Principes de la sélection de modèles d'apprentissage

(ou comment gagner sur Kaggle)

BELHADJ Ghilas

MASTER DAC
Université de Pierre et Marie Curie

21 Septembre 2015

1. Créer de nouvelles *features* à partir de connaissances métier

Définition

Exploiter les connaissances que l'on a dans le domaine d'où proviennent les données pour sélectionner les *features* pertinentes, et en créer d'autres pour améliorer la prédiction, en supprimer celles qui sont susceptible de fausser la prédiction, etc...

Explication

Aucun algorithme ne peut décider de manière efficace si une *feature* contient de l'information pertinente ou pas. C'est l'intervention humaine sur les données d'apprentissage qui permet à la fin d'obtenir des modèles plus efficaces.

2. Ne pas sous-estimer les modèles simples

Définition

Simple ne rime pas forcément avec inefficaces. Les débutants ont tendance à oublier cette règle, et se servent d'algorithmes très complexes du premier coup, alors que les meilleures solutions se trouvent parmi les modèles simples de types arbres de décision.

Explication

Les modèles simples sont des modèles que l'on peut facilement comprendre, et donc qu'on peu facilement adapter selon les situations, contrairement aux autres modèles dont on ne comprend pas toujours le résultat.

3. Utiliser les combinaisons de modèles

Définition

Affiner encore plus la classification en utilisant plusieurs niveaux de modèles, chacun apportant sa contribution à la décision finale.

Explication

Généralement, on peut toujours mieux faire qu'un modèle seul, même si celui-ci est bien optimisé, Car on peut toujours obtenir plus de précisions en le combinant avec d'autres modèles qui viendront combler les manques du premier modèle.

4. 1er dans le classement publique ne veut pas dire premier dans le classement final

Définition

La *leaderboard* peut aider à améliorer son modèle, car le score est calculé sur des données qui ne sont pas incluse dans les données fournies par le challenge, leur utilisation est donc très recommandée. Par contre, elle peuvent facilement nous induire en erreur, si l'on se colle trop à ses résultats.

Explication

L'explication est que *Kaggle* fait en sorte à prendre des données de natures différentes pour le classement publique et pour le classement final. Ce explique l'écart énorme qu'il y a souvent entre les deux classements.

5. Optimiser la bonne variable avec la bonne fonction objectif

Définition

Utiliser systématiquement les algorithmes d'apprentissage sur les problèmes tels qu'ils sont posés dans les challenges ne sont pas toujours fructueux. il est important de revoir le problème pour le reformuler sous un meilleur angle.

Explication

Il existe différentes fonctions objectif, et toutes ne se valent pas selon le type de problème que l'on veux résoudre. De plus il se peut que la prédiction soit plus correcte, si l'ont prend le problème depuis un autre angle, en essayant d'optimiser une variable dérivée du problème, plutôt que la variable que l'on nous donne à optimiser.

6. Quelques autres conseils

- Lire les forums, c'est riche en informations, surtout pour les débutants.
- Laisser à la fin l'étape d'optimisation des paramètres.
- Apprendre des gagnants des compétitions passées.
- Savoir quel outil est adapté pour chaque type de problème.