Université de Mons

Faculté des Sciences

Le phénomène de double descente au sein de l'Apprentissage Automatique

Raphael Palmeri

Sous la direction de Souhaib Ben Taieb

1 ère Master en Sciences informatiques

Je soussigné, Palmeri Raphael, atteste avoir respecté les règles éthiques en vigueur

Année universitaire 2020-2021

Table des matières

1 Introduction

Dans le cadre de notre Master en Sciences Informatiques en horaire décalé, il nous est demandé de réaliser un mémoire sur un sujet proposé par les enseignants de l'UMons de la section Sciences Informatiques. Mon choix de sujet s'est porté sur "Le phénomène de double descente au sein de l'Apprentissage Machine (*'The double descent phenomenon in machine learning'*)".

Dans un premier temps, je vais présenter ce qu'est l'apprentissage Machine ainsi qu'expliquer brièvement les différentes sous-catégories existantes au sein de celui-ci.

Dans une deuxième partie, je vais expliquer les notions de biais et de variance, ainsi que le phénomène de compromis entre ces deux notions au sein de l'apprentissage machine.

Dans la troisième partie, je vais démontrer mathématiquement le compromis biais-variance et expliquer cette démonstration plus en détails.

Dans la quatrième partie, à l'aide de simulations, nous montrerons l'effet de ce phénomène ainsi que son comportement en influençant sur différents paramètres de la simulation.

2 Apprentissage automatique ("Machine Learning")

L'apprentissage automatique, c'est la capacité d'un ordinateur à "apprendre" en se basant sur des données mises à sa disposition. Le terme "apprendre" dans ce cas-ci désigne la capacité à détecter/trouver des répétitions (*'patterns'*) dans ces données. Ces répétitions permettront ensuite à la machine de donner une expertise par rapport à un problème donné ou une réponse à une certaine question. [5][2]

L'apprentissage automatique est notamment utilisé dans différents domaines tels que le domaine de l'automobile avec ses voitures sans conducteurs, le domaine des Finances afin de notamment détecter les fraudes mais aussi le domaine de la santé avec la possibilité d'essayer de produire un diagnostic sur base des informations disponibles à propos d'un patient, ...

Afin de réaliser de l'apprentissage automatique, il est nécessaire d'avoir deux choses : un algorithme d'apprentissage (voir sous-section 2.1) et des données (voir sous-section 2.3).

2.1 Algorithme d'apprentissage

Il existe bien des algorithmes différents utilisés en machine learning. il existe différents types d'algorithme d'apprentissage :

2.1.1 Apprentissage supervisés

Dans ce type d'apprentissage, la machine reçoit un ensemble de données avec les classes de tout les exemples existants. Par exemple, un expert aura déjà défini les différentes classes possibles pour la reconnaissance d'objets via des images ('Chaise', 'Table', 'Chien', 'Chat', ...). Les algorithmes de cette famille vont dès lors se basé sur ces classes déjà définies afin de pouvoir attribuer une classe (que l'on espère correcte) à une nouvelle donnée encore inconnue. Dans l'exemple ci-dessus, on parlera de **"Classification"**.
Il existe aussi des problèmes de **"Régression"**, ceux-ci tentent de liés une nouvelle donnée à un nombre réel. Par exemple, dans le cadre de l'estimation de prix de maison.

Étant donné le fait que le compromis Biais-Variance (voir section 3) n'existent qu'au sein de cette famille d'algorithme, il est logique que celle-ci soit la famille que nous étudierons le plus dans ce rapport.

2.1.2 Apprentissage non-supervisés

Dans le cas d'algorithme non-supervisé, les données d'entrainement et de test sont mélangés. Le modèle n'aura aucun exemple pour s'aider à détecter un pattern, il devra le faire par lui même en étudiant les similarités entre les différentes données et les ranger par groupes afin qu'un expert puisse les utiliser dans le cadre de recherche par exemple. Dans le cadre de l'utilisation de cette famille, on parlera de **"Clustering"**.

2.1.3 Apprentissage semi-supervisés et Apprentissage actif

Dans la plupart des situations, il est impossible de classifié l'ensemble des données d'apprentissage. Dans ce genre de cas, la machine doit dès lors apprendre des classes qui lui sont fournies mais aussi des données non labellisées, c'est ce que l'on appelle l'apprentissage semi-supervisé. Dans le cadre où ce n'est pas un expert qui donne les classes mais bien la machine qui tente de les labellisées, on se trouve dans le cas de l'apprentissage actif.

2.1.4 Apprentissage par transfert et apprentissage multitâche

L'idée principale derrière l'apprentissage par transfert est d'aider le modèle à s'adapter à des situations qu'il n'as pas rencontrés précédemment. Cette forme d'apprentissage s'appuie sur le fait de tuner un modèle générale pour lui permettre de travailler dans un nouveau domaine.

2.1.5 Apprentissage par renforcement

L'apprentissage par renforcement se base sur l'idée de maximisé une récompense selon une ou plusieurs actions. On va dés lors définir en fonction des actions, si elles sont encouragées ou au contraire, découragés.

2.1.6 Exemples d'algorithmes

— Prédicteurs linéaires (*'Linear Predictors'*) : tels que la régression linéaire, perceptron ...
— Boosting
— Support Vector Machines
— Arbres de décision (*'Decision Trees'*)
— Voisin le plus proche (*'Nearest Neighbor'*)
— Réseau de neurones (*'Neural Networks'*)
— ...

2.2 Notation mathématiques

Dans cette sous-section, je fais expliquer les différentes notations mathématiques nécessaires à la bonne compréhension des prochains chapitres.

Dans le cadre d'un apprentissage automatique supervisé, on cherche à prédire un résultat $y \in \boldsymbol{Y}$ à partir d'une donnée $x \in \boldsymbol{X}$ où les paires (x, y) proviennent d'une distribution inconnue $\boldsymbol{D}$.

Le problème d'apprentissage automatique consiste à apprendre une fonction $f' : \boldsymbol{X} \to \boldsymbol{Y}$ à partir d'un ensemble de données d'entrainement fini $\boldsymbol{S}$ contenant m variables indépendantes et identiquement distribuées ($i.i.d$) provenant de $\boldsymbol{D}$.

f' peut aussi être vue comme étant une hypothèse $h \in \boldsymbol{H}$, choisie à partir d'une classe d'hypothèses $\boldsymbol{H}$ contenant des fonctions possibles pour le modèle.

Dans un cadre idéal, la fonction f' serait équivalente à la fonction f, la 'vrai' fonction qui régit $\boldsymbol{X} \to \boldsymbol{Y}$.

2.2.1 La fonction de perte quadratique

Elle s'exprime comme suit :

$$(y - y')^2$$

où **y** représente la valeur véritable de la vrai fonction et **y'** représente la valeur estimée par le modèle.

L'erreur quadratique moyenne quant à elle n'est que la moyenne des erreurs sur l'ensemble des données :

$$MSE = \frac{1}{n} \sum_{(x,y) \in D} (y - y')^2$$

2.2.2 Espérance mathématique d'une variable aléatoire

Dans les preuves de la décomposition du biais-variance, on peut y trouver une notation statistique appelée l'espérance mathématique qui se note $E(x)$ pour une variable aléatoire x.

L'espérance mathématique représente la moyenne pondérée des valeurs que peut prendre cette variable.

2.3 Données

Afin de permettre le bon fonctionnement de l'apprentissage automatique, il est nécessaire d'avoir des données. Celles-ci doivent être présentes en quantité et, dans le meilleur des cas, elles doivent êtres "nettoyées" c-à-d, il faut parfois retirer des attributs inutiles, en modifier certains pour qu'il soit compréhensibles pour l'algorithme et certains sont inutilisables car incomplets.

Ces données peuvent être distinguées en 2 catégories :

2.3.1 Données d'apprentissage

Ces données sont des exemples déjà traitées par un expert dans le domaine qui peuvent être utilisés comme exemple d'apprentissage pour les algorithmes supervisés. Grâce à celles-ci, l'algorithme pourra générer un modèle qui pourra estimer la valeur (ou la classe) en fonction d'une donnée inconnue.

2.3.2 Données de test

Ces données sont destinés à valider le modèle crée par l'algorithme d'apprentissage. l'idée est de fournir des données non vues précédemment à la machine afin de vérifier et valider son comportement. Si le modèle produit des résultats extrêmement éloignés de la vérité, c'est qu'il n'est pas encore prêt. Il faut donc repasser par une phase d'apprentissage en fournissant potentiellement plus de données d'apprentissage et/ou en les rendant plus précises afin que la machine établissent un nouveau modèle dont les réponses seront plus correctes.

3 Le compromis Biais-Variance

Afin de mieux comprendre la notion du compromis de Biais-Variance, il est important de comprendre sa version plus général qui est le compromis Approximation-Estimation (voir sous-section 3.1)

3.1 Le compromis approximation-estimation

L'erreur totale au sein de l'apprentissage machine est constitué de 3 choses :

$$error_{total} = error_{generalization} + error_{training} + error_{irreductible}$$

1. l'erreur de généralisation (*'generalization error'*) : cette erreur est la conséquence de la sélection d'un sous-ensemble que l'on considère comme étant représentatif. De part cette sélection, on induit une possible erreur.

2. l'erreur d'entrainement (*'training error'*) : cette erreur est la conséquence de l'apprentissage, dans nos données sélectionnées pour l'apprentissage de la machine, on peut avoir des cas spécifiques qui ne se présentent que dans notre ensemble d'apprentissage. ce qui mènera le modèle à un 'biais d'apprentissage' et peut diminuer la précision de celui-ci lors de l'utilisation de données de test.

3. l'erreur irréductible (*'irreductible error'*) : cette erreur est la conséquence d'un traitement peu efficace des données en amont de l'apprentissage, les données qui seront utilisées par l'algorithme doivent êtres nettoyés avant d'être utilisées. cette erreur ne dépends donc pas de l'algorithme directement.
"Garbage In, Garbage Out", ce qui signifie que si les données en entrée ne sont pas correctes, les résultats ne sauraient l'être.

3.2 Le compromis Biais-Variance

Le compromis biais-variance permet de quantifier le compromis approximation-estimation lorsque l'on utilise la fonction de perte quadratique (ou perte *L2*) et plus particulièrement, l'erreur quadratique moyenne (*MSE*).

3.2.1 Le biais

Le biais est la mesure qui montre à quel point le modèle établi par l'algorithme d'apprentissage supervisé est proche de la 'vrai' fonction d'un problème donné.

La figure 1 montre une représentation graphique du biais. f représente la 'vrai' fonction d'un problème donné, H représente l'ensemble des hypothèses choisies et le point noir représente une hypothèse sélectionnée parmi H.

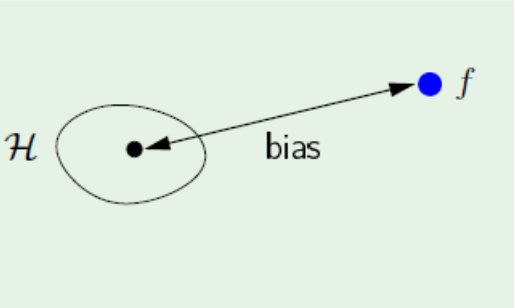

FIGURE 1 – Représentation du Biais.
[4]

3.2.2 La Variance

La variance est la variation entre la valeur d'un ensemble de données de test par rapport à la valeur donnée par le modèle choisi.

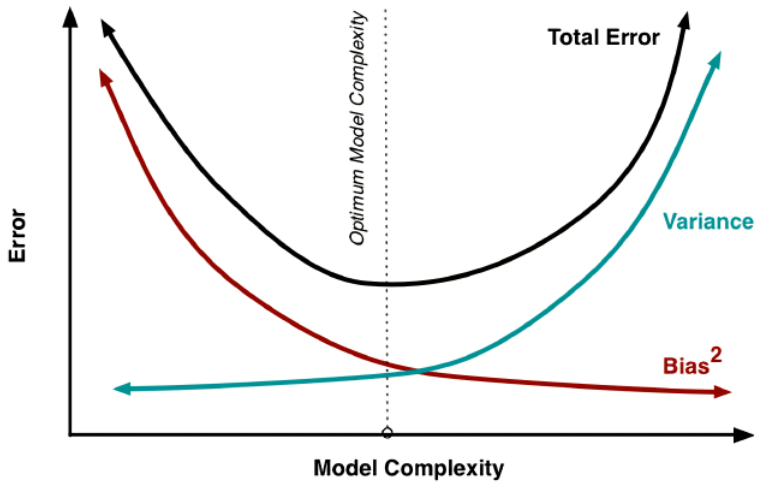

FIGURE 2 – Représentation de l'erreur en fonction de la complexité du modèle [4]

La figure 2 montre l'impact du biais et de la variance sur l'erreur d'un modèle et ce en fonction de sa complexité. Elle montre aussi très bien le point d'optimisation de la complexité du modèle lié au compromis entre le biais et la variance.

Un exemple concret du compromis biais-variance est celui du tir sur cible :

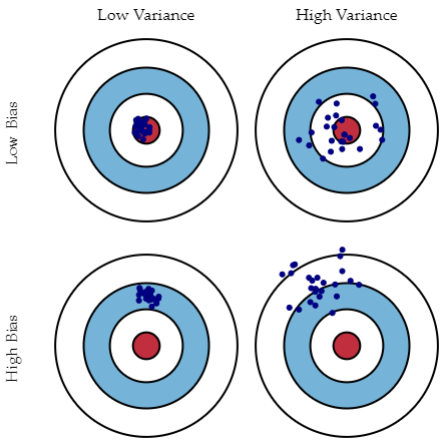

FIGURE 3 – Exemple concret du Compromis Biais-Variance.
[6]

Dans la figure 3, on a quatre cibles selon deux axes différents, le biais et la variance, chacun de ces axes peut-être soit faible, soit élevé.

Le premier cas (biais faible et variance faible) représente un excellent tireur, il vise toujours le centre et ses tirs sont fortement groupés.

Le second cas (biais faible et variance élevée) représente un 'bon' tireur, il vise le centre mais ses tirs sont assez dispersés.

Le troisième cas (biais élevé et variance faible) représente un tireur moyen, il ne vise pas le centre mais n'est pas non plus hors de la cible ou au bord de celle-ci et ses tirs sont fortement groupés.

Le quatrième cas (biais élevé et variance élevée) représente un mauvais tireur, il ne vise pas le centre et ses tirs sont très dispersés.

3.3 Décomposition du biais-variance

Considérons le modèle suivant :

$$y = f(x) + \epsilon \tag{1}$$

où :
— x ∼ p(x)
— f est une fonction fixée inconnue
— ϵ est du bruit aléatoire tel que :
  — $E[\epsilon|x] = 0$
  — $Var(\epsilon|x) = \sigma^2$

étant donné un set de données $D = (x_i, y_i)_{i=1}^n$ où (x_i, y_i) est un échantillon provenant de (1), et un ensemble d'hypothèses H, on calcule :

$$g^{(D)} = argmin_{h \in H} E_{in}(h) := \frac{1}{n} \sum_{i=1}^{n} L(y_i, h(x_i))$$

étant donné D, l'erreur au carré hors-échantillon de $g^{(D)}$ est :

$$E_{out}(g^{(D)}) = E_{x,y}[(y - g^{(D)}(x))^2]$$

considérons

$$E_D[E_{out}(g^{(D)})] = E_{x,y,D}[(y - g^{(D)}(x))^2] \tag{2}$$

représentant l'espérance moyenne sur les variables x, y et D.

En prenant $\bar{g} = E_D[g^{(D)}(x)]$, on peut décomposer (2) comme suit

$$\underbrace{E_x[(f(x) - \bar{g}(x))^2]}_{Biais} + \underbrace{E_{x,D}[(\bar{g}(x) - g^{(D)}(x))^2]}_{Variance} + \underbrace{\sigma^2}_{Variance irreductible}$$

On sait que le compromis biais-variance s'exprime comme suit :

$$E_{x,y,D}[(y - g^{(D)}(x))^2]$$

En sachant que le g moyen est :

$$\bar{g}(x) = E_D[g^{(D)}(x)] \tag{3}$$

Prouvons qu'il est égal à :

$$E_x[(f(x) - \bar{g}(x))^2] + E_{x,D}[(\bar{g}(x) - g^{(D)}(x))^2] + \sigma^2$$

4 Simulation

5 Conclusion

Annexes

A Preuve mathématique

Reprenons l'expression du compromis :

$$E_{x,y,D}[(y - g^{(D)}(x))^2] \tag{4}$$

On peut exprimer (4) comme suit :

$$E_x[E_{y,D}[(y - g^{(D)}(x))^2|x]] \tag{5}$$

En fixant x, on peut simplifier (5) :

$$E_{y,D}[(y - g^{(D)}(x))^2] \tag{6}$$

En ajoutant $-f(x) + f(x)$ à l'équation 6, on obtient :

$$E_{y,D}[(y - f(x) + f(x) - g^{(D)}(x))^2] \tag{7}$$

En considérant $y - f(x)$ comme étant a et $f(x) - g^{(D)}(x)$ comme b et en appliquant la formule $(a + b)^2 = a^2 + b^2 + 2ab$ dans l'équation (7), on obtient :

$$E_{y,D}[(y - f(x))^2] + E_{y,D}[(f(x) - g^{(D)}(x))^2] + 2E_{y,D}[(y - f(x))(f(x) - g^{(D)}(x))] \tag{8}$$

En utilisant (1) pour remplacer y dans l'équation (8), on obtient :

$$E_{y,D}[(f(x)+\epsilon-f(x))^2]+E_{y,D}[(f(x)-g^{(D)}(x))^2]+2E_{y,D}[(y-f(x))(f(x)-g^{(D)}(x))] \tag{9}$$

En simplifiant l'équation (9), on obtient :

$$E_{y,D}[(\epsilon)^2] + E_{y,D}[(f(x) - g^{(D)}(x))^2] + 2E_{y,D}[(y - f(x))(f(x) - g^{(D)}(x))] \tag{10}$$

En utilisant la définition de la Variance de ϵ de la section 3.3, on peut simplifier l'équation (10) comme suit :

$$\sigma^2 + E_{y,D}[(f(x) - g^{(D)}(x))^2] + 2E_{y,D}[(y - f(x))(f(x) - g^{(D)}(x))] \tag{11}$$

En utilisant (26) (voir section B) dans l'équation (11), on obtient :

$$\sigma^2 + E_{y,D}[(f(x) - g^{(D)}(x))^2] + 0 \tag{12}$$

En ajoutant $-\bar{g}(x) + \bar{g}(x)$ à l'équation (12) dans le terme $E_{y,D}[(f(x) - g^{(D)}(x))^2]$, on obtient :

$$\sigma^2 + E_{y,D}[(f(x) - \bar{g}(x) + \bar{g}(x) - g^{(D)}(x))^2] + 0 \tag{13}$$

En considérant $f(x) - \bar{g}(x)$ comme étant a et $\bar{g}(x) - g^{(D)}(x)$ comme b et en appliquant la formule $(a + b)^2 = a^2 + b^2 + 2ab$ dans l'équation (13), on obtient :

$$\sigma^2+E_{y,D}[(f(x)-\bar{g}(x))^2]+E_{y,D}[(\bar{g}(x)-g^{(D)}(x))^2]+2E_{y,D}[(f(x)-\bar{g}(x))(\bar{g}(x)-g^(D)(x))] \tag{14}$$

En vérifiant les espérances, on peut encore simplifier l'équation (14) en :

$$\sigma^2+(f(x)-\bar{g}(x))^2+E_D[(\bar{g}(x)-g^{(D)}(x))^2]+2E_{y,D}[(f(x)-\bar{g}(x))(\bar{g}(x)-g^(D)(x))] \tag{15}$$

En utilisant la preuve intermédiaire (31) (voir section C), on obtient l'équation suivante :

$$\sigma^2 + (f(x) - \bar{g}(x))^2 + E_D[(\bar{g}(x) - g^{(D)}(x))^2] + 0 \tag{16}$$

et finalement en ré-appliquant l'espérance de x que nous avions retiré pour faciliter la notation, on obtient :

$$\sigma^2 + E_x[(f(x) - \bar{g}(x))^2] + E_{x,D}[(\bar{g}(x) - g^{(D)}(x))^2] \tag{17}$$

ce qui prouve bien que $E_{x,y,D}[(y - g^{(D)}(x))^2]$ est équivalent à (17)

B Preuve intermédiaire de $2E_{y,D}[(y-f(x))(f(x)-g^{(D)}(x))]$

Prouvons que $E_{y,D}[(y - f(x))(f(x) - g^{(D)}(x))]$ est $= 0$

$$E_{y,D}[(y - f(x))(f(x) - g^{(D)}(x))] \tag{18}$$

On peut distribuer dans l'équation (18), on obtient :

$$E_{y,D}[yf(x) - yg^{(D)}(x) - f^2(x) + f(x)g^{(D)}(x)] \tag{19}$$

En utilisant (1) dans l'équation (19), on obtient :

$$E_{y,D}[(f(x) + \epsilon)f(x) - (f(x) + \epsilon)g^{(D)}(x) - f^2(x) + f(x)g^{(D)}(x)] \tag{20}$$

En séparant les différents éléments et en simplifiant dans (20), on obtient :

$$E_{y,D}[f^2(x)+\epsilon f(x)]-E_{y,D}[f(x)g^{(D)}(x)+\epsilon g^{(D)}(x)]-E_{y,D}[f^2(x)]+E_{y,D}[f(x)g^{(D)}(x)] \tag{21}$$

En vérifiant les espérances, on peut encore simplifier l'équation (21) en :

$$f^2(x) + \epsilon f(x) - E_D[f(x)g^{(D)}(x) + \epsilon g^{(D)}(x)] - f^2(x) + E_D[f(x)g^{(D)}(x)] \tag{22}$$

En utilisant (3) dans l'équation (22), on obtient :

$$f^2(x) + \epsilon f(x) - f(x)\bar{g}(x) + \epsilon g^{(D)}(x) - f^2(x) + f(x)\bar{g}(x) \tag{23}$$

Pour faciliter la notation, nous avions fixé x, l'équation (23) donne en réalité :

$$E_x[f^2(x) + \epsilon f(x) - f(x)\bar{g}(x) + \epsilon g^{(D)}(x) - f^2(x) + f(x)\bar{g}(x)] \tag{24}$$

En utilisant la définition de l'espérance de ϵ dans l'équation (24), on obtient :

$$E_x[f^2(x) + 0f(x) - f(x)\bar{g}(x) + 0g^{(D)}(x) - f^2(x) + f(x)\bar{g}(x)] \tag{25}$$

En simplifiant l'équation (25), on obtient finalement :

$$E_x[f^2(x) - f(x)\bar{g}(x) - f^2(x) + f(x)\bar{g}(x)] = E_x[0] = 0 \tag{26}$$

On a donc prouvé mathématiquement que (18) est bien égale à 0

C Preuve intermédiaire de $2E_{y,D}[(f(x)-\bar{g}(x))(\bar{g}(x)-g^{(D)}(x))]$

Prouvons que $E_{y,D}[(f(x) - \bar{g}(x))(\bar{g}(x) - g^{(D)}(x))]$ est $= 0$

$$E_{y,D}[(f(x) - \bar{g}(x))(\bar{g}(x) - g^{(D)}(x))] \tag{27}$$

On peut distribuer dans l'équation (27), on obtient :

$$E_{y,D}[f(x)\bar{g}(x) - f(x)g^{(D)}(x) - g^2(x) + \bar{g}(x)g^{(D)}(x)] \tag{28}$$

en vérifiant les espérances dans l'équation (28), on obtient :

$$f(x)\bar{g}(x) + E_D[-f(x)g^{(D)}(x)] + E_D[\bar{g}(x)g^{(D)}(x)] - g^2(x) \tag{29}$$

en appliquant la formule du g moyen (3), on obtient l'équation suivante :

$$f(x)\bar{g}(x) - f(x)\bar{g}(x) + \bar{g^2}(x) - \bar{g^2}(x) \tag{30}$$

Finalement, en simplifiant l'équation (30), on obtient :

$$f(x)\bar{g}(x) - f(x)\bar{g}(x) + \bar{g^2}(x) - \bar{g^2}(x) = 0 \tag{31}$$

Références

[1] Belkin M., Hsu D., Ma S., Mandal S., Reconciling modern machine learning practice and the bias-variance trade-off *arXiv :1812.11118v2*, November 1-4, 2015, pp. 337-350.

[2] Fernandes de Mello R., Antonelli Ponti M., *Machine Learning A practical Approach on the Statistical Learning Theory*, Springer, Cham, 2018.

[3] Geman S., Bienenstock E., Doursat R., Neural Networks and the Bias/Variance Dilemma *Neural Computation* 4, 1-58, 1992 `http://direct.mit.edu/neco/article-pdf/4/1/1/812244/neco.1992.4.1.1.pdf`

[4] Neal B., On the Bias-Variance Tradeoff : Textbooks Need an Update *arXiv :1912.08286v1*, December 2019

[5] Shalev-Shwartz S., Ben-David S., *Understanding Machine Learning From Theory to Algorithms*, Cambridge University Press, 2019 (12th printing).

[6] `http://scott.fortmann-roe.com/docs/BiasVariance.html`, consulté le 18 Juin 2021 à 09 :25

Table des figures