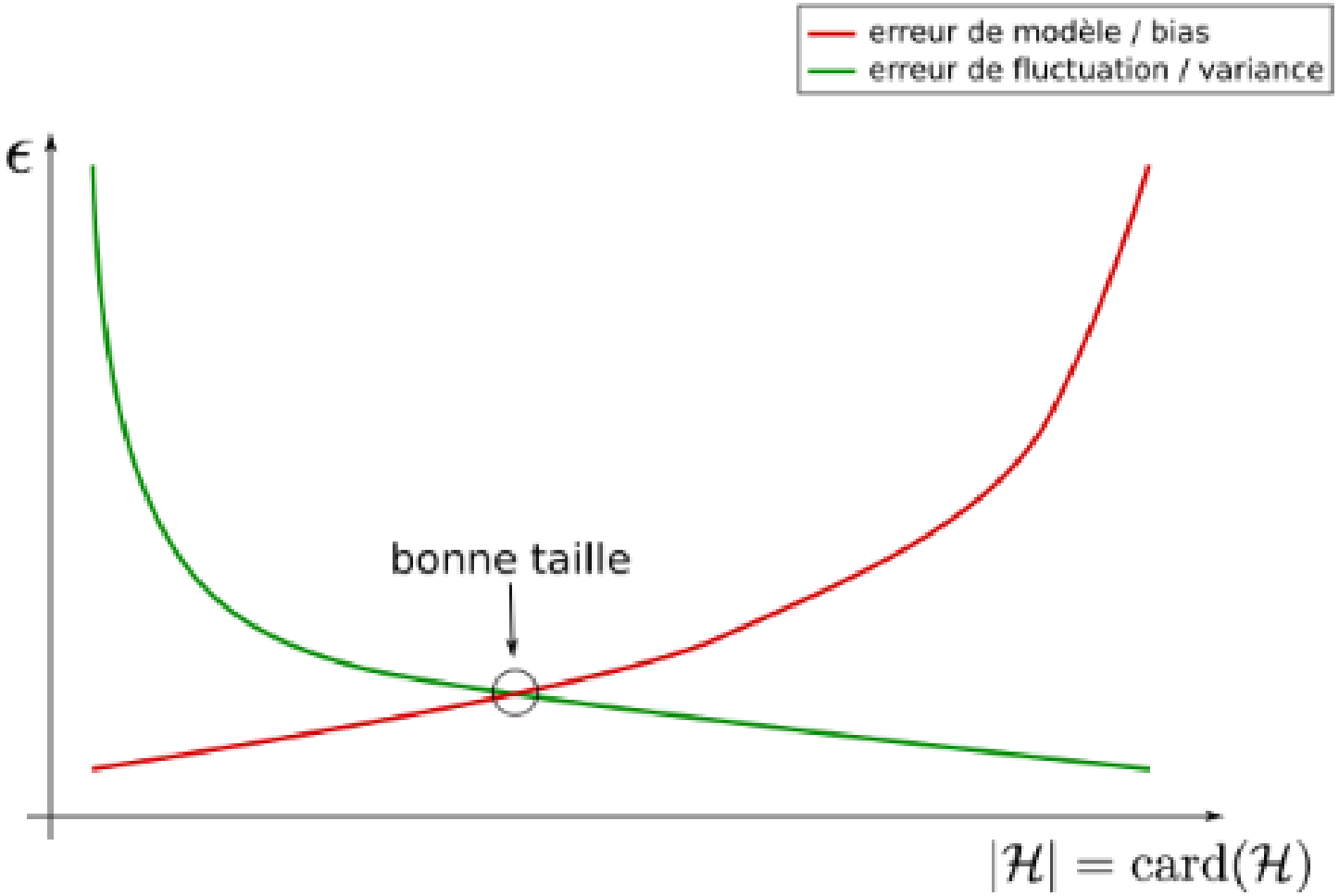
Avec

$f_o = \operatorname{argmax}_{h \in H} R(h)$: la "meilleure" approximation au sein de $\mathcal{H}$ qui minimise le risque généralisation.

$\hat{f} = \operatorname{argmax}_{h \in H} \hat{R}_e(h)$: la "meilleure" approximation au sein de $\mathcal{H}$ qui minimise le risque empirique.

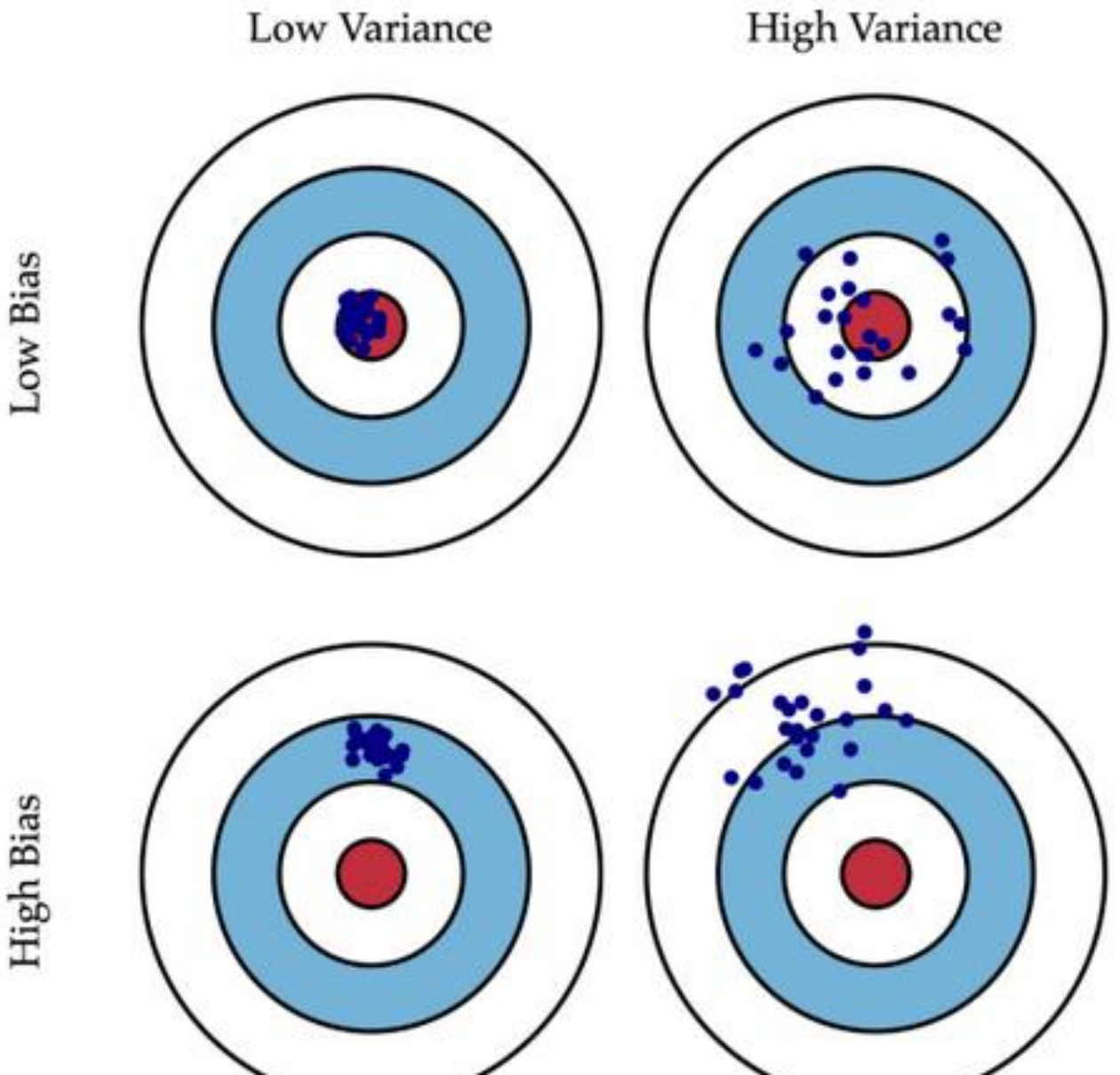
- $\boldsymbol{R(f_o)}$: l'erreur d'approximation qui représente l'erreur de généralisation minimale au sein de la classe de fonctions$\mathcal{H}$
- $\boldsymbol{\max_{h \in H} |R(h) - \hat{R}_e(h)|}$: erreur maximum de *fluctuation* entre le risque empirique et le risque moyen de n'importe quel prédicteur $h \in \mathcal{H}$.

Apprentissage supervisé : Dilemme biais variance



Dilemme biais-variance : le terme de biais décroit tandis que celui de fluctuation croit avec le cardinal de $\mathcal{H}$

Apprentissage supervisé : Dilemme biais variance



Apprentissage supervisé : Dilemme biais variance

Illustration dans le cas de la régression

Soit $S = \{(x_i, y_i)\}_{i \leq n}$ un ensemble d'apprentissage.

On suppose qu'il existe une relation fonctionnelle en entre les x_i et les y_i :

$$y_i = f(x_i) + \epsilon_i$$

Avec ϵ_i est un bruit de moyenne nulle et d'une variance σ^2

L'objectif est d'estimer la fonction $\hat{f}$ qui minimise le risque moyenne (risque de généralisation)

Dans ce cas, la forme du la fonction de cout est :

$$\mathcal{L}(y_i, \hat{f}(x_i)) = (y_i - \hat{f}(x_i))^2$$

Apprentissage supervisé : Dilemme biais variance

Ainsi le risque empirique est:

$$\hat{R}_e(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2$$

Le risque moyen est:

$$R(\hat{f}) = E[\mathcal{L}(y, \hat{f}(x)] = \text{Biais } [\hat{f}]^2 + \text{Var}[\hat{f}] + \sigma^2$$

Avec

$$\text{Biais } [\hat{f}] = \text{E } [\hat{f}] - f$$

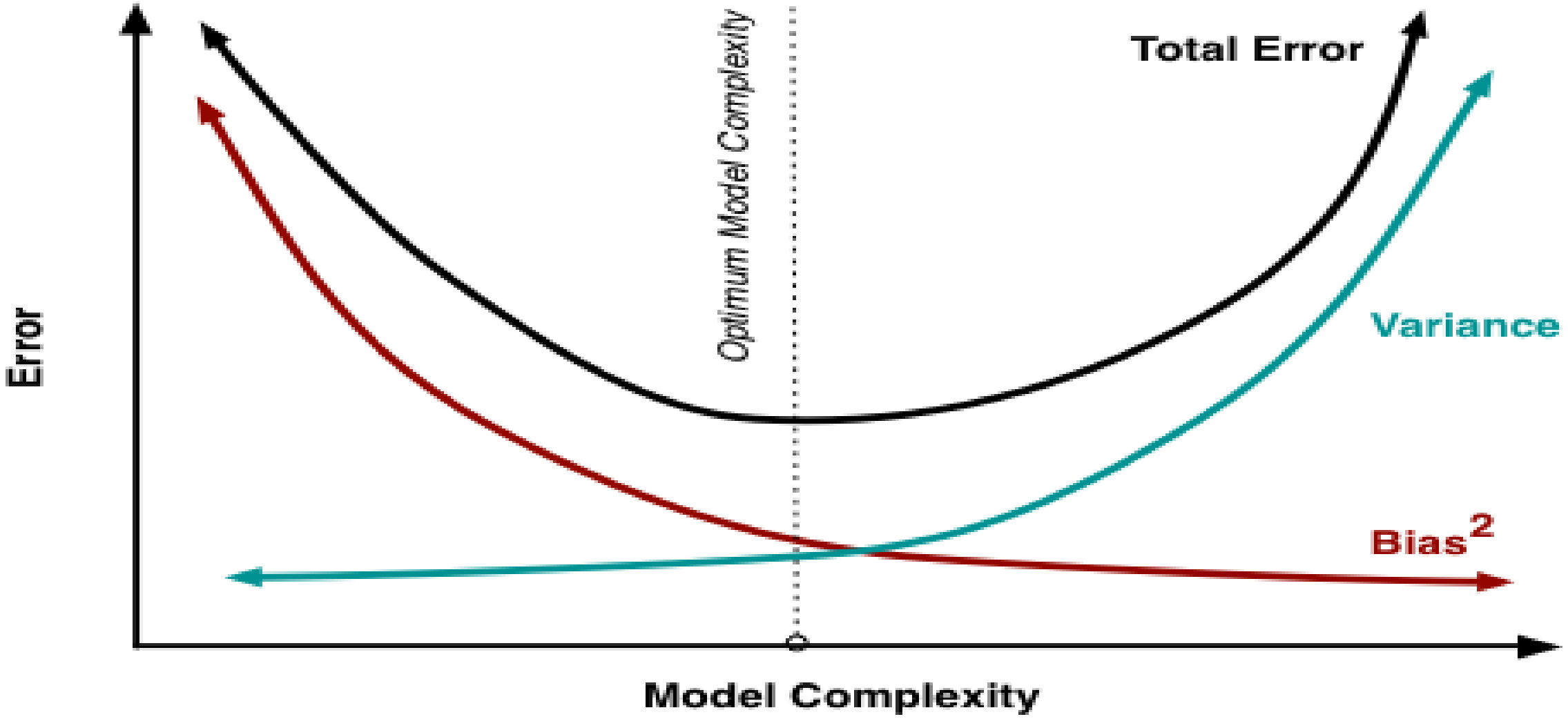
$$\text{Var}[\hat{f}] = \text{E } [\hat{f} - \text{E}[\hat{f}])^{\,2}]$$

Apprentissage supervisé : Dilemme biais variance

Démonstration

Apprentissage supervisé : Dilemme biais variance

$$R(\hat{f}) = E[\mathcal{L}(y, \hat{f}(x)] = \text{Biais } [\hat{f}]^2 + \text{Var}[\hat{f}] + \sigma^2$$



Apprentissage supervisé : Dilemme biais variance

Modèle complexe variance ↗ Biais ↘

Modèle moins complexe Biais ↗ variance ↘

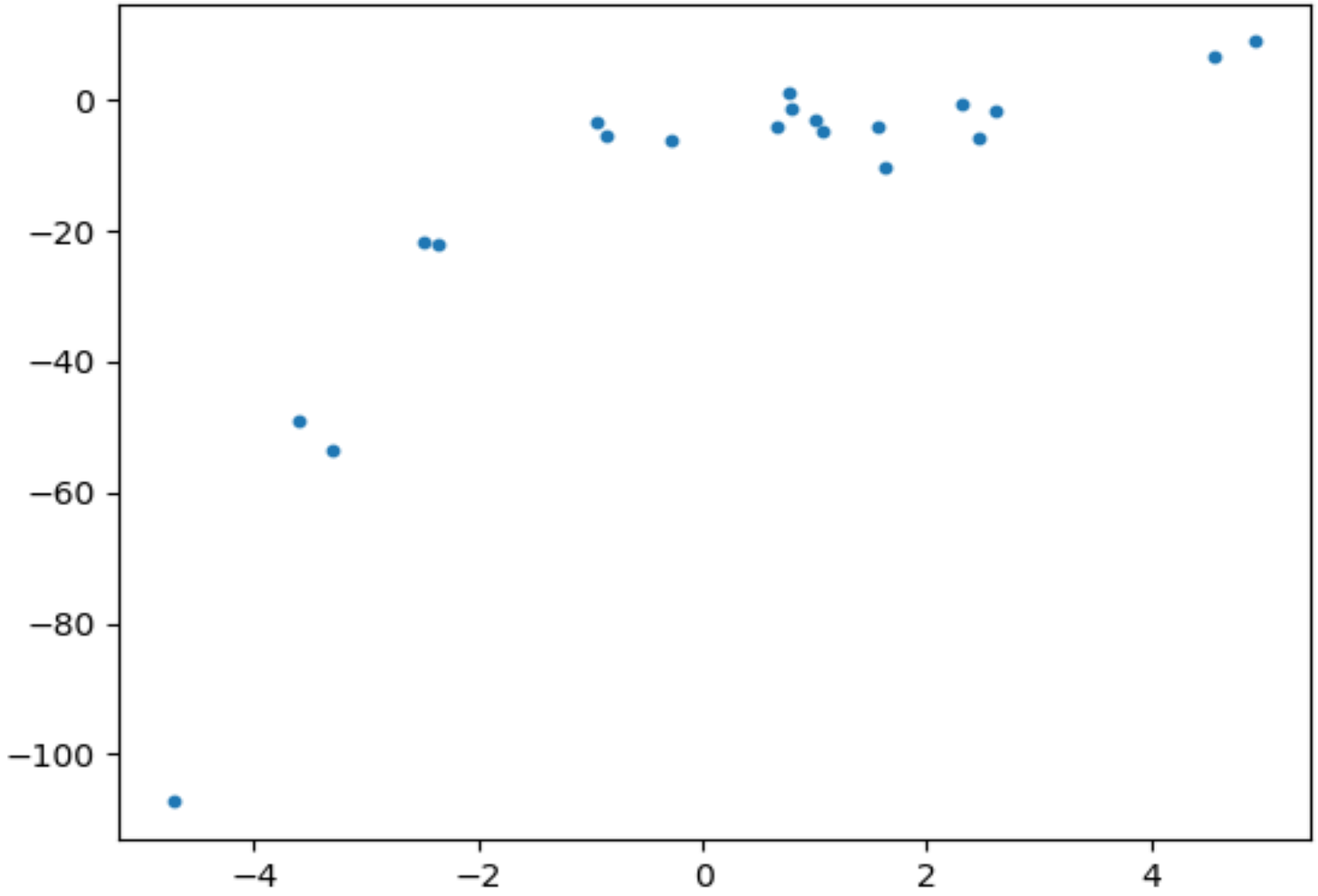
Complexité ↗ variance ↗ Biais ↘

Objectif : on cherche un modèle qui permet de minimiser le biais et la variance.

Le biais et la variance sont antagonistes ⇒ il faut trouver un compromis entre le biais et la variance.

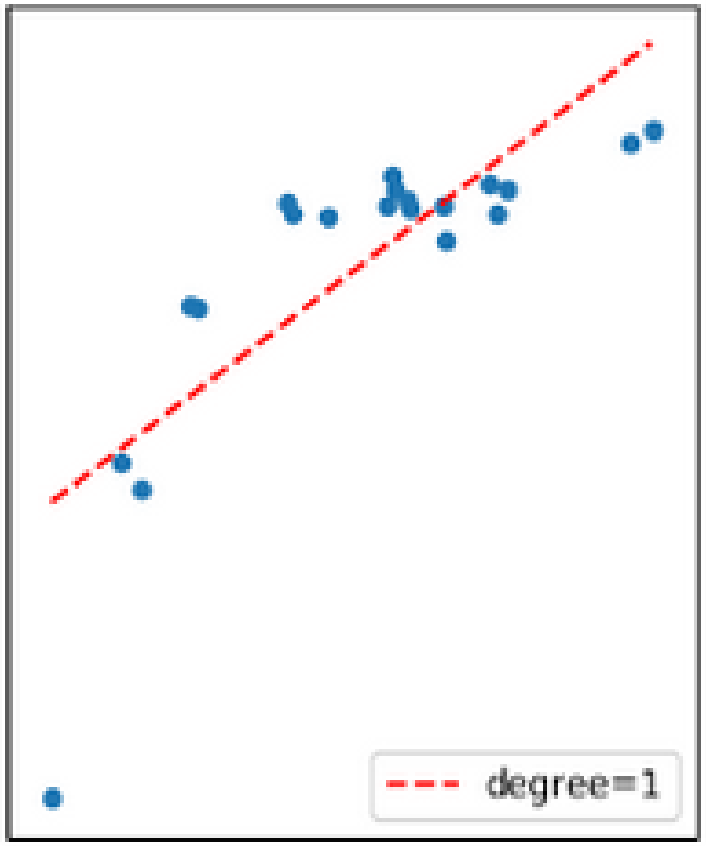
Apprentissage supervisé : Dilemme biais variance

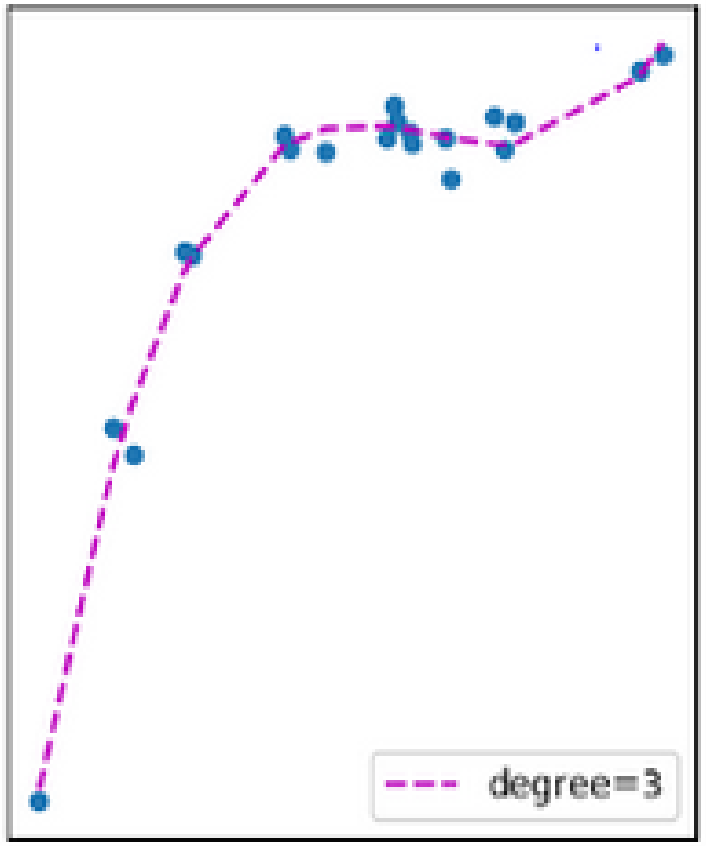
Illustration dans le cas de la régression

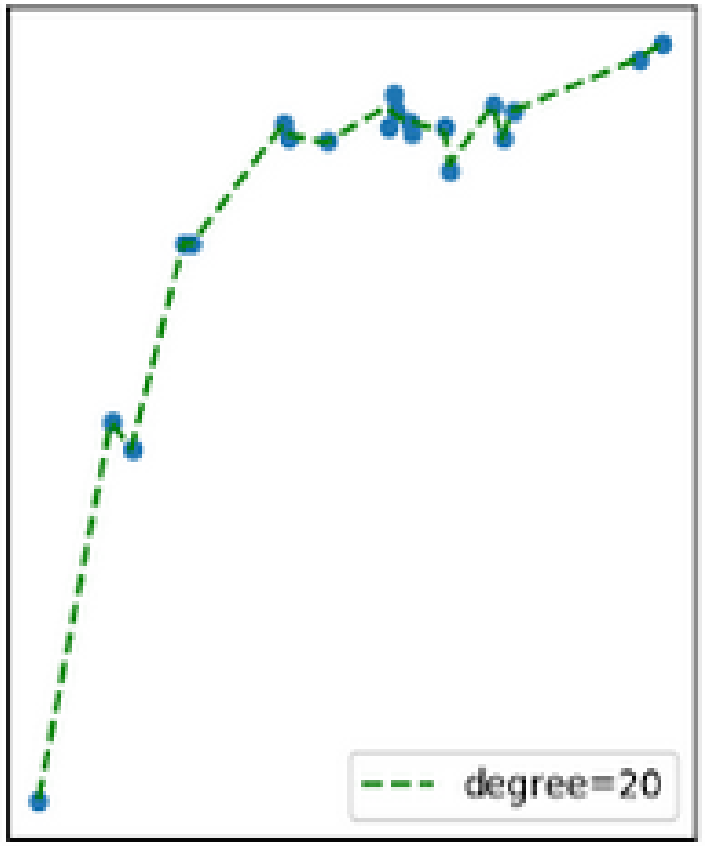


Apprentissage supervisé : Dilemme biais variance

Minimiser le risque empirique (données d'apprentissage, points bleus)

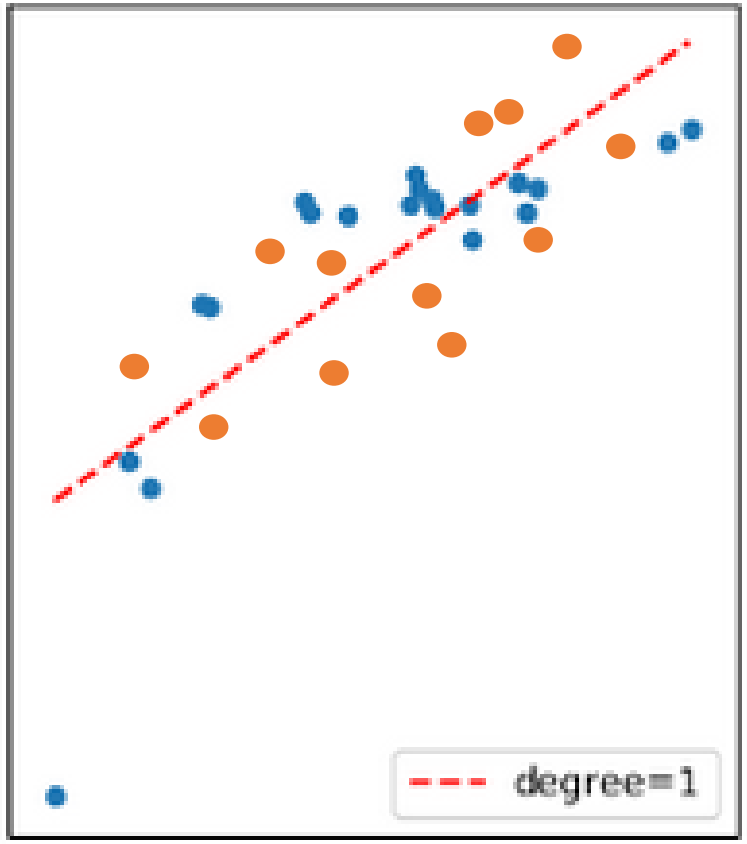


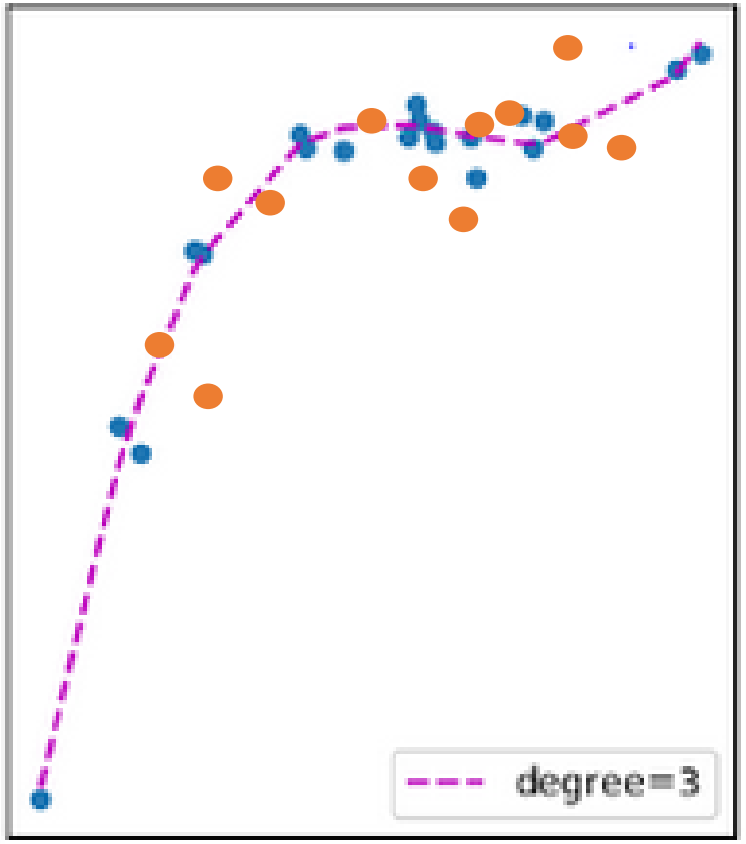


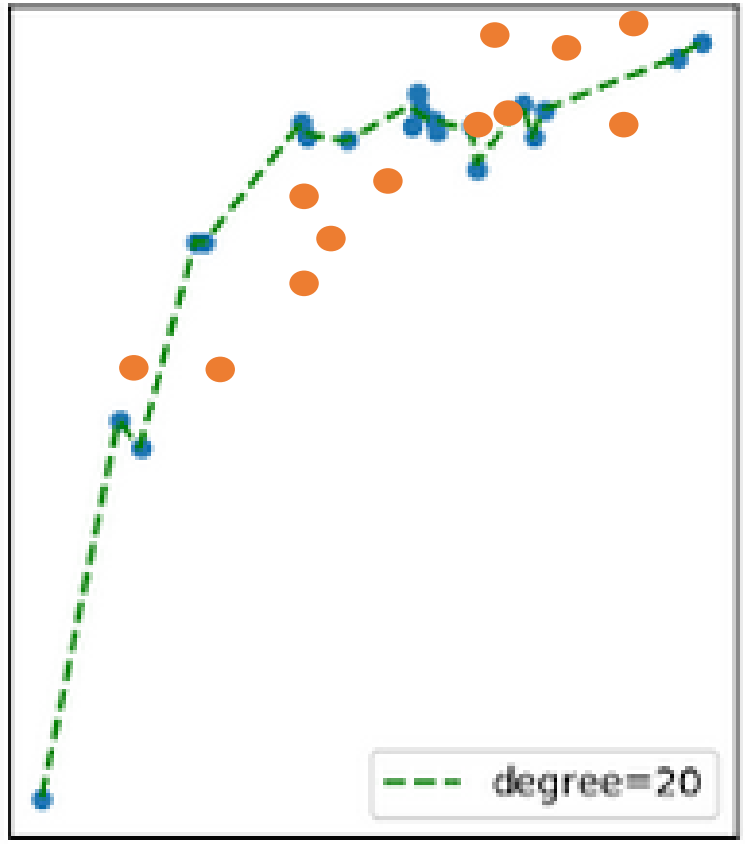


Apprentissage supervisé : Dilemme biais variance

Estimation du risque généralisation (données de test, points oranges)

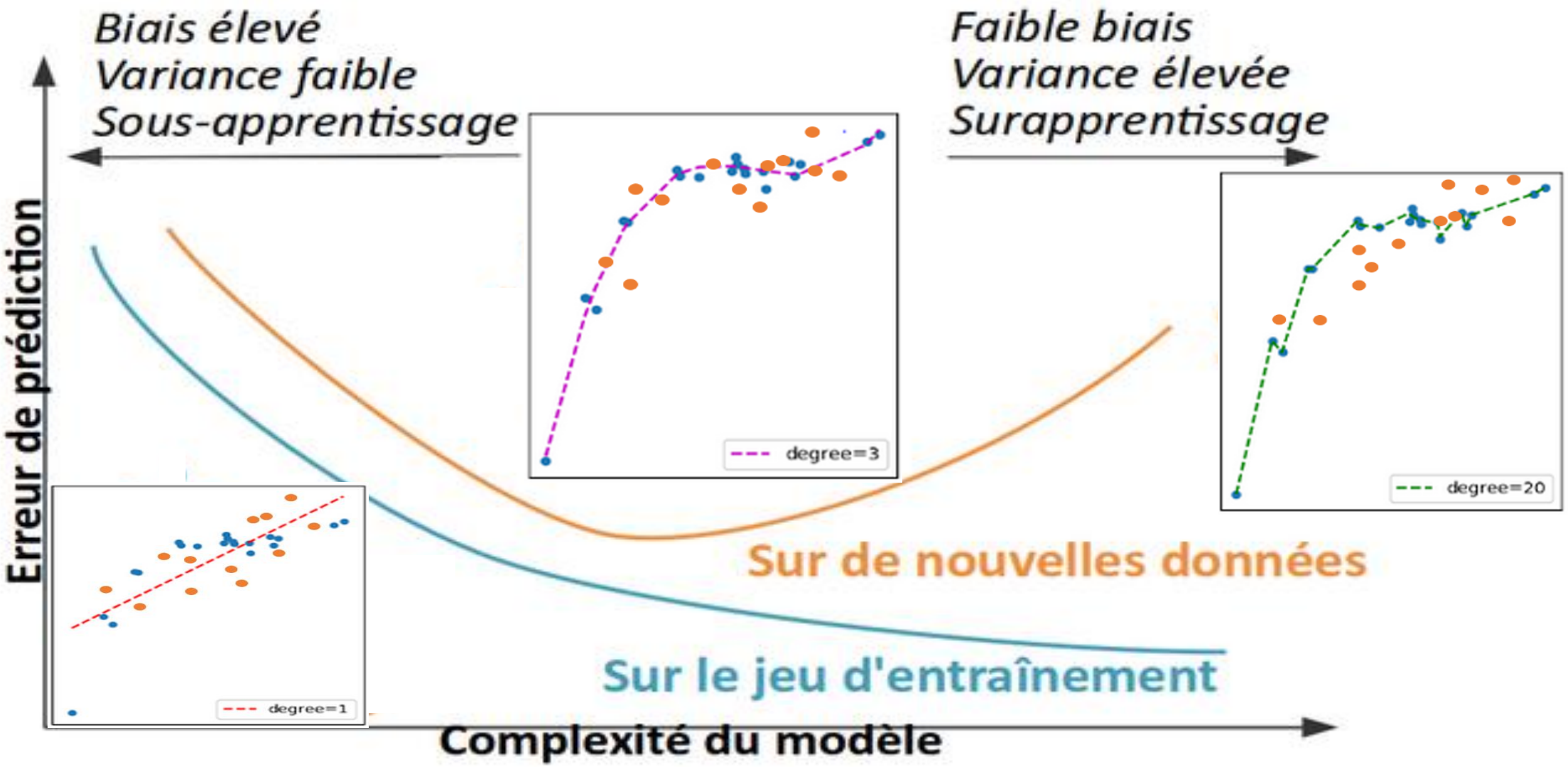






Apprentissage supervisé : Dilemme biais variance

Estimation du risque généralisation (données de test, points oranges)



Apprentissage supervisé : Dilemme biais variance

Estimation du risque de généralisation par les méthodes de re-échantillonnage

La méthode de validation est une méthode de vérification a posteriori des performances d'un modèle dont le principe est simple mais efficace (après avoir construit le modèle, on estime ses performances de généralisation). Il est crucial de mesurer l'erreur de généralisation sur des exemples qui n'ont pas servi lors de la construction du modèle. pour cela, on divise l'ensemble des données disponibles en
deux parties :

- un sous ensemble d'entrainement (**S**) dont les données serviront à l'apprentissage (ou construction) du modèle ;
- un sous ensemble de validation (**V**) dont les données seront utilisées uniquement pour évaluer la performance du modèle entrainé.

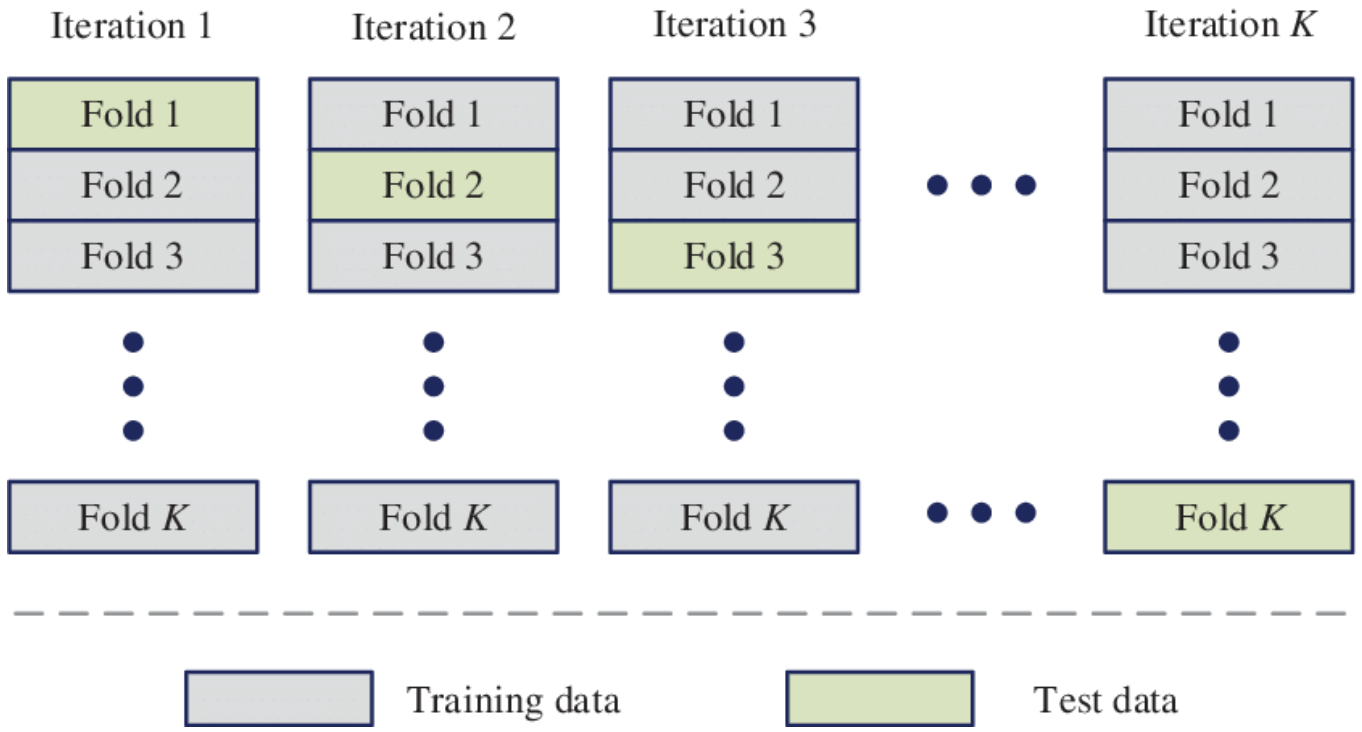
A partir d'un ensemble d'échantillons de taille N, il existe de nombreuses méthodes (appelées techniques de ré-échantilonnage) pour estimer la qualité de l'apprentissage. Ces méthodes sont le plus souvent appliquées avec l'hypothèse que la base de donnée utilisée est constituée de réalisations i.i.d de $p(\boldsymbol{x}, y)$.

- **Validation simple ou Hold out** : Cette technique de validation consiste à diviser les données disponibles en deux ensembles (les ensembles d'apprentissage (**S**) et de validation (**V**)) sans qu'aucune donnée ne soit commune. Un nombre assez conséquent de données est nécessaire dans l'ensemble de validation pour estimer l'erreur de généralisation avec une bonne précision, réduisant d'autant le nombre de données disponibles pour l'apprentissage. Souvent on garde 2/3des données dans l'ensemble d'apprentissage pour en réserver 1/3 pour la validation.

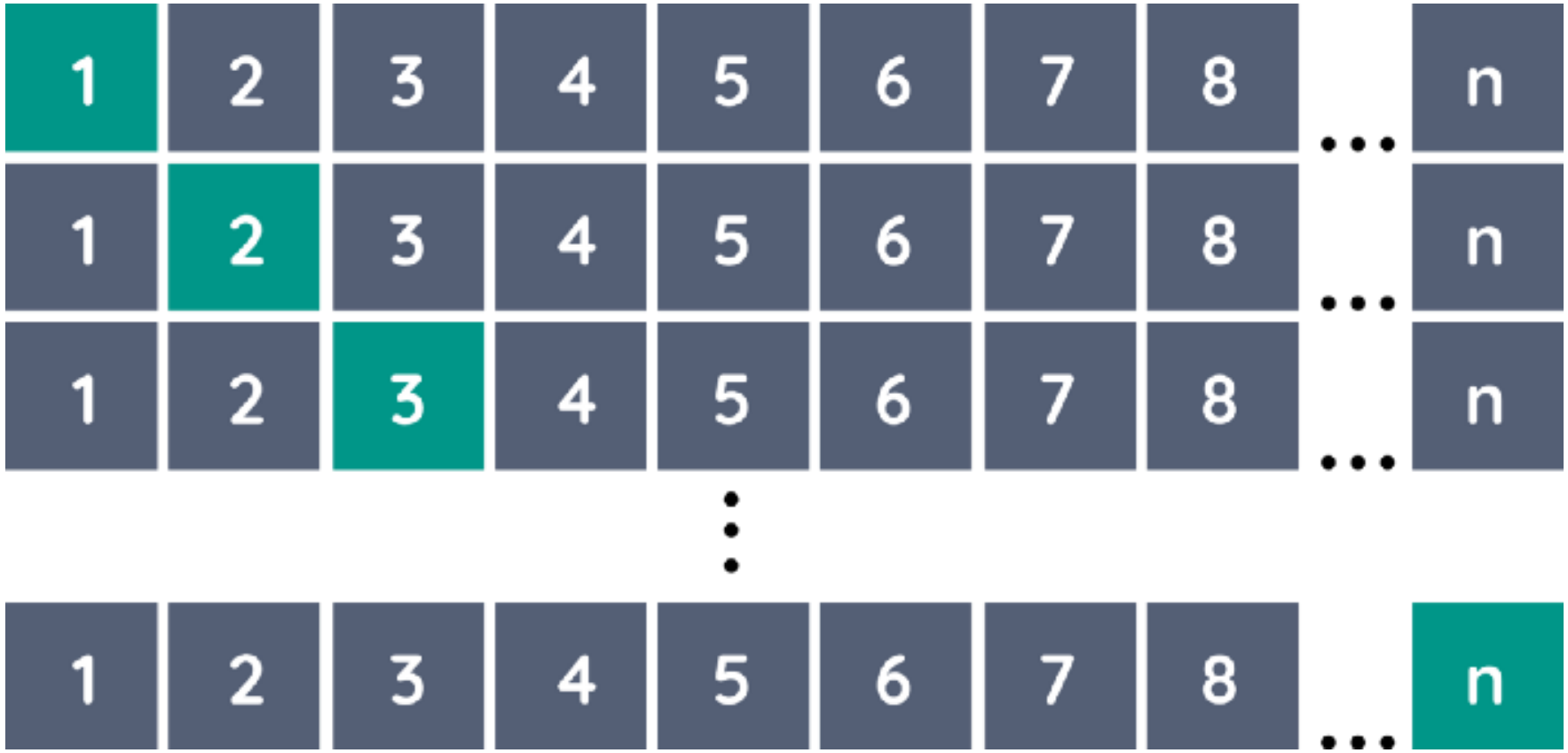
Pour éviter autant que possible tout problème lié à une dérive des données (et donc à la remise en cause de leur caractère i.i.d.), les ensembles d'apprentissage et de validation sont tirés aléatoirement de façon exhaustive parmi les *n* données disponibles.

Apprentissage supervisé

K-validation croisée (*K-fold cross validation*): l'échantillon initial est partitionné en K sous-ensembles disjoints de tailles approximativement identiques $\frac{n}{K}$. On utilise tour à tour chacun de ces sous-ensembles comme jeu de validation alors que les $(K-1)$ autres servent pour l'apprentissage. Le modèle optimal est alors défini comme étant celui qui présente le meilleur score (erreur de généralisation minimum) de validation croisée.



Leave One Out : cette méthode est un cas particulier de la validation croisée pour lequel K = n . Cette technique nécessite de relancer n fois la méthode de classification sur (n − 1) échantillons la rendant très prohibitive en temps de calcul



Apprentissage supervisé : Dilemme biais variance

Estimation du risque de généralisation par pénalisation

Utilisation d'une expression analytique de l'erreur de généralisation, incluant deux termes :

- Un terme de lie à l'erreur empirique (d'apprentissage) et
- Un terme de pénalisation reflétant la complexité du modèle et la dimension de l'espace des données.

$$\hat{R}(\hat{f}) = \hat{R}_e(\hat{f}) + Mcomp(\hat{f})$$

$Mcomp(\hat{f})$ c'est un terme de pénalisation qui est une mesure de complexité de $(\hat{f})$

Apprentissage supervisé : Dilemme biais variance

Exemple de pénalisation

Cp de Mallows (cas de régression):

$$\hat{R}(\hat{f}) = \hat{R}_e(\hat{f}) + \frac{d}{n}\hat{\sigma}^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2 + 2\,\frac{d}{n}\hat{\sigma}^2$$

$\hat{\sigma}^2$: une estimation de la variance.

Apprentissage supervisé : Dilemme biais variance

Exemple de pénalisation

Cas ou $\hat{f}_\theta$ est une densité, θ paramètres de la densité .

$$\hat{R}_e(\hat{f}) = -\frac{1}{n}\sum_{i=1}^{n} log\left(\hat{f}_\theta(x_i)\right) = -\frac{1}{n} log \prod_{i=1}^{n} \hat{f}_\theta(x_i)$$

Critère d'information bayésien

$$\begin{aligned}\hat{R}(\hat{f}) &= \hat{R}_e(\hat{f}) + Mcomp(\hat{f}) \\ &= -\frac{1}{n} log \prod_{i=1}^{n} \hat{f}_\theta(x_i) + \frac{d}{2n} log(n)\end{aligned}$$

Critère d'information d'Akaike

$$\begin{aligned}\hat{R}(\hat{f}) &= \hat{R}_e(\hat{f}) + Mcomp(\hat{f}) \\ &= -\frac{1}{n} ln \prod_{i=1}^{n} \hat{f}_\theta(x_i) + \frac{d}{2n}\end{aligned}$$

Apprentissage supervisé : Dilemme biais variance

Estimation du risque de généralisation par régularisation

Dans le cas de régression polynomiale

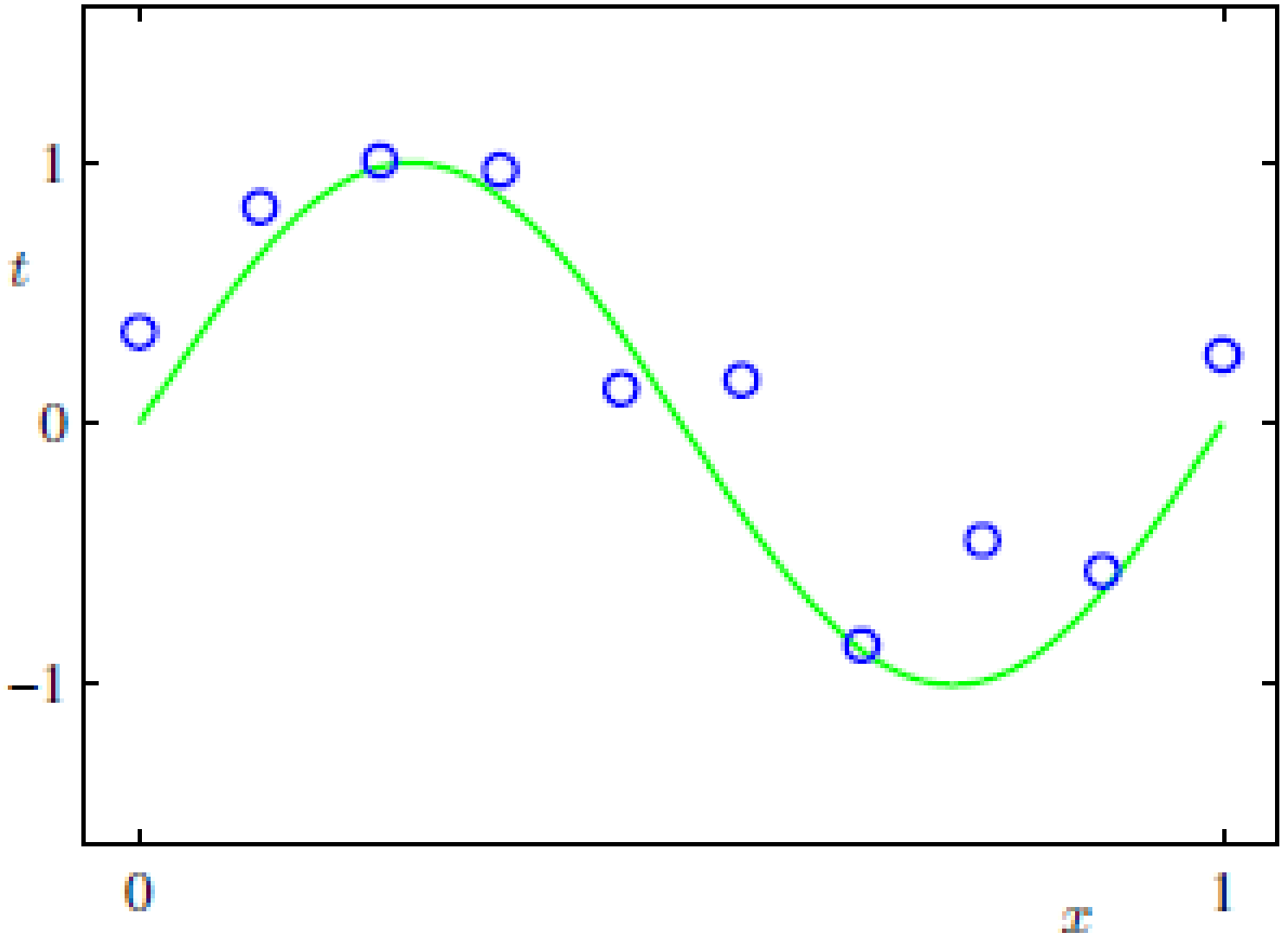
$$y = f(x,w) + \epsilon$$
$$= w_0 + w_1 x^1 + w_2 x^2 + \cdots + w_p x^P + \epsilon$$

Le risque de généralisation estimé s' écrit :

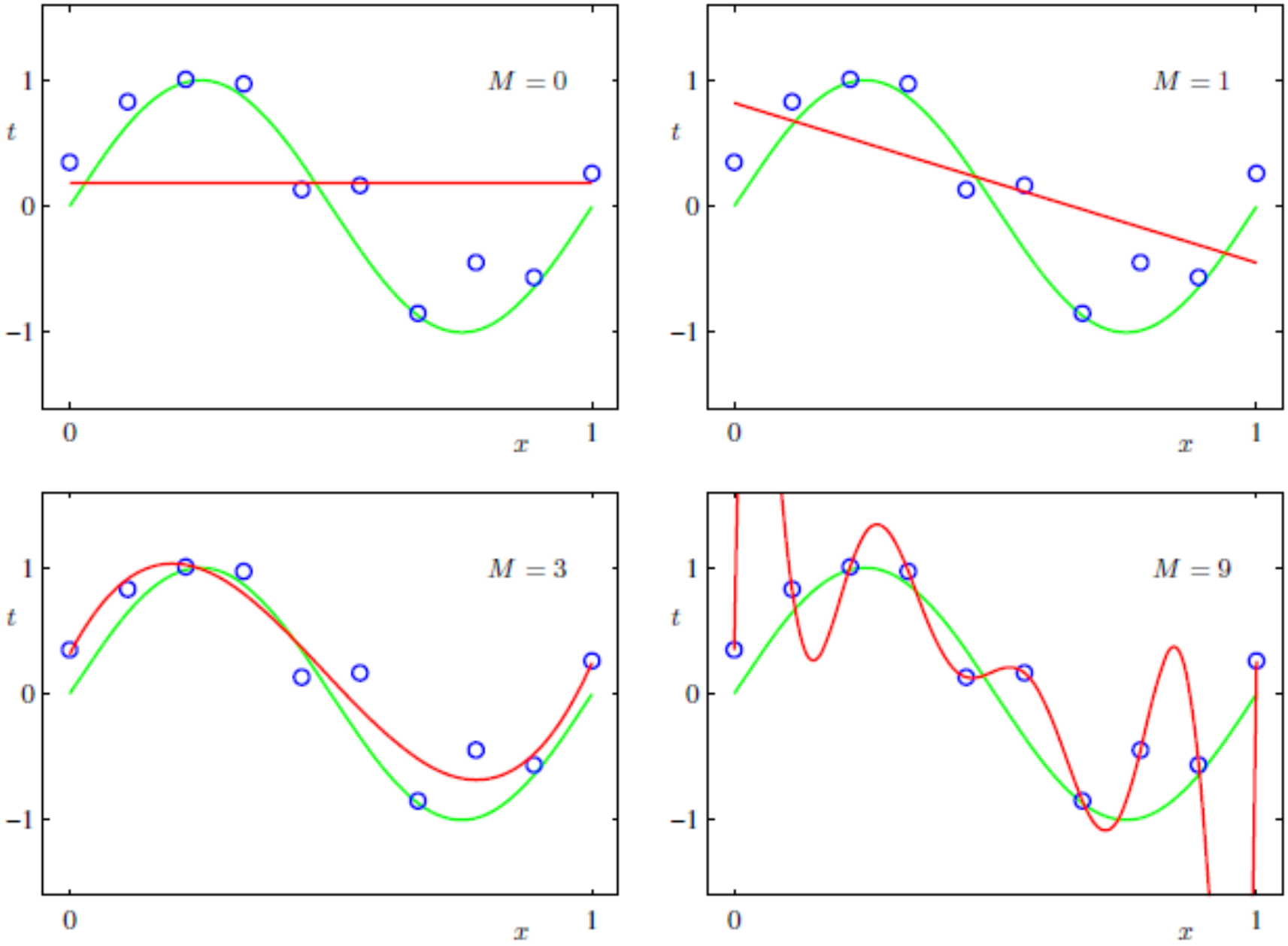
$$\hat{R}(\hat{f}) = \hat{R}_e(\hat{f}) + \lambda \, ||w||^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i, w) - y_i)^2 + \lambda \, ||w||^2$$

λ paramètre de régularisation
P l'ordre du polynôme

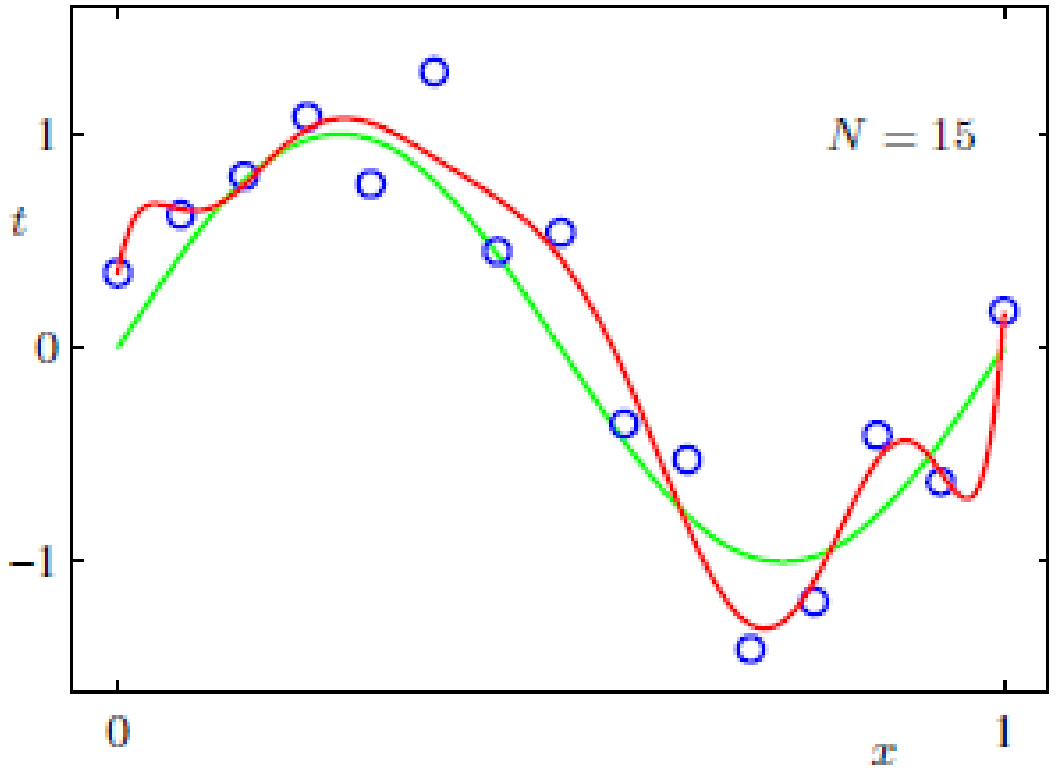
Apprentissage supervisé : Dilemme biais variance

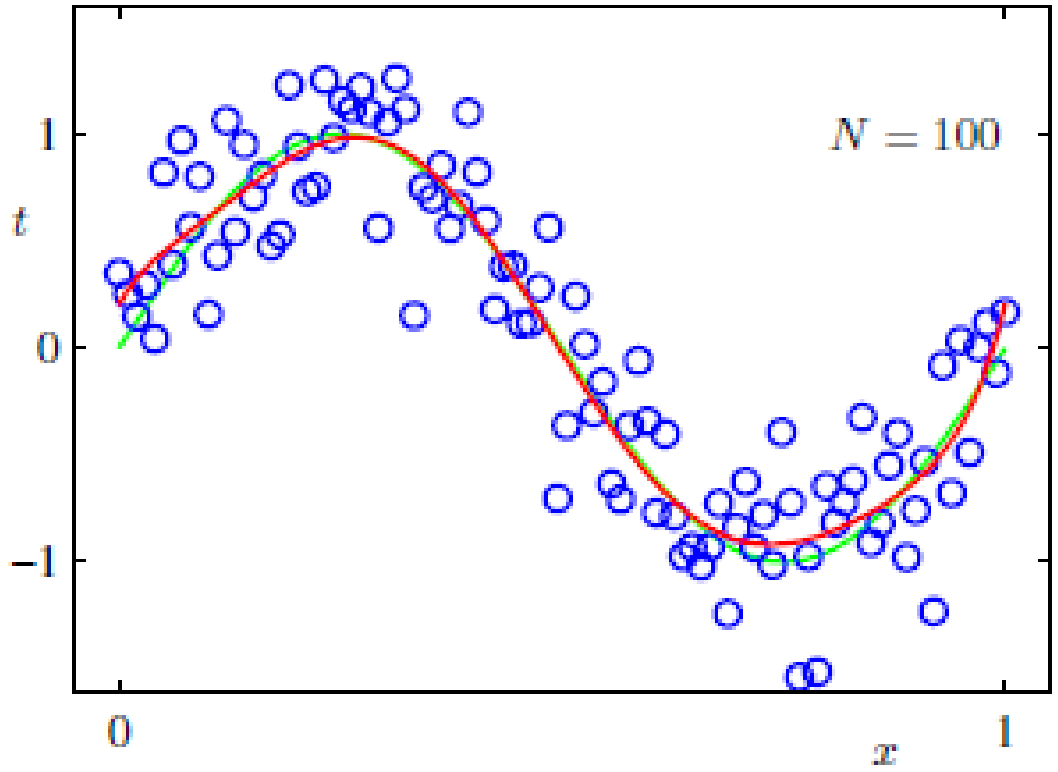


Apprentissage supervisé

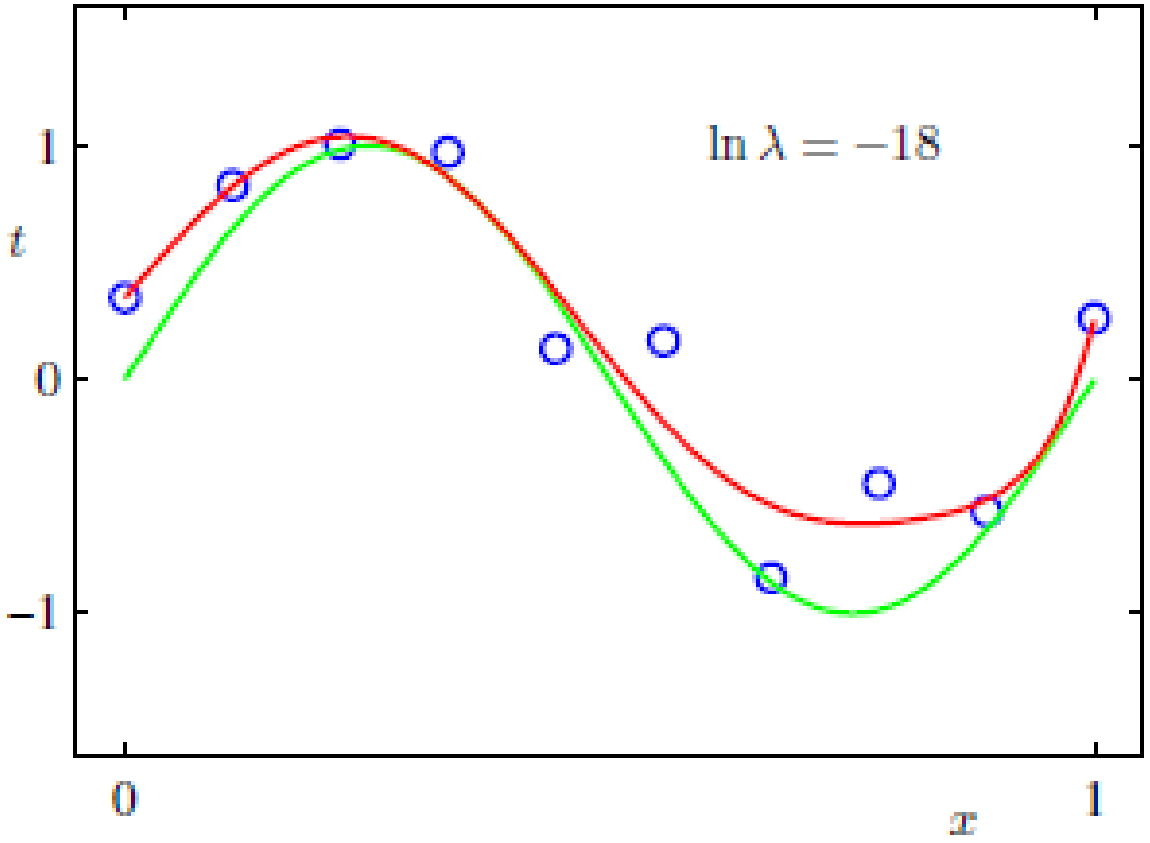


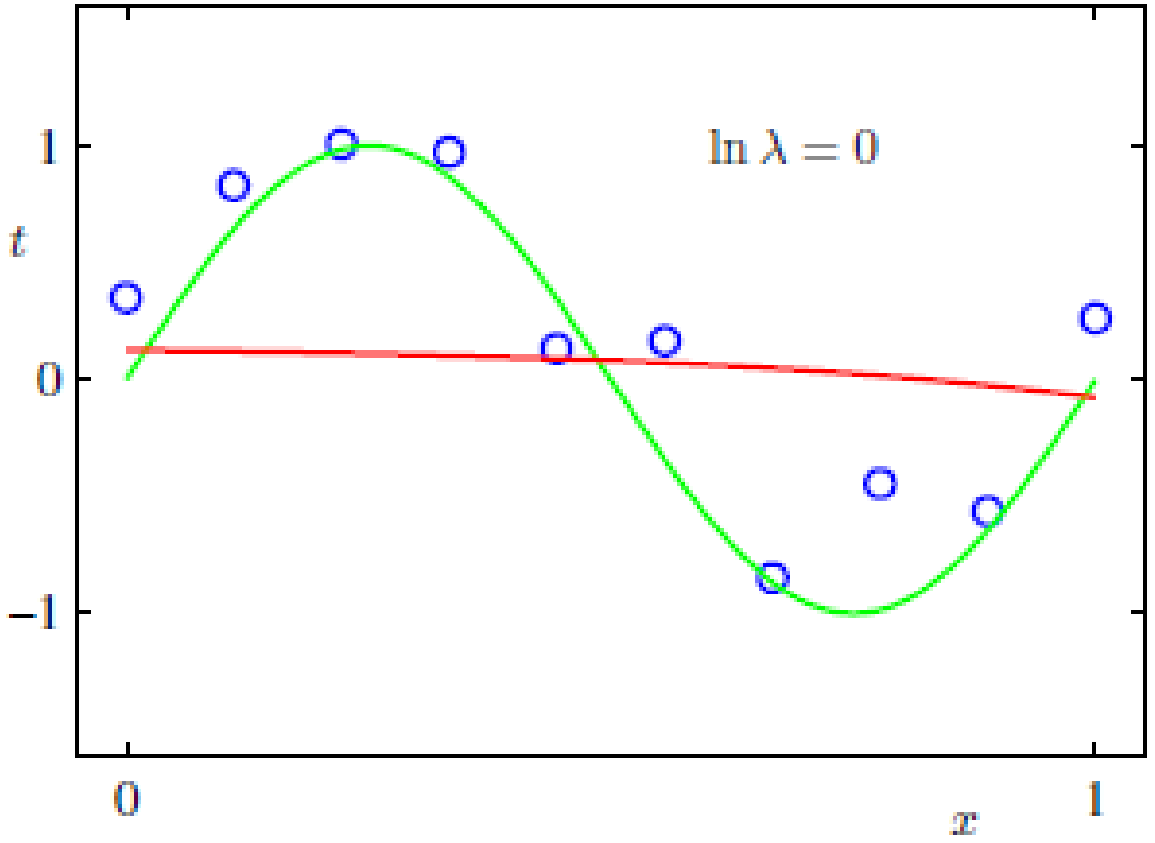
Apprentissage supervisé : Dilemme biais variance





Apprentissage supervisé : Dilemme biais variance





Apprentissage non supervisé

Définition: L'apprentissage non supervisé traite le problème de la capacité des systèmes à apprendre et à représenter certains régimes d'une manière qui reflètent la structure statistique de l'ensemble des motifs d'entrées. Contrairement à l'apprentissage supervisé, dans l'apprentissage non supervisé il n'y a pas de variables de sorties ou d'évaluations environnementales associées à chaque variable d'entrée X. Le seul élément que l'apprentissage non supervisé fait intervenir est la variable d'entrée X qui est souvent supposée issue d'une loi de probabilité $p(X)$ inconnue.

L'algorithme de l'apprentissage non supervisé apprend à partir d'une base d'apprentissage défini de la façon suivante :

$$S = \{(x_i)\}_{i \leq n}$$

Apprentissage non supervisé : Le regroupement automatique

Définition Le problème du regroupement automatique ou le Clustering en anglais consiste à regrouper les données en sous groupes ayant des caractéristiques statistiques semblables. Autrement dit, séparer les données d'entrées en groupes d'individus qui présentent des caractéristiques similaires.

Le regroupement automatique est riche par ces méthodes :

- les approches hiérarchique Classification Ascendante Hiérarchique, CAH),
- L'algorithme des k-means ,
- Les modèles de mélange.
- ….

L'intérêt du regroupent automatique réside dans le fait qu'il permet d'avoir les mêmes sorties qu'un apprentissage supervisé en se basant juste sur les variables d'entrées.

Apprentissage non supervisé : La réduction de la dimension

La réduction de la dimension de l'espace a pour objectif de faciliter l'analyse des données, améliorer les performances des classificateurs, sélectionner les ensembles de données les plus cohérents à la fois pour plus de précision et pour la réduction du temps de calcul des algorithmes, enlever les informations redondantes ou non pertinentes, etc . . .

Les méthodes de la réduction de la dimensionnalité sont généralement regroupées en deux approches :

- Les approches linéaires (ACP Analyse en Composantes Principales, ALD Analyse Linéaire Discriminante, MDS Multi-Dimensional Scaling)
- Les approches non linéaires (Isomap, LLE (Locally Linear Embedding)).

Sélection de caractéristiques

Définition La sélection des caractéristiques est un terme couramment utilisé dans le domaine de la fouille de données pour décrire les outils et techniques disponibles pour réduire la taille des données mises en jeu dans une problématique donnée. La sélection des caractéristiques implique non seulement la réduction de la cardinalité de l'ensemble des variables d'intérêt, qui signifie la sélection arbitraire ou prédéfinie du nombre d'attributs qui peuvent être considérés lors de la construction d'un modèle d'apprentissage, mais aussi le choix des attributs les plus pertinents.

Sélection de caractéristiques

soit $D = \{(x_i^p)\}_{i \leq n}^{p \leq d}$ un ensemble de caractéristiques de taille *d*, où *d* est le nombre total de caractéristiques. On définit une fonction *Ev* qui évalue la pertinence des sous-ensembles sélectionnés et cette fonction est maximale pour le "meilleur ’’ sous-ensemble.

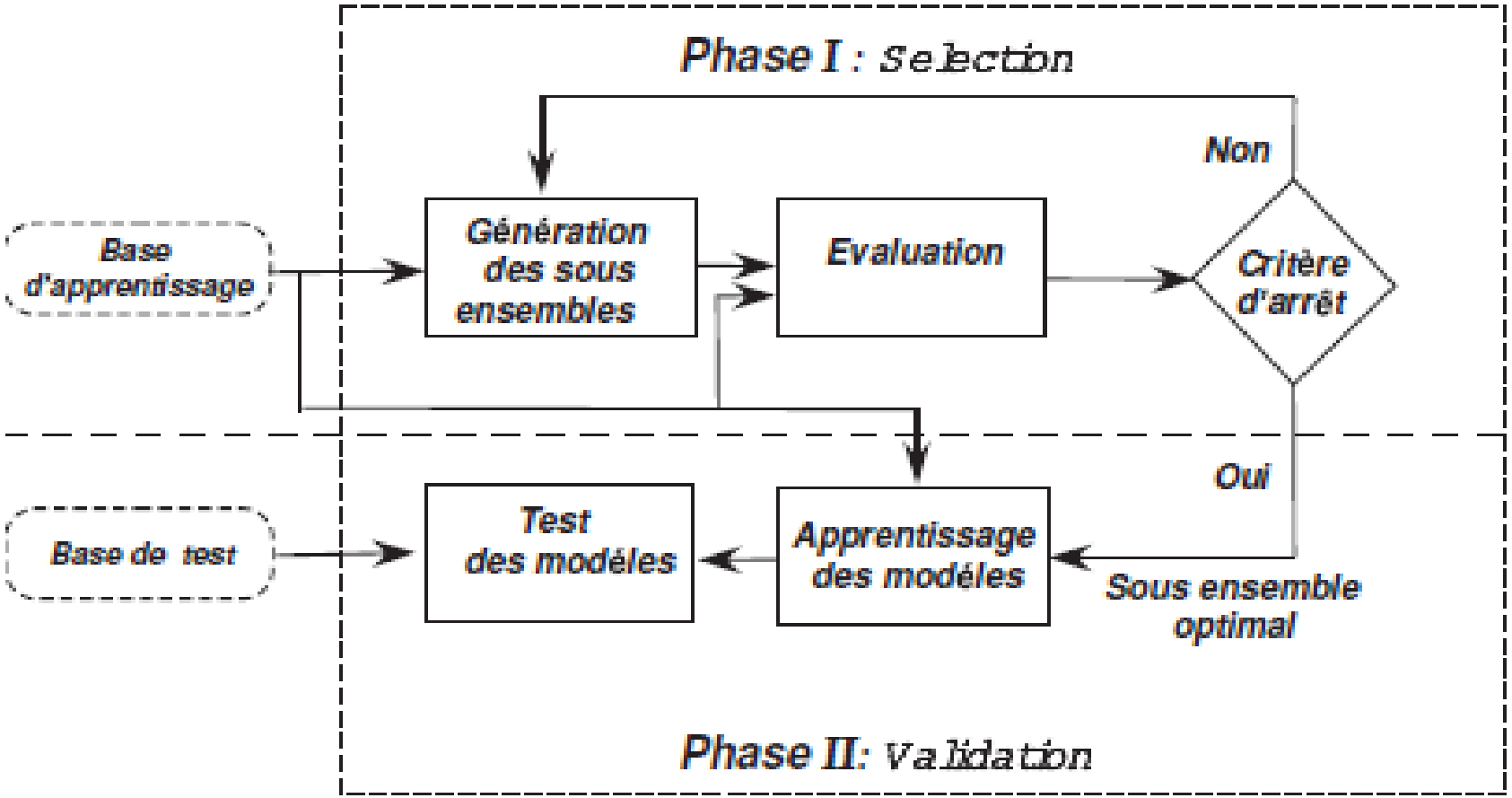
L’objectif de la sélection de caractéristiques est de trouver un sous-ensemble optimal $D'(D' \subset D)$ de taille $d'(d' < d)$ de telle sorte que l’égalité suivante soit vérifiée :

$$Ev(D') = \underset{Z \subset D}{argmax}\, Ev(Z)$$

Avec $\|Z\| = d'$ et d' est un nombre prédéfini par l’utilisateur ou bien contrôlé par le critère utilisé dans le processus de sélection.

Sélection de caractéristiques

Un processus de sélection de caractéristique peut être structuré en deux phases :
Phase de sélection des meilleurs sous ensembles,
Phase de validation



Processus de sélection des caractéristiques.

Sélection de caractéristiques

Phase de sélection

La phase de sélection est subdivisée en trois étapes :

- la génération d'un sous-ensemble à partir de l'ensemble initial selon une stratégie de recherche, qui peut être complète, séquentielle ou aléatoire,

- La phase d'évaluation de tout les sous-ensemble générés dans la première phase selon un critère d'évaluation, qui peut être filtre, l'enveloppe ou hybride

- La phase d'arrêt selon un critère d'arrêt.

Sélection de caractéristiques

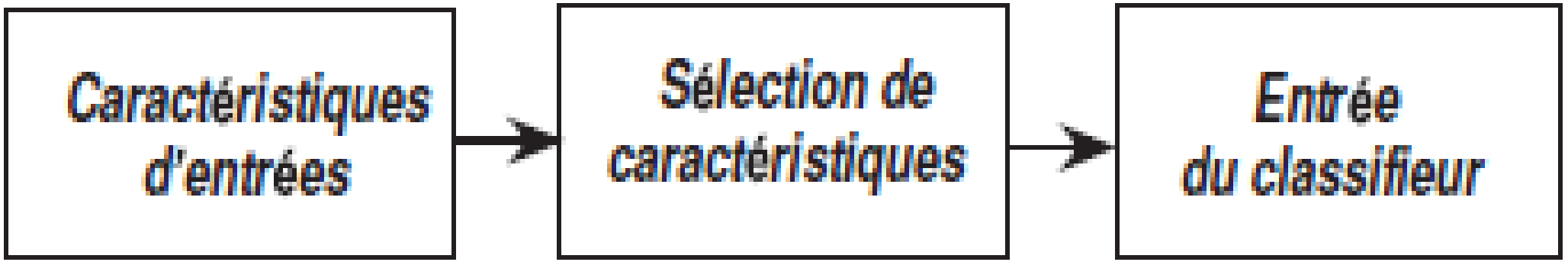
Stratégies de recherche

- Stratégie exhaustive
- Stratégie heuristique :
  - Ascendante(ou forward en anglais)
  - Descendante (ou backward en anglais)
  - Descendante-ascendante (ou stepwise en anglais)
- Stratégie aléatoire

Sélection de caractéristiques

Phase de sélection (Évaluation : approche filtre)

Approche filtre (ou *Filter* en anglais) l'approche filtre repose sur l'analyse et l'évaluation des caractéristiques générales des données sans la mise en jeu d'un algorithme d'apprentissage selon une mesure *M*. Cette approche est considérée comme un prétraitement de données elle intervient en aval de l'algorithme d'apprentissage.



Approche filtre

Sélection de caractéristiques

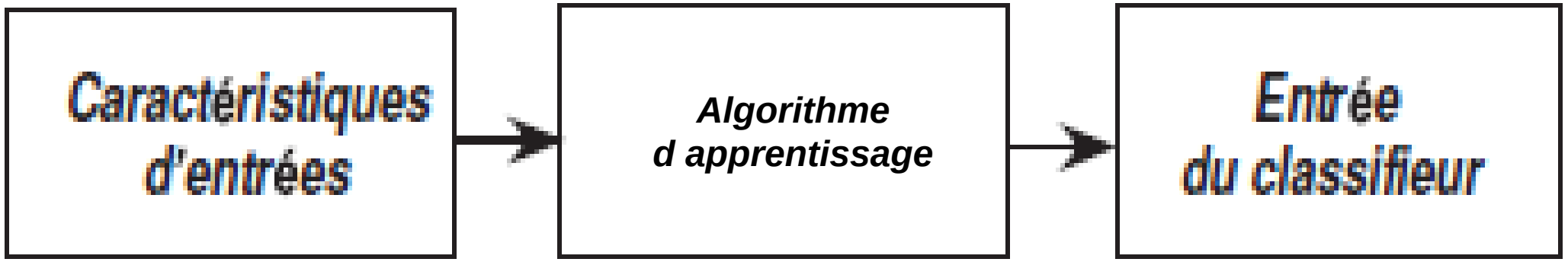
Phase de sélection (Évaluation : approche filtre)

- L’algorithme SFS (Sequential Forward Selection) qui est l’un des premier algorithme de sélection de caractéristiques.
- L’algorithme SBS (Sequential Backward Selection).
- L’algorithme focus
- L’algorithme Relief

Sélection de caractéristiques

Phase de sélection (Evaluation : approche Wrapper)

Dans l'approche Wrapper, un algorithme d'apprentissage est utilisé afin d'évaluer les performances du sous ensemble généré. L'approche de Wrapper est très similaire à l'approche filtre sauf que dans la première, un algorithme d'apprentissage *A* est utilisé à la place d'une mesure *M* pour l'évaluation du sous-ensemble généré.

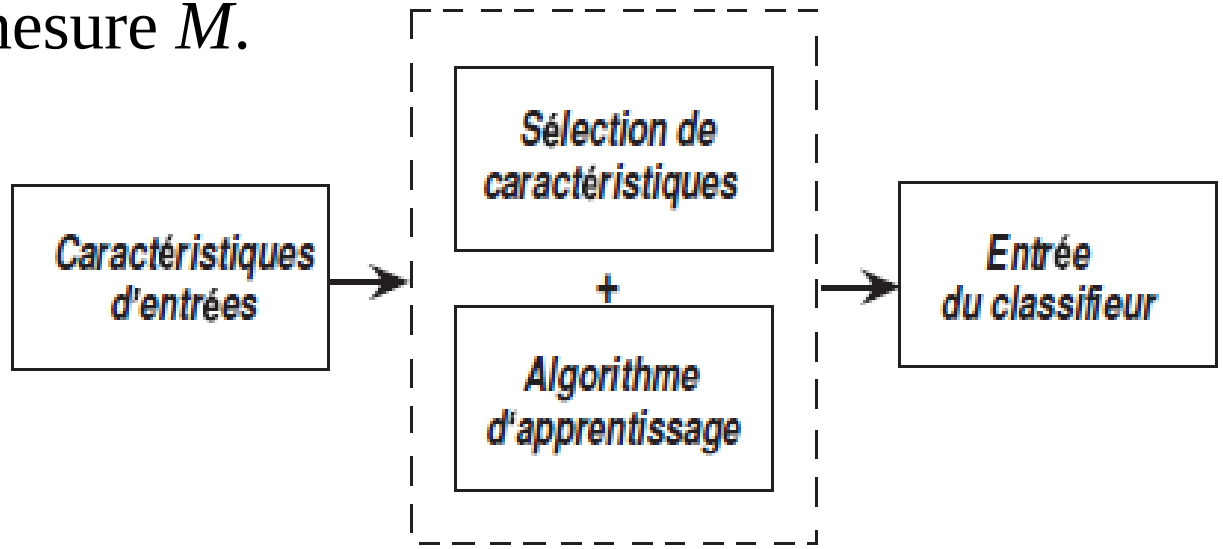


Approche Wrapper

Sélection de caractéristiques

Phase de sélection (Evaluation : approche Hybride)

Approche Hybride L'approche hybride a été récemment proposé pour traiter de grands ensembles de données. Les algorithmes de l'approche hybride utilisent à la fois une mesure *M* et un algorithme d'apprentissage *A* pour évaluer les sous-ensembles générés. La mesure *M* est utilisée pour sélectionner les meilleurs sous-ensembles d'une cardinalité donnée et l'algorithme d'apprentissage *A* est utilisé pour sélectionner le dernier meilleur sous-ensemble parmi les meilleurs sous-ensembles sélectionnés par la mesure *M*.



Approche Hybride

Sélection de caractéristiques

Phase de sélection (Critère d'arrêt)

Un critère d'arrêt détermine le moment où le processus de sélection des caractéristiques doit s'arrêter. Selon l'application et l'objectif visé. Le critère d'arrêt peut être de diverses natures :

- Un nombre minimum de caractéristiques
- Le nombre maximum d'itérations
- Un taux de bonne classification
- Un temps de calcul, etc …

Sélection de caractéristiques

Phase de validation

Un moyen simple pour la validation des caractéristiques sélectionnées est de le comparer à un ensemble de caractéristique connu si nous connaissons les caractéristiques pertinentes à l'avance comme dans le cas de données synthétiques. Dans les applications réelles, souvent on ne dispose pas des connaissances a priori sur les données. Par conséquent, des méthodes indirectes peuvent être utilisées pour évaluer la pertinence des caractéristiques sélectionnées. Par exemple, le taux d'erreur de classification peut être utilisé comme indicateur de performance pour la validation des sous-ensembles sélectionnés.

Mesure de performance d'un classifieur

Matrice de confusion : Dans le domaine de l'apprentissage automatique, une matrice de confusion, est un tableau de contingence qui permet la visualisation des performances d'un algorithme d'apprentissage. Elle permet aussi de voir facilement si l'algorithme d'apprentissage confond deux classes. Chaque colonne de la matrice de confusion représente les instances d'une classe estimé, tandis que chaque linge représente des instances d'une classe réelle.

		Classe réelle		
		Positive	Negative	Total
Classe estimée	Positive	V_p	F_p	$V_p + F_p$
	Negative	F_n	V_n	$F_n + V_n$
	Total	$V_p + F_n$	$F_p + V_n$	N

Mesure de performance d'un classifieur

l'Accuracy : L'accuracy est la métrique la plus courante pour l'évaluation des performances des algorithmes d'apprentissage automatique. Elle mesure la proportion d'exemples classés correctement. Dans un problème de classification binaire, l'accuracy est définie comme suit :

$$Accuracy = \frac{Vp + Vn}{Vp + Vn + Fp + Fn}$$

avec

- Vp représente les vrais positifs ;
- Vn représente les vrais négatifs ;
- Fn représente les faux négatifs ;
- Fp représente les faux positifs.

Mesure de performance d’un classifieur

Le rappel : Le rappel est une métrique qui permet d’évaluer la proportion de solutions pertinentes qui sont trouvées. En d’autre termes, il mesure la capacité du système à donner toutes les solutions pertinentes. l’expression de cette métrique est la suivante :

$$rappel = \frac{Vp}{Vp + Fn}$$

.

Mesure de performance d'un classifieur

La précision : La précision est une métrique qui permet d'évaluer la proportion de solutions trouvées qui sont pertinentes. Elle mesure la capacité du système à refuser les solutions non-pertinentes :

$$Precision = \frac{Vp}{Vp + Fp}$$

.

Mesure de performance d'un classifieur

La F-mesure : Dans l'analyse statistique d'un classification binaire, la F-mesure est une métrique de performance qui est basée sur la moyenne harmonique de la précision et du rappel. En d'autres termes elle mesure la capacité du classifier à donner toutes les solutions pertinentes et à refuser les autres. La formule explicite générale de la F-mesure peut être donnée comme suit :

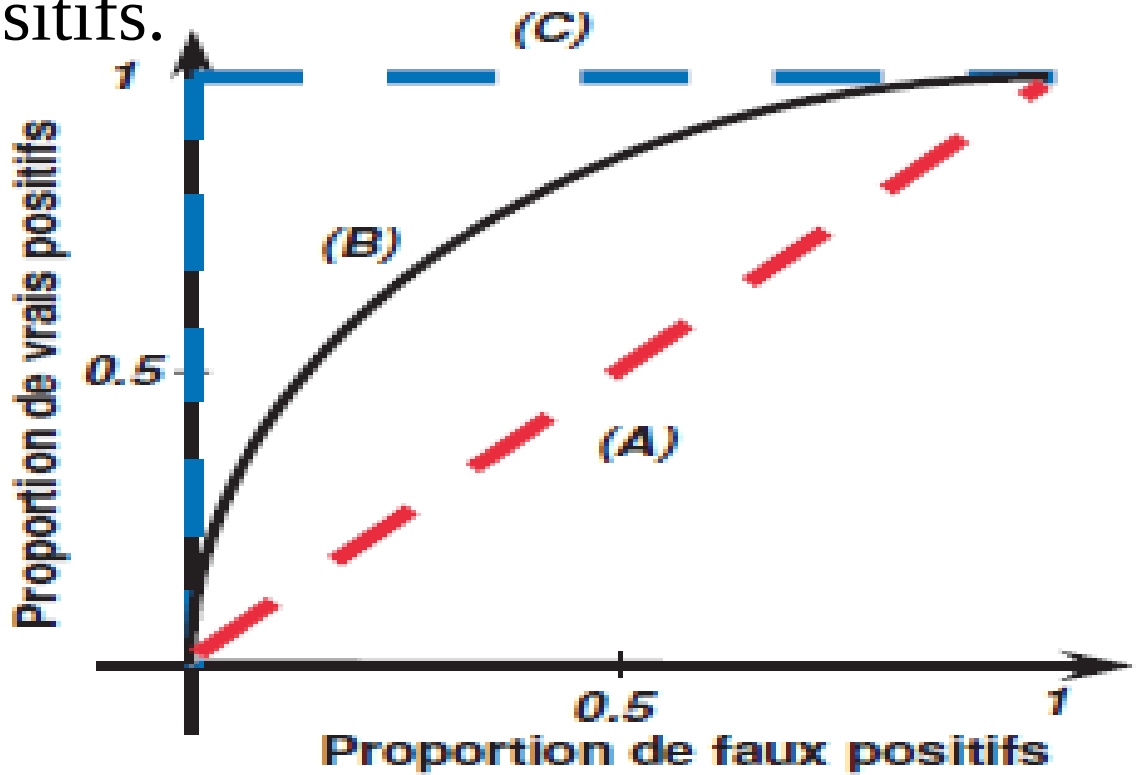
$$F_\beta\text{-}mesure = (1 + \beta^2).\ \frac{rappel.precision}{\beta^2 precision + rappel}$$

β un facteur qui contrôle le degré d'importance de rappel/précision. prend ses valeurs dans .

Mesure de performance d’un classifieur

La courbe COR ou (ROC en anglais) :

La courbe COR est une méthode graphique qui permet de visualiser les performances des algorithmes d’apprentissage statistiques. Sur cette courbe est représentée en abscisse la proportion de faux négatifs et en ordonnée la proportion de vrais positifs.



Mesure de performance d’un classifieur

Une courbe diagonale (la courbe(A)) représente le classifieur aléatoire. Si on obtient une courbe sous la diagonale, cela est souvent synonyme d’un problème sur les sorties du classifieur (présence d’un biais généralement).

Remarque Un classifieur sera donc d’autant meilleur que sa courbe se situera proche de la courbe(C) et loin de la diagonale. Souvent, à partir de cette courbe le critère (*AUC*) (aire sous la courbe COR) est calculé. Ce critère permet de connaître le lien entre le taux de vrais positifs et le taux de faux positifs. Plus l’aire est importante, plus le classifieur possède un pouvoir discriminant sur les classes. L’utilisation du critère (*AUC*) peut se révéler fort intéressante dans le cas où une classe est sous représentée par rapport à une autre.