[FDMS Devoir Maison] 5 grands principes de séléction de modèles statistiques

Rémi Cadène

Université Pierre et Marie Curie

remi.cadene@etu.upmc.fr

18 septembre 2015

Sommaire

Features engineering is important

- Il y a souvent des features cachées pouvant beaucoup augmenter les performances. En NLP par exemple, l'extraction de features du texte est cruciale (nombre d'occurences, part-of-speech tagging, etc.).
- Néanmoins, avoir trop de features peut diminuer les performances du modèle : Curse of Dimensionality / features redondantes / features inutiles.
- Par ailleurs, avec les données d'images ou de textes, les features peuvent être extraites automatiquement par des modèles de deep learning commme les convnets.

Beware of overfitting

- Le leaderboard public contient de l'information, mais les paramètres et hyper paramètres du modèles doivent être appris en cross validation sur les données d'apprentissage afin d'effectuer le meilleur score sur le leaderboard privé en généralisant le mieux.
- Bien sûr dans le cas où le leaderboard public contient des données issues d'une distribution différente, il convient d'adapter sa méthode de validation. Par exemple, si les données d'apprentissage contiennent des exemples issus d'un échantillonnage sur les années 1970-1990, que le public leaderboard 1990-2000 et que le private leaderboard 2000-2010.

Simple models are useful

- Très utile pour créer une baseline et pour comprendre les données.
- Parfois la solution est dans le prétraitement/extraction de features des données. Alors, une simple régression linéaire peut être un modèle gagnant malgré sa simplicité.

Ensembling for a higher accuracy

- Intuition : effectuer la moyenne des prédictions de plusieurs modèles permet d'être plus sûr de la bonne prédiction
- Marche très bien quand les modèles individuels sont moins corrélés
- Bending de plusieurs modèles simples peut être contre productif
- Combiner plusieurs modèles différents et bien construits marche mieux
- Optimiser un modèle individuel et l'insérer dans l'ensemble peut faire diminuer les résultats, il faut trouver les bons réglages pour l'ensemble.

Predicting the right thing

- Le choix du bon critère d'évaluation (ou fonction d'erreur) est très important, car il oriente l'optimisation des paramètres.
- Notamment, pour une même tâche de classification, les moindres carrés est un critère qui peut très bien fonctionner sur un certain type de données, alors qu'il faudra utiliser Hinge Loss pour d'autres.
- Parfois, entraîner ses modèles sur des critères différents de ceux proposés par les organisateurs de la compétition peut amener à de meilleurs résultats.