Credit scoring, statistique et apprentissage

Gilbert Saporta

Chaire de Statistique Appliquée & CEDRIC
Conservatoire National des Arts et Métiers
292 rue Saint Martin
75141 Paris Cedex 03
saporta@cnam.fr

Les accords dits « Bâle 2 » sur la solvabilité des banques ont remis au goût du jour les techniques de scoring en imposant aux banques de calculer des probabilités de défaut et le montant des pertes en cas de défaut. Nous présentons dans cet exposé les principales techniques utilisées et les problèmes actuels.

Le terme *credit scoring* désigne un ensemble d'outils d'aide à la décision utilisés par les organismes financiers pour évaluer le risque de non-remboursement des prêts. Un score est une note de risque, ou une probabilité de défaut.

Le problème semble simple en apparence, puisqu'il s'agit d'une classification supervisée en deux groupes, les « bons payeurs » et les « mauvais payeurs ».

Les classifieurs linéaires sont les plus classiques et souvent les seuls utilisables en raison de contraintes légales : on doit pouvoir expliquer la décision de refus. Ces classifieurs doivent être adaptés au cas de prédicteurs qualitatifs, que l'on rencontre le plus souvent en crédit à la consommation. On rappellera que l'usage de variables qualitatives remonte à des travaux très anciens de Fisher. La régression logistique est devenue un standard dans la profession, que l'on oppose souvent à tort à l'analyse discriminante.

La théorie de l'apprentissage statistique apporte alors des justifications à l'usage de techniques de réduction de dimension (méthode Disqual de discrimination sur composantes factorielles, régression PLS) et de régularisation (régression ridge). La régression PLS se révèle équivalente à une technique méconnue : l'analyse discriminante barycentrique qui est le pendant additif du classifieur naïf de Bayes qui est multiplicatif.

Le choix entre méthodes ou algorithmes ne peut reposer sur des critères statistiques de type vraisemblance, inadapté à des problèmes de décision mais sur des mesures de performance en généralisation. La courbe ROC et les indices associés (AUC, Gini, Ki) sont les plus utilisés.

Un des problèmes épineux est celui du biais de sélection : en effet l'ensemble d'apprentissage ne contient que des individus dont la demande de prêt a été accordée. On sait que si les variables d'acceptation sont différentes des variables disponibles, on ne peut trouver de solution sans biais. La prise en compte des dossiers refusés (*reject inference*) donne lieu cependant à une abondante littérature, sans guère de résultats convaincants.

La discrimination entre défaillants et non-défaillants n'est plus le seul objectif, surtout pour des prêts à long terme : le « quand » devient aussi important que le « si ». De nombreux travaux s'orientent actuellement vers l'utilisation de modèles de survie pour données censurées dont nous donnerons un aperçu.