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

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. [1][2]

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 2.2).

2.1 Algorithme d'apprentissage

Il existe bien des algorithmes différents utilisés en machine learning. On les classe en deux catégories principales :

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.

Afin de permettre l'apprentissage supervisé, il est nécessaire de fournir deux types de données à l'algorithme :
les données d'apprentissage qui est, comme expliqué ci-dessus, un ensemble de données ayant déjà reçus un label (ou une classe), cet ensemble est considéré comme complet, ç-à-d, tous les classes possibles existent en son sein et il existe au moins une donnée pour chacune des classes existantes.

les données de test qui est un ensemble de données qui ne possèdent pas encore de classes qui serviront à tester et valider le modèle de la machine afin de vérifier que ses prédictions sont correctes. Dans le cas idéal, ces données n'ont jamais été traités par l'algorithme de manière à tester correctement le modèle (Si toutefois, il n'existe que très peu de données, il sera alors nécessaire d'utiliser de la validation croisée (*'cross-validation'*) afin d'obtenir des résultats concluants.)
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 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 :

2.2.1 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.

2.2.2 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

Le compromis Biais-Variance n'est présent que dans le cadre d'un apprentissage supervisé!!!

Le compromis biais-variance est valide uniquement lorsque l'on utilise les erreurs au carré (squared errors)

Le biais est la différence entre les prédictions attendues du modèle trouvé et les vraies valeurs. Dans la figure 1, elle est indiqué comme étant la courbe de *'Training Risk'*.

La variance indique l'écart des résultats en fonction de l'ensemble de données utilisées. Plus la variance est élevée et moins on pourra être certain des résultats d'un ensemble de données à un autre. Dans certains cas, les résultats seront excellents et dans d'autres cas, les résultats seront médiocres. Dans la figure 1, elle est indiqué comme étant la courbe de *'Test risk'*.

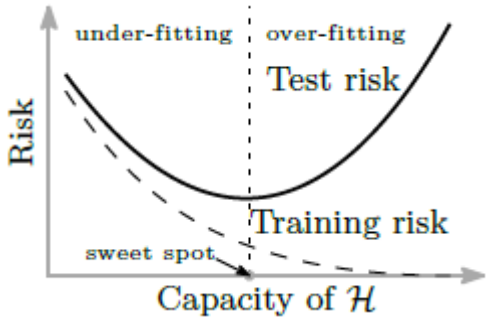

FIGURE 1 – Représentation Compromis Biais-Variance. [3]

3.1 Démonstration mathématique

L'erreur totale au sein de l'apprentissage machine est constitué de 3 choses (le compromis approximation-estimation qui est la version généralisé) :

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

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

2. l'erreur d'entrainement ou la variance (*'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"*

En sachant que le biais s'exprime comme suit :

$$Biais(\hat{f}(x)) = E(\hat{f}(x)) - f(x)$$

$\hat{f}(x)$ étant le modèle choisi (qui possiblement se rapproche le plus de la réelle fonction $f(x)$ qui est inconnue).

et que la variance s'exprime ainsi :

$$Var(\hat{f}(x)) = E(\hat{f}(x)^2) - E(\hat{f}(x))^2$$

3.2 Décomposition du biais-variance

Considérons le modèle suivant :

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

où :
— x ∼ p(x)
— f est un 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

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

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 :

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

3.2.1 Preuve

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 3.2, 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 3.2.2) 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 (voir section **??**), 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 :

$$\sigma^2 + E_x[(f(x) - \bar{g}(x))^2] + E_{x,D}[(\bar{g}(x) - g^{(D)}(x))^2] + 0 \tag{17}$$

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

3.2.2 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

4 Simulation

5 Conclusion

Références

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

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

[3] 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.

Table des figures