Loïc Herbelot

Lecture de l'article "Statistical Modeling: The Two Cultures" de L. Breiman. Cet article donne les clefs pour comprendre les deux types d'approche classiques centrées données: la "**data modelling culture**" qui correspond aux approches 'modélisation' historiques (et sur lesquelles sont par exemple basés les modèles de type météo), et l'approche ici appelée "*Algorithmic Modeling Culture"* qui correspond grosso modo à l'approche **machine learning** et qui tend à prendre une place de plus importante dans le traitement de données.

Travail à rendre: Vous ferez un résumé (1 à 2 pages) des principes, différences, avantages et inconvénients des deux types d'approches. Vous donnerez des exemples d'applications pour lesquelles vous pensez que l'une ou l'autre des approches et la plus appropriée.

Pour extraire de l'information à partir de données brutes, on peut distinguer deux approches distinctes : l'approche orientée modélisation statistique : « Data Modeling », plutôt théorique, et l'approche algorithmique : « Algorithmic Modeling », plus pratique.

L'approche « Data Modeling » consiste à trouver le modèle probabiliste qui pourrait avoir généré les données que l'on a. Par exemple, dans le cas d'une régression, on pourrait supposer que la variable cible soit le résultat d'une combinaison linéaire des variables explicatives, perturbée par un bruit, supposé normal, centré et réduit. Cela permet de construire un modèle théorique assez simple, de trouver ensuite quelles sont les relations entre les variables explicatives et la variable cible pour déterminer quelles variables sont importantes ou au contraire lesquelles n'ont aucune influence. À partir d'un modèle statistique on peut aussi dériver beaucoup de résultats, comme trouver un estimateur efficace, trouver son biais, sa variance, déterminer des intervalle de confiance… On peut aussi tester la validité de certaines hypothèses, l'utilisation de modèles théoriques est plutôt efficace dans l'étude des effets de médicaments placebo, car on a à notre disposition des intervalles de confiance qui nous aident à tirer des conclusions sur l'efficacité de produits. Le problème de cette approche est qu'elle est assez théorique, il y a beaucoup de façons d'obtenir des résultats erronés. Le choix du modèle notamment est crucial, il ne faut pas sortir un modèle de nulle part, car une fois que l'on a trouvé des paramètres qui font que notre modèle « colle » à la réalité, il est difficile de douter de la validité de son modèle, même si le modèle choisi n'est pas le bon.

L'approche algorithmique est beaucoup plus pratique et ne s'intéresse pas trop à la théorie qui régit les données que l'on a. On pourrait penser que cette méthode est un peu du bricolage et qu'on essaie juste de balancer des algorithmes et différents paramètres au hasard

jusqu'à avoir un résultat satisfaisant, sans rien comprendre aux données, mais ce n'est pas le cas. Des algorithmes simples permettent d'obtenir de très bons résultats, mais aussi d'avoir une idée de l'influence des variables explicatives sur le résultat. Par exemple des arbres de décision permettent de visualiser très facilement l'importance de certaines variables sur le résultat. On peut aussi utiliser des 'Random Forest' afin d'estimer l'importance des variables pour sélectionner les variables les plus influentes et éliminer les variables inutiles. L'utilisation de méthodes algorithmiques a l'avantage d'être très rapide à mettre en œuvre et d'être très facilement adaptable. Cependant, le but dans ce type de problème est souvent de pouvoir à la fois comprendre les relations entre les variables et de pouvoir fournir des estimations précises. Et des modèles simples type k plus proches voisins ou arbres de décisions, qui sont très faciles à comprendre, ne fournissent pas d'aussi bons résultats que des réseaux de neurones ou des random forest, qui sont des boîtes noires. Ces méthodes algorithmiques sont très utilisées dans le monde des sciences de données et fournissent de très bons résultats, notamment dans des tâches complexes telle que la reconnaissance de voix ou d'image, ou le traitement automatique du langage, qui sont des domaines où la théorie n'est pas encore très fructueuse.

J'ai l'impression que la communauté des data-scientists abandonne l'aspect théorique « data modelling » au profit de méthodes algorithmiques, notamment dans le cadre des compétitions type Kaggle. On voit beaucoup de méthodes algorithmiques donner de très bons résultats : sélection de variables avec PCA et NMF, estimation avec divers algorithmes, sans passer par la conception d'un modèle statistique qui pourrait régir les données.

Lecture de l'article "Model Selection in Data Analysis Competitions" qui donne des "recettes de cuisine" classiques utilisées dans le cadre des compétitions de data science.

Travail à rendre: Vous résumerez les 5 points principaux (3.1 à 3.5) en une ou deux phrases

Loïc Herbelot

1. Feature engineering is the most important part of predictive machine learning

Le plus important afin d'avoir de bons résultats n'est pas le choix de l'algorithme de l'estimateur, mais plutôt la gestion de l'information. Pour être efficace il faut savoir générer de nouvelles informations à partir des données, mais aussi savoir modifier ou supprimer certaines informations.

2. Simple models can get you very far

Les modèles simples sont facilement et rapidement modifiables, ce qui permet de les adapter rapidement à une situation. Cela facilite l'amélioration des résultats, contrairement à un modèle plus complexe de type « black-box ».

3. Ensembling is a winning strategy

Combiner plusieurs estimateurs qui sont décorrélés permet d'améliorer la précision des résultats

4. Overfitting to the leaderboard is a real issue

Le leaderboard fournit certaines informations sur la qualité du résultat, mais il vaut mieux s'appuyer sur des résultats de cross-validation pour juger de la qualité de sa méthode plutôt que d'essayer à avoir le meilleur score sur le leaderboard. Dans une compétition on a pu voir des gens très haut placé sur le leaderboard public avoir des résultats moyens sur le leaderboard privé.

5. Predicting the right thing is important

Il ne faut pas forcément essayer de prédire directement le résultat attendu, mais plutôt une variable dérivée qu'on utilisera pour avoir le résultat : exemple du challenge de prédiction des heures d'arrivée des avions, il valait mieux prévoir la durée du trajet, et ensuite déduire l'heure d'arrivée.

Il faut aussi savoir choisir les bonnes métriques : fonctions d'erreur, distance…