Les deux types d'approches classiques centrées des données

Cet article traite les deux types d'approches dans l'utilisation de la modélisation statistique centrées données, l'une se basant sur l'hypothèse que les données sont générées par un modèle stochastique (modélisation historique) et l'autre de type modélisation algorithmique (machine learning).

L'objectif de l'article est d'amener la communauté des statisticiens, à travers des exemples concrets pratiques dont l'auteur Leo Breiman a été amené à résoudre, à se pencher de plus en plus sur l'importance de la modélisation algorithmique au niveau du traitement des données et de ses retombées scientifiques et pratiques enregistrées récemment en les incitant à travailler sur de nouveaux problèmes porteurs dont la modélisation algorithmique au niveau de l'apprentissage machine a fait ses preuves en termes informatif et de précision comme alternatives à la modélisation des données. Le but n'est pas l'interprétabilité, mais bien l'information précise. Plus la précision prédictive est grande, plus elle est associée à de l'information crédible du mécanisme des données inhérentes (conclusions claires). Les modèles algorithmiques peuvent donner une précision de prédiction meilleure que celle des modèles de données, et fournir en conséquence des informations plus crédibles.

Il en ressort qu'il faut en vue de trouver de bonnes solutions, de s'intéresser davantage aux données avant de s'y mettre dans la modélisation, de tenir compte des deux types de modélisation statistique et aussi seuls les tests basés sur la précision prédictive sont des indices corrects pour la validation du modèle. Pour résoudre les problèmes centrés données, un ensemble d'outils plus large est nécessaire où les données et le problème guident la solution, d'où l'importance de la modélisation algorithmique. Les problèmes actuels à complexité croissante (tels en sciences et commerce) utilisant l'apport informatique en stockage et manipulation de grandes base de données (big data) sont mieux et plus pris en charge par des non statisticiens. Le développement des méthodes algorithmiques s'est fait en dehors des statisticiens.

Modélisation des données	Modélisation algorithmique
Principe : modélisation stochastique (traditionnelle) des données où les valeurs des paramètres du modèle sont estimées conjointement par les données et le modèle, ensuite utilisées comme information et/ou prédiction. La validation de ce type de modèle, tel la régression linéaire, la régression logistique et le modèle Cox, se fait selon des tests de conformité et le contrôle résiduel. Différence : portant beaucoup plus sur le mécanisme et la construction du modèle stochastique et sa paramétrisation. Avantages : Modèles mathématiques théoriques génériques (statistiques) rigoureux avec traçabilité, lisibilité (algo. Viterbi). -Précision et information : pour des modèles paramétriques simples en présence de données générées par des systèmes complexes (données médicales, financières, etc.). Inconvénients : -Théorie non assez	Principe : les entrées/sorties prédites sont liées par une fonction/algorithme. La validation de ce type de modèle, tel les arbres de décision et les réseaux neuronaux, se mesure par la précision prédictive (système boite noir). Différence : portant sur la structure de la relation entre entrées/sorties, avec plus d'information crédible que sur les modèles de données . Avantages : Prédiction de problèmes complexes et réels où la modélisation de données n'est pas évidente : Reconnaissance de la parole, reconnaissance d'images, prédiction de séries temporelles non linéaires, reconnaissance de l'écrit, prédiction dans les marchés financiers, etc. -Théorie de la modélisation algorithmique : SVM (Vapnik) basée sur la généralisation de l'erreur qui dépend aussi de la capacité de l'algorithme. Approche plus précise que RN en classification et régression. -La multiplicité de bons modèles, plusieurs

significative (ne reflétant pas suffisamment la réalité des données d'où conclusions scientifiques obtenus restant à interpréter et confirmer). - Pas de liens effectifs sur le type et nature de données et le degré de similitude entre les données et le modèle. -Modèles statistiques difficilement interprétables : avec conclusions non suffisamment fiables. -L'arbitraire dans l'analyse résiduelle : c'est au modèle de s'autodéterminer par rapport aux données réelles tout en restant le plus possible linéaire (régression linéaire). Les données réelles observables ne peuvent-elles être engendrées par un modèle supposé généralement être linéaire. -La validation des modèles de données selon l'analyse résiduelle non encourageante, surtout au-delà de 4 à 5 dimensions. -<u>Multiplicité des modèles :</u> plusieurs bons modèles (autour de 5% de précision) peuvent représenter un même ensemble de données. -<u>Aucun modèle compréhensif</u> capable d'accorder plus d'importance à l'interprétation des résultats surtout pour les modèles de survie (Cox modèles). <u>-L'estimation de la précision prédictive est biaisée en présence de modèles ayant plusieurs paramètres</u> : d'où la validation croisée pour y remédier surtout le cas de larges bases de données. -Pas de standards d'estimation et de comparaison de modèles à partir de la précision prédictive, contrairement à l'apprentissage machine. <u>L'apparition de modèles de données plus complexes :</u> tel l'hybridation entre Réseaux Bayésiens et Mcarlo.Cmarkov d'où tentative dans la représentation du mécanisme de la nature des données (identification de gènes : ADN, génome). Techniques stochastiques à base de distributions de densité de probabilité – pdf (GMM), et aussi la logique floue.	solutions f(x) pour le même taux d'erreur. Instabilité ou légère perturbation dans les données ou la construction du modèle, résultent la multiplicité dans les modèles. -Compromis entre simplicité (l'interprétabilité) et précision qui nécessite généralement des méthodes de prédiction plus complexes d'où difficilement interprétables de par leurs conclusions (exemple entre approches de décision par forêts (bon prédicteur mais complexe) et arbres (interprétation simple mais moins bon prédicteur). -La dimensionnalité entre malédiction (Bellman) et pourquoi pas bénédiction : généralement, on procède par élagage de variables les moins pertinentes afin de réduire la dimensionnalité, car on cherche des paramètres ou fonctions de variables de prédiction qui contiennent le plus d'information tout en étant le moins nombreux possible, rendant ainsi le système moins complexe. En principe, plus il y a de variables de prédiction, est plus il y a d'information. Algorithmes à métaphore biologique (soft-computing), méta-heuristiques : R. Neurones, A. Génétiques, R. Immunitaires, R. Neurones impulsionnels, deeplearning, Et autres algorithmes de classification (clustering) : Kppv, k-means, quantification vectorielle, et la programmation dynamique-DTW et bien sûr les SVM. <u>Inconvénients :</u> la plupart des algorithmes associés sont des méta-heuristiques, non démontrables mathématiquement (boites noires) cherchant à simuler la nature (monde biologique). Or plus on s'approche de la nature, est plus on se rend compte qu'on se rapproche moins. La nature se laisse approcher qu'en termes de probabilité. <u>Conclusion :</u> les méthodes de la modélisation algorithmique sont des contributions importantes comme utilitaires pour les statisticiens, en termes de prédiction et d'information, n'empêche que leur apport, encore palpable sur le plan scientifique/académique, est incontournable.