Université de Montréal

Examen de synthèse pré-doctoral

Question 2

Romain Martinez

Été 2018

Table des matières

> Question du Dr Jonathan Tremblay :
>
> *L'apprentissage automatique repose sur l'analyse d'une quantité de données pour en extraire un ensemble de relations (apprendre) qui pourront à leur tour être exploitées pour prédire de nouvelles données. Cette approche ne distingue toutefois pas cause et corrélation de par sa construction mathématique, et est incapable d'aller au-delà du cadre imposé par ses données. Quelles sont les conséquences de ce type de limite dans la généralisation de modèles d'apprentissage automatique, particulièrement dans le domaine de la santé humaine, et quelles sont des solutions pour limiter leurs impacts.*

1 Introduction à l'apprentissage automatique

Résumé

Dans cette section, nous introduisons l'apprentissage automatique. Nous présentons d'abord un contexte général puis nous donnons un bref aperçu de certains principes et techniques de base de l'apprentissage automatique. Cette section ne se veut pas être une revue de littérature exhaustive, mais davantage une définition du cadre théorique nécessaire pour faire une analyse critique de l'apprentissage automatique.

Il s'agit de l'un des domaines techniques qui connaît la croissance la plus rapide, à l'intersection des statistiques et de l'informatique, l'apprentissage automatique est devenu omniprésent et indispensable pour résoudre des problèmes complexes dans la majorité des domaines scientifiques. L'apprentissage automatique s'est considérablement développé au cours des dernières décennies (Fig. 1), passant d'une science marginale à une technologie intégrée dans la plupart des industries (Manyika et al. 2011). Aujourd'hui, environ 17,000 articles scientifiques sur l'apprentissage automatique sont publiés par année (Fig. 1)—soit 47 articles par journée, en moyenne. Le développement constant de nouveaux algorithmes et théories, ainsi que la disponibilité accrue à une forte quantité de données et capacité de calcul, ont permis de stimuler cette croissance majeure.

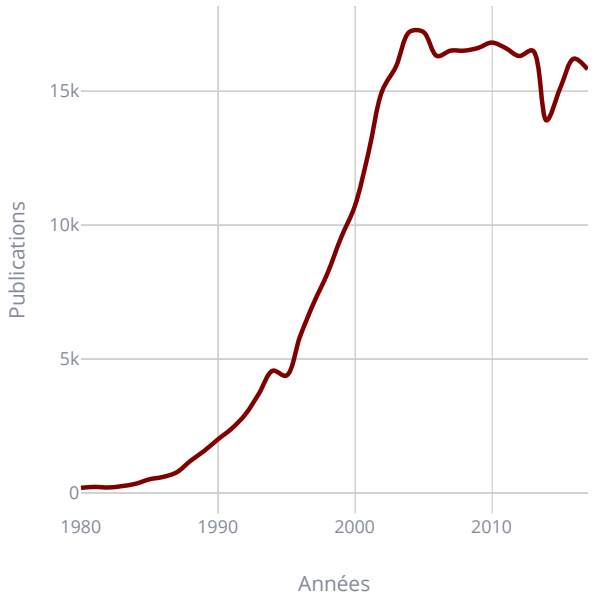

Figure 1 – Publications par année en lien avec l'apprentissage automatique et l'apprentissage profond entre les années 1980 et 2017. Les données ont été récoltées sur Google Scholar avec la requête `"machine learning"OR"deep learning"`.

Plusieurs domaines, scientifiques et industriels, ont déjà adopté l'apprentissage automatique. En astronomie, les algorithmes examinent des millions d'images issues d'études de télescopes pour classer les galaxies et trouver des supernovas (Kremer et al. 2017). Les données de trafic automobile sont utilisées pour améliorer le contrôle des axes routiers et réduire la congestion (Wiering 2000). En santé, les dossiers médicaux sont analysés pour construire des plans de traitement personnalisés (Kononenko 2001) et améliorer les diagnostics médicaux (Obermeyer et Emanuel 2016).

L'une des forces de l'apprentissage automatique est d'intégrer un nombre important de variables et de les combiner de manière non linéaire. Cette association est réalisée de manière automatique et de nombreux utilisateurs réalisent qu'il est plus facile d'entrainer un modèle en lui montrant des exemples (*i.e.* comme un enfant apprendrait à reconnaitre des animaux) plutôt que de le programmer manuellement. Il est ainsi possible d'utiliser de nouveaux types de données (*e.g.* images et vidéos) dont la complexité rendait leur analyse auparavant inimaginable.

Bien qu'il existe un large éventail d'algorithmes à la disposition des utilisateurs, la plupart de ces modèles sont constitués des trois mêmes éléments. Un programme informatique apprend de l'expérience E en ce qui concerne une tâche T et une mesure de performance P, si sa performance sur T, mesurée par P, s'améliore avec l'expérience E (Mitchell et Others 1997). Par exemple, pour dépister une blessure chez un athlète, la tâche T consiste à assigner une étiquette « blessé » ou « non-blessé »; la performance P peut être la précision de ce classificateur de blessures et l'expérience P une collection de données étiquetée sur une variété d'athlètes.

Les types de modèles les plus utilisés sont les méthodes par apprentissage supervisé (John Lu 2010). Le but de l'apprentissage supervisé, qui inclut notre classificateur de blessure, est de produire une prédiction y en réponse à une collection de variables X. La prédiction y est généralement formulée à partir d'une fonction $f(X) = y$ qui produit une sortie y pour chaque entrée X. La fonction f est construite à partir d'exemples étiquetés de y. Il existe plusieurs fonctions f (*e.g.* arbre de décision, régression linéaire, réseau de neurones) qui reflètent divers besoins des applications de l'apprentissage automatique, chacune ayant certains avantages et inconvénients. En apprentissage supervisé, il est possible de prédire des valeurs discrètes (*i.e.* classification) ou des valeurs continues (*i.e.* régression). Contrairement à l'apprentissage supervisé, l'apprentissage non-supervisé, comme son nom l'indique, est capable d'inférer des prédictions sans données étiquetées (Hinton et Salakhutdinov 2006). L'apprentissage non supervisé est principalement utilisé pour regrouper (*e.g.* k-Means), visualiser et réduire les dimensions de données (*e.g.* analyse en composantes principales) (Hinton et Salakhutdinov 2006). Enfin, une troisième famille de modèles, l'apprentissage par renforcement, permet à un agent d'observer l'environnement, d'exécuter des actions, et d'obtenir des récompenses ou des pénalités. Il apprend ainsi, par lui même, quelle est la meilleure stratégie pour obtenir le maximum de récompenses (Mnih et al. 2015).

Conclusion

Dans cette section, nous avons d'abord énoncé la popularité de l'apprentissage automatique. Puis, nous avons présenté les éléments de base qui permettent d'apprendre à un modèle numérique: une expérience, une tâche et une mesure de performance. Enfin, nous avons décrit les trois principales techniques: l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement.

2 Le compromis entre biais et variance

Résumé

Dans cette section, nous décrivons les principaux types d'erreurs que peut faire un modèle en apprentissage automatique. Nous définissons d'abord ce qu'est une erreur de généralisation, puis nous décomposons cette erreur en biais et variance. Enfin, nous évoquons trois moyens de limiter les erreurs de généralisation.

L'objectif fondamental de l'apprentissage automatique est de généraliser au-delà des exemples de l'apprentissage. En effet, il est facile pour un modèle de mémoriser toutes les données qu'on lui fournit. La prédiction que nous obtenons sur les exemples d'entrainement n'est ainsi qu'une approximation de l'erreur que nous obtiendrons sur un nouvel échantillon. Cette dernière erreur est appelée erreur de *test*, à distinguer de l'erreur d'*entrainement*. L'erreur la plus courante chez les débutants est d'évaluer un modèle sur les exemples d'entrainement et d'avoir l'illusion de réussir. Malheureusement, le modèle a simplement appris « par cœur » les données et est incapable de généraliser. Lutter contre ce phénomène, appelé *overfitting* ou variance, est le combat quotidien des utilisateurs d'apprentissage automatique (Domingos 2012). À l'inverse, imaginons un modèle qui prédit systématiquement que tous les joueurs de Hockey marqueront 30 buts dans la prochaine année. Ce modèle, très éloigné de la réalité, n'a aucune variance et l'erreur totale du modèle est attribuée à l'*underfitting* ou biais. Le compromis entre biais et variance est fondamental et inévitable. C'est à l'utilisateur de décider dans quelle direction aller, dépendamment de la question de recherche. Pour illustrer ce compromis, imaginons une partie de fléchettes (Fig. 2). Le vainqueur de la partie est le joueur qui minimise un critère des moindres carrés (*i.e.*, $\min_S S = \sum_{i=1}^{n} (d_i)^2$). Le biais est associé à la tendance systématique du joueur à dévier du centre (Fig. 2, ligne verte) et la variance est représentée par la déviation des lancers individuels par rapport à leur moyenne (Fig. 2, lignes bleues). En apprentissage automatique, nous décomposons l'erreur d'un modèle en biais et variance parce que l'utilisateur à typiquement plus de contrôle sur le biais.

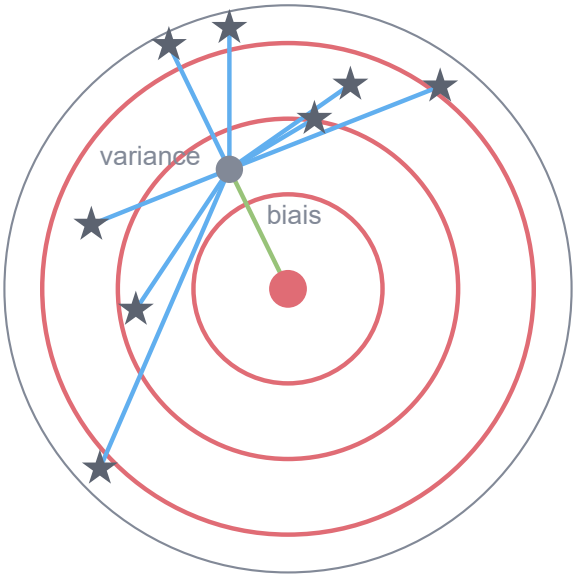

Figure 2 – Analogie du compromis biais-variance avec une partie de fléchettes. La tendance systématique du joueur à dévier du centre (ligne verte) représente le biais tandis que la déviation des lancers individuels par rapport à leur moyenne (lignes bleues) représente la variance.

Les modèles d'apprentissage automatique nécessitent souvent de régler et de tester plusieurs paramètres, appelés *hyperparamètres*, comme le taux d'apprentissage ou le critère d'arrêt. Il n'existe pas de recette universelle pour déterminer les hyperparamètres. La définition des hyperparamètres s'effectue souvent de manière itérative et par optimisation (*e.g.* minimiser l'erreur). Cependant, pour évaluer ces réglages, nous ne pouvons pas utiliser les données d'entrainement (l'évaluation serait optimiste) ni les données de test (notre évaluation finale serait optimiste). Il est donc nécessaire d'avoir un troisième jeu de données—l'ensemble de *validation*. La formation de trois jeux de données (entrainement, validation et test) peut poser problème sur des petites bases de données, ce qui est typiquement le cas en biomécanique. Heureusement, il est possible de recycler le jeu de données d'entrainement pour créer un ensemble de validation. Imaginons un ensemble d'entrainement diviser aléatoirement en 5 sous-ensembles. Lors de la première itération, nous entrainons le modèle sur les quatre premiers sous-ensembles et nous l'évaluons sur le sous-ensemble restant. Lors des itérations suivantes, le modèle est évalué à chaque fois sur un sous-ensemble différent (cinq itérations au total). La performance globale est ensuite calculée en faisant la moyenne de chaque évaluation. Ce processus, appelé validation croisée en sous-ensemble (Browne 2000), permet d'utiliser l'ensemble des données

pour entrainer et évaluer un modèle, sans produire une estimation faussée. La validation croisée permet de lutter contre l'*overfitting*, en choisissant par exemple des modèles ou des hyperparamètres adaptés, mais ce n'est en aucun cas un remplacement de l'ensemble de test (Shao 1993). De plus, c'est une technique qui peut devenir couteuse en temps de calcul comme nous entrainons le modèle pour chaque sous-ensemble.

Dans le cas où notre algorithme *overfit* l'ensemble d'entrainement, deux solutions s'offrent à nous: bâtir un meilleur algorithme ou rassembler plus de données. Les scientifiques se sont surtout intéressés au défi technique de créer des algorithmes toujours plus performants. Mais en tant qu'utilisateur, le moyen le plus direct d'améliorer la généralisation d'un modèle est souvent de collecter plus de données. Une grande quantité de données offre une barrière contre l'*overfitting*. Au plus notre échantillon grossit, au plus il sera représentatif de la population et au moins il aura tendance à apprendre les tendances locales (*e.g.* le bruit) de l'ensemble d'entrainement. Si bien que, en règle générale, un modèle « stupide » avec beaucoup de données généralisera mieux qu'un modèle « intelligent » avec moins de données (Banko et Brill 2001). L'apprentissage automatique n'est pas une solution magique—il ne peut créer du savoir à partir de rien. Outre la quantité de données, c'est aussi la qualité des variables qui importe.

En plus de la validation croisée et l'ajout de données, la régularisation permet également de lutter contre l'*overfitting* en contraignant la complexité d'un modèle. Si on prend l'exemple d'une régression linéaire, celle-ci produira toujours des coefficients non nuls. À l'inverse, la version régularisée de la régression linéaire, aussi appelée régression lasso, aura tendance à réduire les faibles coefficients à zéro. Cette régularisation est intégrée dans la fonction d'optimisation. Cette dernière consiste désormais à réduire la somme des erreurs au carré (comme une régression classique), mais également à minimiser la somme des coefficients. En réduisant le nombre de degrés de liberté du modèle, on réduit également les chances que notre modèle *overfit* (Fig 3). Ainsi, en apprentissage automatique, un modèle complexe n'est pas forcement plus performant qu'un modèle simple (Domingos et Pazzani 1997).

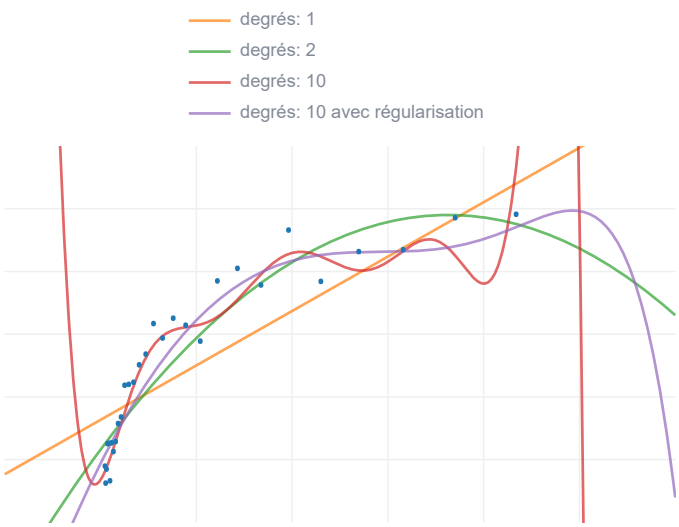

Figure 3 – Régressions linéaires avec différents degrés de polynôme. Une régression linéaire de degrés un (courbe orange) underfit les données tandis que celle de degrés 10 (courbe rouge) overfit. Un polynôme de degrés deux (courbe verte) ou une régression linéaire de degrés 10 avec régularisation (courbe mauve) semblent adaptés. Le code utilisé pour générer cette figure est disponible sur ce lien.

Conclusion

Dans cette section, nous avons discuté des erreurs de généralisation et plus spécifiquement du compromis biais-variance qui la composent. Il n'existe aucun moyen de résoudre ce compromis, cependant, un modèle qui minimise l'erreur de généralisation nécessite généralement trois éléments. Premièrement, il faut pouvoir estimer précisément l'erreur du modèle afin d'en évaluer objectivement les performances (i.e. évaluation sur l'ensemble de validation et de test). Deuxièmement, il faut avoir un contrôle sur le compromis biais-variance (i.e. régularisation du modèle). Troisièmement, il faut avoir accès à une quantité et une qualité de données suffisante pour former des modèles statistiques.

3 Utiliser l'apprentissage automatique en biomécanique

Résumé

Le but de la biomécanique est de comprendre le fonctionnement musculo-squelettique du corps humain, c'est-à-dire être capable de l'expliquer, mais aussi de le prédire. Cependant, la plupart de la recherche en biomécanique se concentre sur l'explication. Ce choix a conduit au développement de modèles théoriques rarement capable de prédire efficacement les mécanismes musculo-squelettiques. La prédiction doit être considérée comme un objectif complémentaire de l'explication, un objectif qui pourrait même encourager notre compréhension. Dans les prochaines sections, nous discutons des bénéfices à utiliser et s'inspirer de l'apprentissage automatique pour améliorer la recherche en biomécanique, notamment concernant l'interprétation et la réplication des résultats.

3.1 Améliorer l'interprétation

L'apprentissage des modèles dont nous avons discuté est, la majorité du temps, basé sur des corrélations entre variables. Cependant, les résultats des modèles d'apprentissage sont parfois faussement interprétés comme des relations causales. L'apprentissage automatique ne résout pas le problème de corrélation et causalité—les prédicteurs ne sont pas les causes (Kleinberg et al. 2015). La corrélation ne signifie pas causalité, mais elle est peut-être un signe de causalité potentielle. Et c'est en tant que guide que nous pourrons utiliser ces corrélations pour investiguer la causalité. Les applications de l'apprentissage profond que nous voyions dans la presse ne représentent qu'un sous-ensemble. La plupart des applications dans l'industrie ou dans les soins de santé concernent la prise de décision. En santé, les modèles prédictifs sont plus souvent utilisés comme des aides à la décision que des prédicteurs. Dans le cas d'un risque de réadmission par exemple, la prédiction est utile parce qu'elle permet de répondre à la question « faut-il renvoyer le patient chez lui? ». Mais l'interprétation est bien plus intéressante, car elle permet d'identifier les raisons d'un risque élevé de réadmission, et donc d'identifier les leviers qui pourraient dimunier ce risque.

Pour illustrer la valeur ajoutée que peut apporter l'interprétation d'un modèle en apprentissage automatique, nous allons utiliser les données d'une expérimentation issue d'une collaboration avec le laboratoire de Julie Côté de l'Université de McGill. Nous suggérons de suivre cet exemple avec le code disponible sur ce lien. Les participants ($n = 81$) devaient réaliser une tâche répétitive de variabilité de motrice jusqu'à un critère d'arrêt (dépasser 45 minutes ou obtenir un score sur l'échelle

de Borg > 8). Les variables récoltées sont la moyenne et l'écart-type des angles de six articulations lors des premières minutes (*i.e.* sans fatigue) et des dernières minutes (*i.e.* avec fatigue). Une question que l'on pourrait se poser est « quel est l'effet de la fatigue sur nos variables cinématique ». Une approche standard en biomécanique serait de tester, statistiquement, la différence entre les variables cinématiques avant et après fatigue. Pour cela, nous utilisons un t-test non paramétrique apparié d'Hotelling (qui est une application du t-test, mais multivariée). Avec 10,000 itérations et $\alpha = 0.05$, le test rejette l'hypothèse nulle ($p = 0.0001$) et l'ensemble des variables semble donc différent avant et après fatigue. Ce test nous autorise à réaliser les tests post-hoc consistant à des t-test non paramétriques appariés. Une fois réalisés, toutes les paires de tests (25) indiquent une différence significative ($p < 0.05$), sauf pour deux variables, l'écart-type de la flexion du coude (AveElbFlexSD) et l'angle moyen du tronc (AveTrYMean). Il est donc difficile d'avoir une conclusion intéressante avec ces tests, hormis « pratiquement toutes les variables cinématiques sont modifiées avec la fatigue pendant une tâche de variabilité motrice ».

Certains scientifiques utilisent les coefficients d'une régression linéaire pour interpréter un modèle. Essayons d'emprunter cette voie en utilisant une régression logistique (qui est une application de la régression linéaire, mais en classification) pour prédire la fatigue. Une fois entrainée sur 80% des données (l'ensemble d'entrainement), celle-ci est capable de prédire la fatigue avec une précision et un rappel de 67% (précision et rappel sont définis dans l'équation 1).

$$\text{precision} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs}} \quad ; \quad \text{rappel} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux négatifs}} \tag{1}$$

Une erreur souvent réalisée est de ne pas normaliser les données avant de les inclure dans un modèle linéaire. Une fois que les données ont été normalisées, notre régression logistique est capable de prédire la fatigue avec une précision et un rappel de 82%. Les coefficients de cette régression indiqueraient que l'angle d'élevation moyen de l'épaule (AveShEleMean) et l'écart-type de la rotation du tronc (AveTrZSD) semblent être des variables importantes pour prédire la fatigue (Fig. 4). En contradiction avec nos tests précédents, les variables précédemment identifiées comme non significativement différentes avant et après fatigue (AveElbFlexSD et AveTrYMean) ont des coefficients non nuls (-0.09 et -0.42, respectivement). Cependant, interpréter les coefficients d'une régression linéaire est un jeu dangereux et nous devons être prudent sur les conclusions d'une telle analyse. Les statisticiens mettent régulièrement en garde les chercheurs contre l'interprétation de coefficients issues de regression linéaire (Shear et Zumbo 2013; Westfall et Yarkoni 2016). Les modèles de régression sont souvent instables et peuvent aboutir à des conclusions trompeuses.

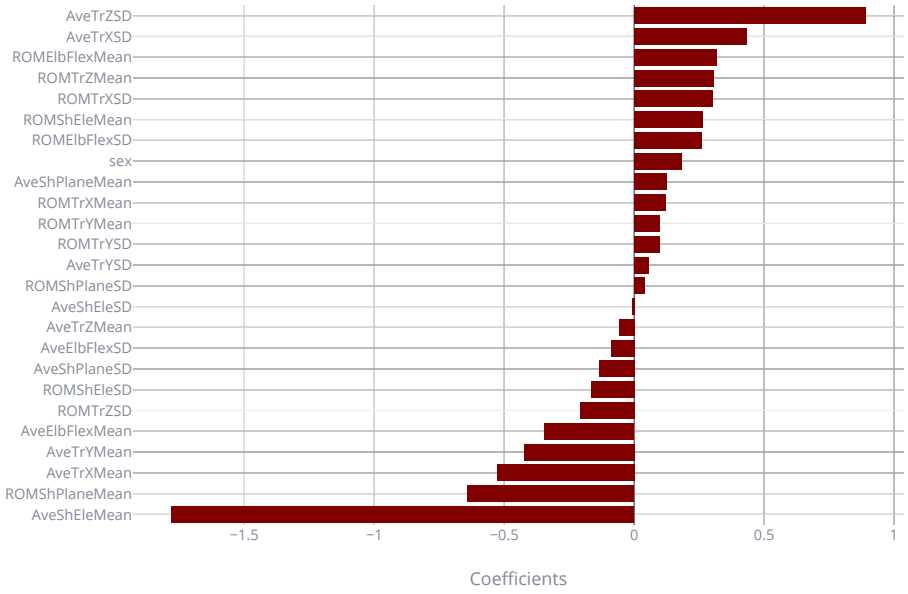

Figure 4 – Coefficient de la régression logistique pour chacune des variables.

Les modèles d'apprentissage automatique sont parfois décrits de manière péjorative comme des «boites noires», impossibles à interpréter. Cette affirmation était vraie pour les réseaux de neurones profonds, mais elle est sujet à changement (Bach et al. 2015; Shrikumar, Greenside, et Kundaje 2017). Pour notre exemple, nous allons tenter de déconstruire ce mythe avec un modèle de *gradient boosting* (Chen et Guestrin 2016). Ce modèle consiste à générer plusieurs arbres de décisions (entre 20 et 100, en moyenne) avec une variation aléatoire de variables et d'observations, à classifier sur chaque sous-ensemble puis à générer une prédiction en calculant la moyenne. Ces méthodes, appelées méthodes d'*ensemble*, permettent de réduire considérablement la variance d'un modèle et donc de diminuer l'erreur de généralisation. Dans un modèle de *gradient boosting*, chaque observation à un poids et celui-ci est modifié pour que le classificateur suivant se concentre sur les exemples générant des erreurs de prédiction. Dans sa première version, notre modèle permet de prédire la fatigue avec une précision de 82% et un rappel de 79%. Des récents efforts de la communauté ont permis de développer des outils qui aide à l'interprétation des modèles. L'un d'eux, la librairie `shap` (Lundberg et Lee 2017; Lundberg, Erion, et Lee 2018), est particulièrement utile pour les modèles d'ensemble. Pour avoir un aperçu des variables les plus importantes, la librairie `shap` nous permet de calculer les valeurs *shap* (*Shapley Additive Explanations*), représentant l'impact de chaque variable sur la prédiction du modèle (Algo. 1). À partir de cette information (Fig. 5), nous pouvons sélectionner les

six variables les plus utiles et entrainer une nouvelle fois le modèle.

Algorithme 1 : Calcul de l'importance de chaque variable

$ref \leftarrow$ score de prédiction du modèle de base;
$n_{\text{variables}} \leftarrow$ nombre de variables;
pour $i \leftarrow 1$ ***jusqu'à*** $n_{variables}$ **faire**
    $score_i \leftarrow$ score de prédiction du modèle avec la variable i mélangée au hasard;
    $p_i \leftarrow ref - score_i$;
retourner p;

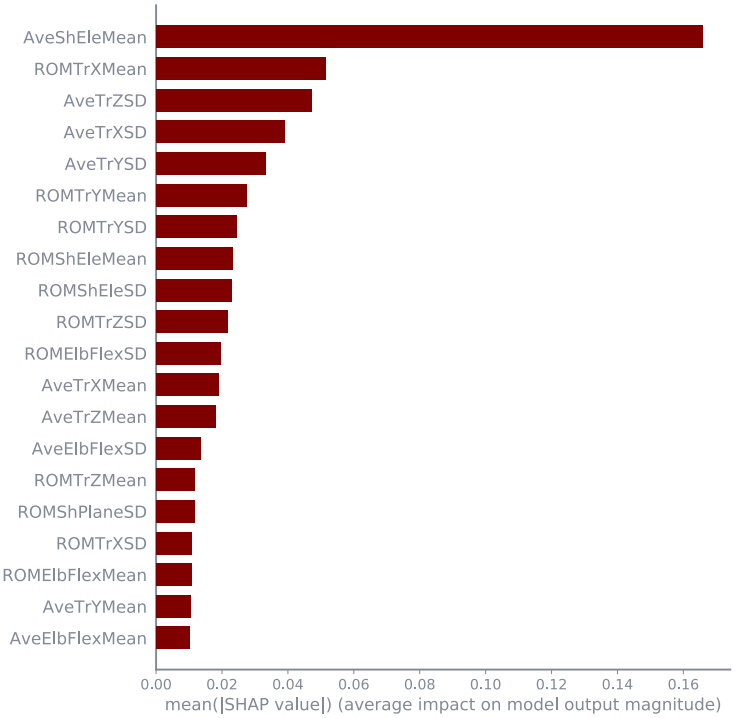

Figure 5 – Importance de chacune des variables sur la prédiction du modèle calculée à partir de la valeur absolue des valeurs *shap*.

Lors de sa deuxième itération, notre modèle prédit la fatigue avec une précision de 86% et un rappel de 85%. L'importance des variables calculées sur le second modèle (Fig. 6, panneau A) nous indique que l'angle d'élévation moyen (AveShEleMean) est la variable la plus importante pour la prédiction de l'état de fatigue. Une valeur élevée diminue la probabilité d'être identifié comme fatigué, et inversement (Fig. 6, panneau B). Le sexe semble peu influencer la prédiction (Fig. 6).

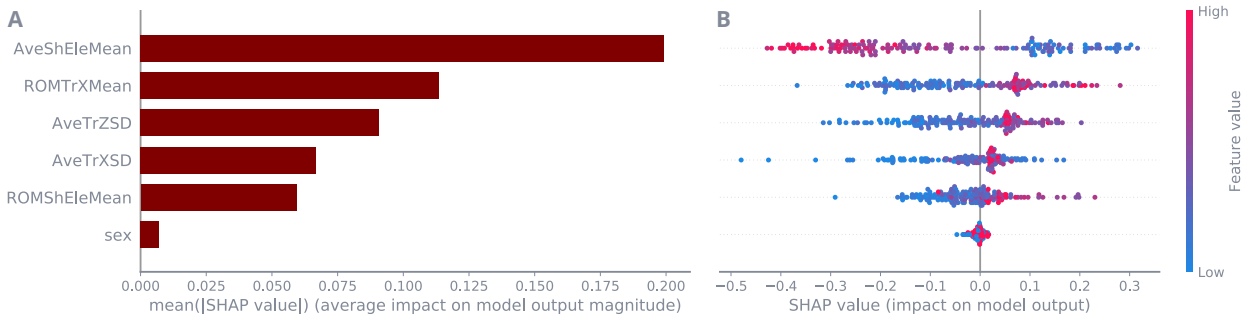

Figure 6 – Importance de chacune des variables sur la prédiction du modèle calculée à partir de la moyenne des valeurs absolues des valeurs *shap* (A) et la valeur *shap* où chaque point est une observation (B).

Un regroupement supervisé (Fig. 7) nous permet de réduire la dimension des données et de les visualiser sur une figure à deux dimensions. Celle-ci nous confirme que l'augmentation de l'angle d'élévation du coude diminue notre probabilité d'être étiqueté fatigué, et inversement pour l'amplitude de mouvement du tronc (ROMTrXMean).

Figure 7 – Regroupement supervisé en analyse de composantes principales (A) et en t-SNE (B). La couleur de la première colonne correspond à la prédiction du modèle, la seconde à la variable AveShEleMean et la troisième à la variable ROMTrXMean.

Pour mieux comprendre la prédiction du modèle, il est aussi possible de visualiser le chemin moyen que prend la prédiction d'une observation particulière parmi les l'ensemble des arbres de décision (Fig. 8). Par exemple, la première observation du jeu de données (Fig. 8, panneau A) est étiquetée comme « fatiguée » surtout en raison d'un angle d'élévation de coude faible (AveShEleMean = 77.9 deg) tandis que pour la deuxième (Fig. 8, panneau B) c'est un faible écart-type d'inclinaison du tronc (AveTrZSD = 0.23 deg) qui influence la prédiction de l'étiquette « non fatiguée ».

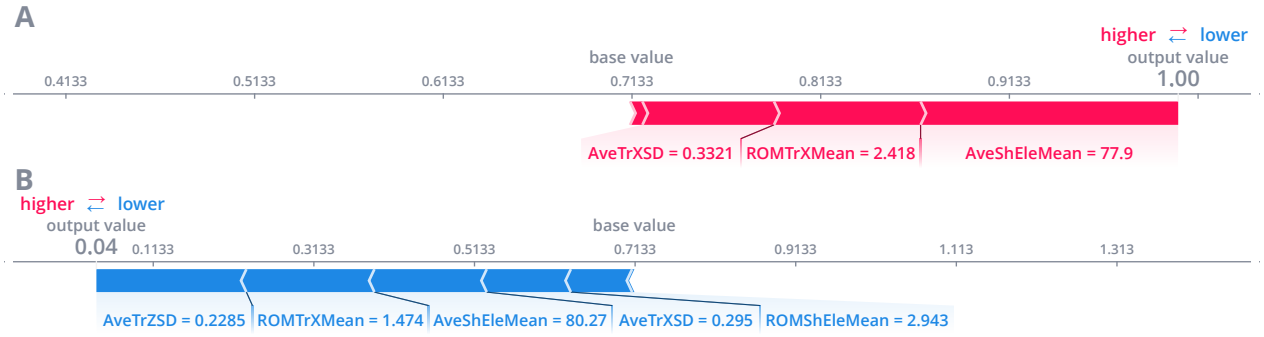

Figure 8 – Chemin moyen parmi les l'ensemble des arbres de décision que prend la prédiction d'une observation particulière étiquetée comme fatiguée (A) et non fatiguée (B).

Enfin, pour visualiser comment le changement d'une variable change la prédiction du modèle, il est possible de tracer la valeur d'une variable par rapport à sa valeur *shap*. Comme la valeur *shap* représente l'impact de la modification d'une variable sur la prédiction du modèle, la figure 9 représente la modification de la fatigue au fur et à mesure des modifications de l'angle d'élévation du coude. La couleur représente l'interaction avec la variable présentant le plus d'effet d'interaction (ROMTrXMean). Dans ce cas, la coloration de la variable ROMTrYMean souligne que AveShEleMean a moins d'impact sur la fatigue lorsque ROMTrYMean diminue et inversement (Fig 9).

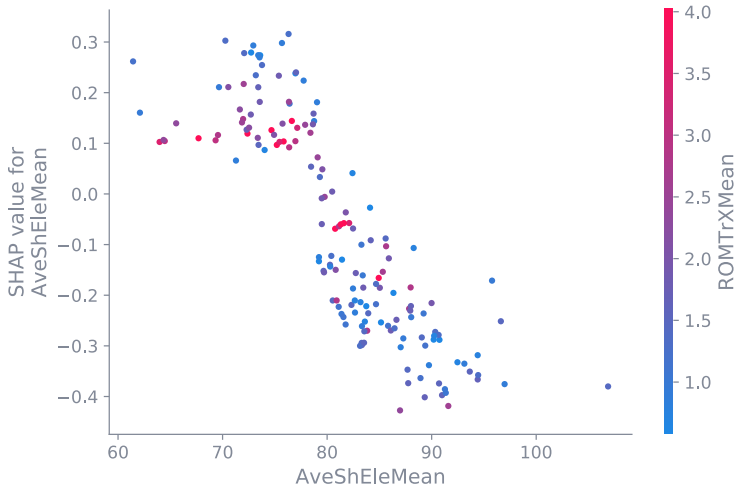

Figure 9 – Valeur de la variable AveShEleMean par rapport à sa valeur *shap*. Chaque point est une observation. La couleur représente l'interaction avec la variable ROMTrXMean.

Ces différentes interprétations nous ont permis de répondre à quatre questions: (1) dans quelle mesure avons-nous confiance en nos prévisions, (2) quelle est l'importance de nos variables, (3) quel est le chemin de l'arbre de décision et (4) comment la variable cible est-elle liée aux variables importantes. Les réponses à ces questions semblent centrales pour la compréhension de la prédiction, mais aussi de l'impact de la fatigue sur les variables cinématique. La classe de modèle utilisé (*i.e.* arbres de décision) est intrinsèquement interprétable et la prédiction est plus facile à imaginer qu'avec un modèle linéaire. Alors que beaucoup d'utilisateurs pensent que les modèles simples sont nécessairement plus interprétables, les récents progrès en apprentissage automatique sur l'interprétation des modèles complexes (Ribeiro, Singh, et Guestrin 2016; Štrumbelj et Kononenko 2014; Datta, Sen, et Zick 2016; Lipovetsky et Conklin 2001; Lundberg et Lee 2017; Lundberg, Erion, et Lee 2018) rendent ces derniers davantage interprétables *et* performants (Tab. 1).

Table 1 – Comparaison de l'interprétabilité et la performance des modèles simples (*e.g.* modèles linéaires) et complexes (*e.g.* modèle d'ensemble).

	Interprétable	Performant
Modèle simple	✓	×
Modèle complexe	✓	✓

Il convient cependant de noter que ces interprétations ne font qu'expliquer le fonctionnement du modèle prédictif et pas nécessairement le fonctionnement de la réalité. Étant donné que le modèle est formé à partir de données d'observation, il ne s'agit pas nécessairement d'un modèle causal. Enfin, les solutions produites par un quelconque modèle (numérique ou théorique) doivent être considérées avec scepticisme.

3.2 Limiter le *p-hacking*

Dans la majorité des études scientifiques, la signification statistique est définie comme $p < 0.05$, c'est-à-dire qu'une différence observée sur vingt serait due au hasard. Bien que ce seuil semble être raisonnable, il est souvent l'objet d'un intense débat au sein de la communauté scientifique, car les données sont souvent sélectionnées ou manipulées pour arriver à ce chiffre. Cette pratique, aussi nommée *p-hacking*, est à l'origine d'une crise de reproductibilité qui touche la plupart des domaines (Simmons, Nelson, et Simonsohn 2011). Les résultats de nombreuses études sont difficiles, voire impossibles à reproduire au cours d'études indépendantes (Open Science Collaboration 2015).

Le compromis biais-variance que nous avons introduit dans la section précédente peut servir à comprendre l'enjeu du *p-hacking*. Du point de vue de l'apprentissage automatique, le *p-hacking* peut être perçu comme une forme d'*overfitting*. Une méthode de recherche qui encouragera une analyse de données préalable et des pratiques de *p-hacking* produira ainsi une forte variance et un faible biais. Que ce soit un modèle numérique ou un scientifique en biomécanique qui conclue sur un phénomène, la problématique est la même: distinguer le signal du bruit. Dans son excellent livre, « *The Signal and the Noise* », le statisticien Nate Silver rapporte:

> « distinguer le signal du bruit nécessite à la fois des connaissances scientifiques et la connaissance de soi: la sérénité d'accepter les choses que nous ne pouvons pas prédire, le courage de prédire ce que nous pouvons et la sagesse de connaitre la différence. »
> – Silver (2012)

Le signal est la vérité et le bruit nous en éloigne. Une analyse *p-hackée* semblera raconter une histoire intéressante, mais bien souvent celle-ci ne sera pas généralisable. Pour lutter contre cette sorte d'*overfitting*, certains laboratoires encouragent la mise en place de procédures standardisées pour communiquer des résultats en biomécanique. Par exemple, l'équipe de Scott Delp du laboratoire de l'Université de Stanford a publié un processus général de vérification et de validation appliqué aux modèles musculo-squelettiques (Hicks et al. 2015). Ce dernier comprend la formulation minutieuse d'une question et de méthodes de recherche, les étapes traditionnelles de vérification et de validation,

la documentation et enfin le partage des résultats pour l'utilisation et la vérification par d'autres chercheurs.

Une autre source d'*overfitting* en biomécanique est la publication de résultats issue d'un nombre de participants réduit. Il est préférable de privilégier de petits effets statistiques sur un grand nombre de participants plutôt que des grands effets sur des petits échantillons. Cependant, la collecte de données est couteuse et il est parfois difficile d'augmenter son échantillon. Une solution serait d'encourager les regroupements de laboratoires et les analyses sur des ensembles de données publiques et disponibles pour la communauté. Le *Grand Challenge to Predict Knee Loads* (Fregly et al. 2012) est un exemple de jeu de données standard de référence accessible au public. En plus d'un ensemble complet de données anthropométriques et de données cinématiques, le Grand Challenge comprend également un problème clair axé sur la validation: la prédiction des charges internes appliquées sur l'articulation du genou. Ce genre d'initiative a joué un rôle central en apprentissage automatique. Avant la publication d'un modèle, les scientifiques évaluent leurs performances sur des données de références, comme les jeux de données MNIST (LeCun, Cortes, et Burges 2010) et ImageNet (Deng et al. 2009).

Continuons sur l'exemple de la modélisation musculo-squelettique. Dans ce domaine, les études prennent rarement l'initiative de vérifier et valider leurs modèles. Une des conséquences est l'incapacité à prédire le fonctionnement musculo-squelettique qu'ils modélisent lorsqu'ils sont testés sur des données indépendantes. Et cela, malgré une formulation théorique ou mathématique attrayante. Pour autant, la validation croisée que nous avons introduite dans la section précédente est absente du domaine. L'adoption de cette technique en biomécanique contribuerait pourtant à améliorer la fiabilité des résultats rapportés. Plutôt que de sélectionner une procédure parce qu'elle permet d'atteindre $p < 0.05$ sur l'ensemble des données, cette approche consisterait à établir une liste de procédures candidates et de les évaluer en validation croisée. Une fois que le chercheur est prêt à rapporter les résultats, il pourra alors évaluer son modèle sur un ensemble de tests indépendant et publier les résultats—significatifs ou non. Bien qu'elle n'est pas incitée par les journaux scientifiques et les organismes subventionnaires, la publication de résultats non significatifs est importante pour la communauté scientifique. Si tous les degrés de liberté du scientifique sont inclus et que le critère de validation croisée est respecté, alors le scientifique pourra essayer toutes les procédures qu'il souhaite car la variance de son modèle sera exposée lors de l'évaluation finale. Bien que la validation croisée ne permettra pas de résoudre complètement le problème du *p-hacking*, c'est une première étape vers une science reproductible et fiable.

Conclusion

Parce qu'une bonne explication du fonctionnement musculo-squelettique implique nécessairement une bonne prédiction de son comportement futur, nous avons suggéré d'utiliser et de s'inspirer de l'apprentissage automatique pour construire et évaluer des modèles biomécaniques. La prédiction ne devrait pas être l'objectif principal du domaine, mais elle devrait au moins être incluse dans le processus d'explication. Dans cette section, nous avons suggéré d'utiliser les modèles prédictifs pour mieux comprendre ses données et interpréter des relations entre variables. Le problème du p-hacking se produit lorsqu' un chercheur influence le processus de collecte de données ou les analyses effectuées afin de produire un résultat statistiquement significatif. Nous avons proposé d'utiliser la validation croisée ainsi que d'encourager la publication de données et les regroupements de laboratoires pour réduire l'effet du p-hacking.

4 Bibliographie

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, et Wojciech Samek. 2015. « On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation ». *PLoS One* 10 (7): e0130140.

Banko, Michele, et Eric Brill. 2001. « Scaling to Very Very Large Corpora for Natural Language Disambiguation ». In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 26-33. ACL '01. Stroudsburg, PA, USA: Association for Computational Linguistics.

Browne, M W. 2000. « Cross-Validation Methods ». *J. Math. Psychol.* 44 (1): 108-32.

Chen, Tianqi, et Carlos Guestrin. 2016. « XGBoost: A Scalable Tree Boosting System ». In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-94. KDD '16. New York, NY, USA: ACM.

Datta, A, S Sen, et Y Zick. 2016. « Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems ». In *2016 IEEE Symposium on Security and Privacy (SP)*, 598-617.

Deng, J, W Dong, R Socher, L Li, Kai Li, et Li Fei-Fei. 2009. « ImageNet: A large-scale hierarchical image database ». In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248-55.

Domingos, Pedro. 2012. « A Few Useful Things to Know About Machine Learning ». *Commun. ACM* 55 (10). New York, NY, USA: ACM: 78-87.

Domingos, Pedro, et Michael Pazzani. 1997. « On the Optimality of the Simple Bayesian Classifier under Zero-One Loss ». *Mach. Learn.* 29 (2): 103-30.

Fregly, Benjamin J, Thor F Besier, David G Lloyd, Scott L Delp, Scott A Banks, Marcus G Pandy, et Darryl D D'Lima. 2012. « Grand challenge competition to predict in vivo knee loads ». *J. Orthop. Res.* 30 (4): 503-13.

Hicks, Jennifer L, Thomas K Uchida, Ajay Seth, Apoorva Rajagopal, et Scott L Delp. 2015. « Is my model good enough? Best practices for verification and validation of musculoskeletal models and simulations of movement ». *J. Biomech. Eng.* 137 (2): 020905.

Hinton, G E, et R R Salakhutdinov. 2006. « Reducing the dimensionality of data with neural networks ». *Science* 313 (5786): 504-7.

John Lu, Z Q. 2010. « The Elements of Statistical Learning: Data Mining, Inference, and Prediction ». *J. R. Stat. Soc. Ser. A Stat. Soc.* 173 (3): 693-94.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, et Ziad Obermeyer. 2015. « Prediction Policy Problems ». *Am. Econ. Rev.* 105 (5): 491-95.

Kononenko, I. 2001. « Machine learning for medical diagnosis: history, state of the art and perspective ». *Artif. Intell. Med.* 23 (1): 89-109.

Kremer, Jan, Kristoffer Stensbo-Smidt, Fabian Gieseke, Kim Steenstrup Pedersen, et Christian Igel. 2017. « Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy », avril.

LeCun, Y, C Cortes, et C J Burges. 2010. « MNIST handwritten digit database ». *[Online]. Available: http://yann. lecun. com/exdb/mnist*.

Lipovetsky, Stan, et Michael Conklin. 2001. « Analysis of regression in game theory approach ». *Appl. Stoch. Mod. Data Anal.* 17 (4): 319-30.

Lundberg, Scott M, Gabriel G Erion, et Su-In Lee. 2018. « Consistent Individualized Feature

Attribution for Tree Ensembles », février.

Lundberg, Scott M, et Su-In Lee. 2017. « A Unified Approach to Interpreting Model Predictions ». In *Advances in Neural Information Processing Systems 30*, édité par I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, et R Garnett, 4765-74. Curran Associates, Inc.

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, et Angela Byers. 2011. « Big data: The next frontier for innovation, competition, and productivity ».

Mitchell, Tom M, et Others. 1997. « Machine learning. 1997 ». *Burr Ridge, IL: McGraw Hill* 45 (37): 870-77.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, et al. 2015. « Human-level control through deep reinforcement learning ». *Nature* 518 (7540): 529-33.

Obermeyer, Ziad, et Ezekiel J Emanuel. 2016. « Predicting the Future — Big Data, Machine Learning, and Clinical Medicine ». *N. Engl. J. Med.* 375 (13). Massachusetts Medical Society: 1216-9.

Open Science Collaboration. 2015. « Estimating the reproducibility of psychological science ». *Science* 349 (6251). American Association for the Advancement of Science: aac4716.

Ribeiro, Marco Tulio, Sameer Singh, et Carlos Guestrin. 2016. « Model-Agnostic Interpretability of Machine Learning », juin.

Shao, Jun. 1993. « Linear Model Selection by Cross-validation ». *J. Am. Stat. Assoc.* 88 (422). Taylor & Francis: 486-94.

Shear, Benjamin R, et Bruno D Zumbo. 2013. « False Positives in Multiple Regression: Unanticipated Consequences of Measurement Error in the Predictor Variables ». *Educ. Psychol. Meas.* 73 (5). SAGE Publications Inc: 733-56.

Shrikumar, Avanti, Peyton Greenside, et Anshul Kundaje. 2017. « Learning Important Features Through Propagating Activation Differences », avril.

Silver, Nate. 2012. *The signal and the noise: why so many predictions fail–but some don't.* Penguin.

Simmons, Joseph P, Leif D Nelson, et Uri Simonsohn. 2011. « False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant ». *Psychol.*

Sci. 22 (11): 1359-66.

Štrumbelj, Erik, et Igor Kononenko. 2014. « Explaining prediction models and individual predictions with feature contributions ». *Knowl. Inf. Syst.* 41 (3): 647-65.

Westfall, Jacob, et Tal Yarkoni. 2016. « Statistically Controlling for Confounding Constructs Is Harder than You Think ». *PLoS One* 11 (3): e0152719.

Wiering, M A. 2000. « Multi-agent reinforcement learning for traffic light control ». In *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000)*, 1151-8. dspace.library.uu.nl.