Apprentissage Statistique de la Classification : Partie I.

Gérard Biau, Laurent Rouvière

Université Paris VI, Université Rennes 2

biau@math.univ-montp2.fr

laurent.rouviere@univ-rennes2.fr

L'*apprentissage statistique* désigne un vaste ensemble de méthodes et d'algorithmes permettant, dans un sens général, d'extraire l'information pertinente de données ou d'apprendre des comportements à partir d'exemples. Les applications de ce paradigme sont très nombreuses, allant de la recherche d'informations dans de grands ensembles de données (fouille de textes ou d'images) à la biologie (reconstruction des réseaux génétiques, puces ADN, etc).

Notre exposé se divisera en trois parties. Dans la première partie, nous rappelons les concepts d'apprentissage supervisé et non-supervisé à travers divers exemples. Dans le premier cas, il s'agit de trouver une fonction g (appelée *règle*) qui permet d'expliquer une variable à partir d'observations d'autres variables. Dans le second cas, c'est-à-dire en l'absence d'une variable à expliquer, l'objectif est la recherche d'une typologie permettant de regrouper les observations en classes homogènes. Il existe de nombreuses approches permettant de traiter ces deux domaines de la statistique, comme par exemple :

- La discriminante de Fisher, les SVM (Support Vector Machines), les plus proches voisins, la règle du noyau, les arbres de classification dans le cas supervisé ;
- La méthode des k-means dans le cas non-supervisé.

Les différentes méthodes de classification énoncées ci-dessus dépendent en général d'un ou plusieurs paramètres dont le choix se révèle crucial. Dit autrement, étant donnée une famille de règles candidates, comment choisir, à partir des seules observations, la "meilleure" règle (ou encore une règle qui ne s'éloigne pas trop de l'optimum) en un sens à préciser ? Dans la deuxième partie de l'exposé, nous abordons cette problématique via la théorie de Vapnik-Chervonenkis qui propose des techniques de sélection basées sur la minimisation de critères empiriques. Nous mettrons en particulier l'accent sur l'intérêt de cette approche :

- D'un point de vue pratique, la procédure est "automatique" : nul besoin pour le praticien de choisir certains paramètres ;

- D'un point de vue théorique, il est possible d'établir des bornes de performance permettant de mesurer la qualité de la règle sélectionnée.

Dans la troisième et dernière partie de l'exposé, nous verrons comment adapter les algorithmes de l'apprentissage à des données plus complexes. En effet, dans de nombreux domaines de la statistique contemporaine, les individus prennent la forme de *courbes* (ou *surfaces*). Ces courbes peuvent, par exemple, représenter la température en un point du globe, le cours d'une action en bourse, le tracé d'un électrocardiogramme ou encore la consommation d'électricité d'une grande ville... Les appareils de mesure ne captant que la valeur de la courbe à certains instants, les données disponibles ne sont en fait que des versions discrétisées de la courbe. Toutefois, le statisticien a intérêt à prendre en compte le caractère *continu* de ces observations. Il devient dès lors pertinent de ne plus considérer ces individus comme des vecteurs de grande dimension, mais plutôt de les appréhender comme des "fonctions", c'est-à-dire comme des objets uniques évoluant dans des espaces de dimension infinie. Il faut alors adapter les méthodes statistiques classiques à ces nouveaux individus. L'approche que nous développons consiste à réduire la dimension infinie des observations en ne considérant que certains coefficients des données décomposées dans une base appropriée (Fourier ou ondelettes par exemple). Nous présentons les propriétés asymptotiques et à distance finie de la méthode envisagée et nous illustrons les performances de cette approche sur des jeux de données réelles et simulées.

Références

Hastie, T., Tibshirani, T. et Friedman, J. (2001) *Elements of Statistical Learning*, Springer.

Vapnik, V.N., (1995) *The Nature of Statistical Learning Theory*, Springer.