Sélection de modèles linéaire et non linéaires

Monrocq Christophe

Université René Descartes
Groupe de Recherche en Imagerie Biomédicale
45 rue des Saint-Pères
75006 PARIS
FRANCE
e-mail : monrocq@citi2 fr

Résumé

Nous allons aborder les problèmes d'estimation des erreurs d'apprentissage et de généralisation pour des modèles linéaires ou non. Il est reconnu que cette sélection doit suivre le principe de parcimonie, i.e. que les modèles les plus simples seront choisis prioritairement grâce à une pénalisation des modèles complexes. Mais le problème majeur qui se pose concerne la forme et l'importance du terme de pénalité.
Les nouvelles règles proposées pour la sélection de modèles, parmi lesquelles la règle GAE, reposent sur les estimations des erreurs d'apprentissage et de généralisation. Ces estimations vont aussi permettre de clarifier les liens qui existent entre ces deux erreurs, d'expliciter le terme pénalité précédent et de présenter des critères semblables à ceux d'Akaike pour les modèles linéaires (FPE et AIC) mais aussi valables pour des modèles non linéaires

1 Introduction

Résoudre un problème de classification (supervisée ou non) ou d'approximation de fonctions ne se limite pas à une étape d'optimisation permettant d'estimer les paramètres du modèle, mais nécessite aussi une seconde étape, dite de validation, qui nous indiquera le comportement du modèle en phase d'utilisation. Il s'agit alors d'évaluer l'erreur en généralisation.

L'apprentissage est réalisé en utilisant une base de données constituée de couples (x_i, y_i), où y est la sortie désirée associée à l'entrée x. Comme cette base est de taille finie, un apprentissage sans erreur serait possible en autorisant une complexité suffisante du modèle[1]. Cependant un apprentissage par cœur n'est pas souhaitable car d'une part cette base d'apprentissage ne représente qu'un échantillon possible de la réalité ; d'autre part si les exemples sont bruités, cet apprentissage va avoir l'inconvénient d'apprendre le bruit.
Par conséquent, l'étape de validation devra nous permettre de trouver le meilleur compromis entre l'erreur à l'apprentissage et la complexité du modèle.

Les résultats présentés dans cet article vont permettre d'introduire et d'expliquer ce compromis. Le premier résultat original (théorème 3), d'où découleront les suivants, met à jour une relation entre les erreurs à l'apprentissage et en généralisation, ce qui aura plusieurs conséquences : i) des résultats connus pour le cas linéaire (critères FPE et AIC d'Akaike) seront étendus au cas non linéaire ; ii) une nouvelle approche, "la règle GAE[2]", est proposée pour déterminer la complexité optimale du modèle.

Les problèmes de classification ou d'approximation de fonctions rentrent dans le cadre de la régression. Il s'agit alors de déterminer la valeur d'une variable aléatoire $\mathbf{Y}$ d'après les informations

1. dans le cas de la régression polynomiale, on peut augmenter le degré du polynôme pour passer par tous les points de la base d'apprentissage.

2. GAE = Graph of the Asymptotic Error. La règle GAE consiste à représenter l'Erreur Asympototique des modèles en fonction d'une mesure de la complexité de ceux-ci (par exemple le nombre de paramètres des modèles.)