Analyse de « Statistical Modeling : The Two Cultures »

Lorsque que l'on désire prédire une sortie Y en fonction de l'entrée X, il est nécessaire d'étudier les données historiques. Dans cet article, Leo Breiman présente et oppose deux « écoles » d'analyse des données et cherche à montrer comment l'ancienne école tend à être dépassée par la nouvelle. L'ancienne école, appelée « The Data Modeling Culture » est celle qui se base uniquement sur les théories statistiques et cherche à utiliser les modèles de prédiction et leurs propriétés pour l'analyse. Pour cette école, il faut étudier la loi, le mécanisme sous-jacent entre l'entrée et la sortie. Pour la nouvelle école appelée « The Algorithm Modeling Culture », la loi est par essence inconnue et complexe. La relation à trouver entre l'entrée et la sortie passe nécessairement par un algorithme qui doit chercher à reproduire le phénomène provoqué par la « nature » et donner la meilleure précision possible. Aussi compliqué et peu transparent soit l'algorithme.

1. The Data Modeling Culture

Le principe est d'utiliser les méthodes de statistiques connues pour prédire la sortie Y. La méthodologie est la suivante : faire l'hypothèse de l'existence d'une loi (par exemple une régression linéaire), estimer les paramètres (maximum de vraisemblance), vérifier la validité de la loi (goodness-of-fit) et à partir des propriétés de la loi aboutir à des conclusions quant au mécanisme qui produit Y en fonction de X (intervalle de confiance)

L'avantages de cette école réside dans le fait que les lois et les méthodologies sont clairement définies. Il est facile d'arriver à des conclusions (accepter ou rejeter le test à 5%) et surtout, il est facile d'interpréter la loi. Pour l'utilisation de la régression linéaire : on peut regarder la valeur de R^2 pour connaître la qualité de notre régression et regarder la p-value ainsi que la valeur des coefficients pour analyser l'impact des facteurs. Il est facile d'arriver à des conclusions. Cette approche est surtout connue de tous.

Les inconvénients sont toutefois nombreux. L'hypothèse qu'il existe une loi statistique décrivant le modèle est trop forte pour certains cas de figures. Pourquoi penser qu'un phénomène très complexe va pouvoir être décrit par une régression linéaire ? C'est pourtant ce qui est annoncé et n'est jamais remis en cause par l'immense majorité des partisans de la première école. Ceci provoque des conclusions hâtives avec certains cas de figures : pour une régression linéaire regarder R^2 alors qu'on a une grande quantité de variables n'est pas cohérent. Regarder la p-value d'un coefficient alors que l'on étudie un groupe très hétérogène n'est pas intelligent pour regarder son influence. La démarche a de plus des limites flagrantes : le test de goodness-of-fit qui est censé vérifier que le modèle est bien ajusté aux données, ne rejette l'hypothèse qu'en cas d'erreur très grande. Et ne marche dans des cas plus complexes (termes non linéaires). Il en va de même pour l'analyse des résidus. De surcroît, l'analyse goodness-of-fit donne une réponse binaire au problème : soit le test passe et l'on admet toutes les conclusions/conséquences, soit le test est rejeté. La cross-validation (le fait d'extraire un troisième set de données pour ajuster les hyper-paramètres) n'est pas beaucoup utilisée dans les démarches (alors qu'elle l'est toujours par la deuxième école). Lorsque l'on recueille des données expérimentales (mesure d'un appareil physique par exemple), il est trop simpliste de penser par exemple, que toutes les variables vont être normalement distribuées. C'est pourtant ce qui est souvent fait. C'est en effet le cas du projet d'analyse météorologique évoqué par Breiman : le projet fut guidé par le choix du modèle « régressions linéaire » contrairement au « data driven » ce qui conduisit le projet à sa perte.

Breiman appelle donc à une démarche moins naïve et préfère la précision des algorithmes.

2. Différences

Face aux problèmes rencontrés, Breiman préconise donc une nouvelle approche et pointe trois principes que les statisticiens de l'ancienne école n'ont pas nécessairement à l'esprit :

- le fait qu'il y ait différents modèles possibles. Ce principe peut se voir sur des régressions linéaires mais aussi sur des neural nets ou des arbres de décision. Il existe une instabilité des modèles : en changeant un peu du trainings set, le modèle peut beaucoup changer.
- La précision (donc la qualité de la prédiction) et la simplicité d'un modèle (donc son interprétabilité) sont deux choses qui ne vont pas ensemble. Il est toujours nécessaire de trouver un compromis pour prédire une sortie.
- Il semble toujours préférable de choisir un nombre restreint de variables explicatives pour la prédiction. Cela rend le modèle plus simple et limite les problèmes d'overfit. Toutefois, chaque variables ou même fonction de variable contient une information que l'on aimerait capturer dans le modèle.

La deuxième école se différencie donc autour de trois grands points :

Data Modeling	Algorithm Modeling
Propriétés mathématiques de la loi stats	Mécanismes complexes de l'algorithme
Interprétabilité	Précision
Mécanisme compréhensible	Boite noire

Les critiques successives de l'article, notamment celle de Cox et Efron, mettent avant tout le doigt sur une différence d'approche : pour la nouvelle école, seule importe la qualité de la prédiction alors que, pour l'ancienne, pouvoir interpréter le phénomène à l'aide d'une loi est ce qui est le plus précieux. Les anciens ont peur du phénomène « boite noire » alors que les nouveaux se veulent très exigeants avec la précision du modèle.

3. The Algorithm Modeling Culture

Le principe est d'utiliser les algorithmes nés avec l'essor de l'informatique, que l'on décrit comme du « machine learning ». On parle de Réseau de neurones, d'arbres de décisions et son extension le random forest. Le critère ultime est la précision obtenue sur le « test set ». La démarche est la suivante : on sépare les données historiques en trois set de variables : à hyper-paramètres fixés, on optimise les paramètres sur le premier set, avec le deuxième set, on choisit les hyper-paramètres et enfin on teste la précision de notre algorithme sur notre troisième set. Les algorithmes reposent sur une succession de calculs automatiques et prédit après des étapes complexes la valeur. Par exemple, pour prédire une classe (patient malade ou non), on choisira un modèle (un arbre de décision), on testera le modèle en faisant avec varier les hyper-paramètres (nombre de couches) et alors chaque nœud nécessitera un calcul et fournira pas-à-pas la prédiction.

L'avantage principale réside donc dans la précision. Que ce soit sur le score du test-set mais aussi sur les problèmes d'overfitting. En séparant trois sets de données et en faisnat attention au nombre de variables, c'esSur la quasi-totalité des cas de figures, les algorithmes donneront un taux d'erreur bien inférieur à celui donné par les modèles « Data ». Que ce soit pour des prédictions qualitatives, quantitatives. De plus, ces algorithmes ouvrent la voie à des problèmes plus complexes : du clustering et d'autres problèmes d'apprentissage non-supervisée. L'approche se veut plus honnête car elle ne cherche pas à calquer une loi sur un mécanisme complexe.

L'inconvénient majeure est le manque d'interprétabilité. Certes, la prédiction est meilleure mais il reste toutefois difficile de tirer des conclusions, que ce soit d'un point de vue économique ou scientifique. Comment adapter le modèle ou comment expliquer un phénomène avec une boite noire ? De plus, les modèles cachent parfois une complexité et font intervenir de nombreux paramètres même lorsque les données ne sont pas si nombreuses. Il semble stupide d'avoir 1000 paramètres lorsque l'on dispose de 100 exemples. De plus, cette complexité provoque une instabilité des modèles : en changeant un exemple dans le training set, ceci va perturber beaucoup les paramètres du modèle. De cette instabilité nait le besoin de recalibrer très souvent le modèle. Contrairement au « Data Modeling », la compréhension du modèle peu s'avérer très délicate.

4. Exemples d'applications

Dans le cas où l'on a un nombre limité de variables explicatives et que l'ensemble à analyser est homogène, il semble plus raisonnable d'utiliser un modèle Data. En effet, prenons l'exemple de donnés X de 1 dimension et essayons une régression linéaire pour prédire Y. Tout d'abord, on pourra vérifier visuellement que la régression est bien ajustée ou l'on pourra décider de rajouter un terme quadratique. Si l'on reprend l'exemple du cours (=la parabole), la visualisation rend la régression linéaire très pertinente. Ainsi, dans ce cas simple, utiliser un modèle simple nous permettra de facilement interpréter les résultats.

Pour un exemple où X est de dimension élevé, la sélection de variables et l'ajout de terme non linéaire sont plus compliqués. Du coup, un random Forest avec classifier permet de réellement gagner du temps et d'améliorer la prédiction. Cela se produit lorsque le problème compte des variables macroscopiques et microscopiques et que leur sélection est plus difficile.

Analyse de « Model Selection in Data Analysis Competitions »

Les cinq grands conseils à suivre pour prédire sont :

1. *Feature engineering is the most important part of predictive machine learning.* Trouver et construire les bonnes variables pour faire marcher l'algorithme est de loin la chose la plus importante. Etant donné qu'il y a une infinité de combinaisons possibles de variables à sélectionner (termes linéaires, quadratiques, etc.), il peut être intéressant de tester de nombreux scenarii.

2. *Overfitting to the leaderboard is a real issue*. Les compétitions Kaggles mettent à disposition plusieurs set de données pour tester. Souvent les compétiteurs ajustent trop leur modèle au test public et overfit. Le taux d'erreur sur le privé, l'erreur devient alors assez élevé. Introduire des set de cross-validation permet d'éviter cet écueil.

3. Simple models can get you very far. Malgré leur simplicité, certains modèles peuvent prédire de manière très précise une sortie. Les modèles sont de plus simples à interpréter, modifier et peuvent donner des informations très précieuses. Il est préférable de ne pas se ruer vers un modèle trop compliqué dès le début.

4. Ensembling is a winning strategy. Combiner différents modèles sera toujours efficace. Ce sera même souvent plus efficace que de vouloir perfectionner un modèle unique et d'y ajouter pleins de variables.

5. Predicting the right thing is important. Avant de commencer et de vouloir prédire la sortie demandée, il peut être très utile de vouloir une sortie dépendante. Certaines techniques audacieuses peuvent être très fructueuses.