Université Paris 1
Panthéon Sorbonne
12, place du Panthéon
75005 Paris

CES
Centre d'Économie de la Sorbonne
UMR 8174

Sujets

Sujet 1. Statistical learning vs machine learning

En économétrie / machine learning se pose la question de la sélection automatique des variables, particulièrement lorsque la base de données contient beaucoup de prédicteurs potentiels. On peut différencier les procédures avec inférence (statistical learning) des procédures sans inférence (machine learning). Par exemple, les procédures stepwise (forward ou backward) utilisent un critère statistique (T-stat, F-test, AIC,…), à l'inverse du LASSO qui utilise une procédure pour contracter les coefficients ayant peu d'influence économique. Après avoir présenté les différentes méthodes, ce mémoire se propose à partir de données simulées, d'estimer la performance de ces dernières, en considérant différents critères d'arrêt. En particulier, seront considérées les méthodes stepwise forward et backward, Forward Stagewise, LARS et LASSO. Pour cette dernière on s'intéressera typiquement au k-fold, leave-one-out (PRESS), CP, etc. On s'intéressera aussi à l'impact de la modalité du choix du paramètre lambda qui diffère selon que nous considérions un critère explicatif ou prédictif. Les données générées le seront sous l'hypothèse nulle d'indépendance, et sous l'alternative de dépendance. Sous cette dernière, seront pris en compte des dépendances linéaires, des dépendances linaires avec rupture(s) et de la multi-colinéarité. Une application empirique conclura le mémoire.

Mots clefs : Forward / backward selection, Statistical learning, Machine learning, Forward Stagewise / LARS / LASSO, k-fold, cross-validation

SAS : proc glmselect, proc iml