SELECTION DES PREDICTEURS ET ESTIMATION DES TAUX D'ERREUR DE CLASSEMENT EN DISCRIMINATION LINEAIRE

Jean-Christophe TURLOT

STID Université de Pau

Résumé :

On montre que les procédures de sélection d'un sous-ensemble de prédicteurs pertinents pour la discrimination penvent engendrer un biais important dans l'estimation des taux d'erreur de classement par rééchantillonnage (validation croisée, jackknife ou bootstap). Le biais de 'sélection' peut conduire à un choix de prédicteurs en partie illusoire, dépendant des fluctuations d'échantillonnage. Il apparaît que la sélection d'un petit nombre de variables exploratoires, complétant l'information apportée par un ensemble de prédicteurs devant intervenir a priori dans l'élaboration de la règle de décision, constitue une protection contre une sélection trop sujette aux fluctuations d'échantillonnage lorsque la taille du fichier des observations est modérée. En réduisant ainsi le biais de sélection, l' estimation de la qualité de la règle par rééchantillonnage s'en trouve plus précise.

Mots-clés: discrimination, sélection de prédicteurs, erreur de classement.