Plus proches voisins

Benjamin Monmege
benjamin.monmege@lsv.ens-cachan.fr

9 février 2012

Exercice 1 (Classification)**.** On se donne des *classes* $\mathcal{C}_1, \ldots, \mathcal{C}_K$ et des données $(x_n)_{1 \leq n \leq N}$ avec $x_n \in \mathbb{R}^D$, chacune appartenant à l'une des classes. On cherche à apprendre les classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$ et à les utiliser pour prédire la classification de nouveaux exemples. Par exemple, si $K = 2$, on peut se demander si un patient a un cancer ou non, en considérant un certain nombre de paramètres médicaux.

1. On va utiliser un exemple avec $K = 2$ et $D = 2$. Charger les données présentes à l'adresse `http://www.lsv.ens-cachan.fr/~monmege/teach/learning/voisins_data.mat` : elles sont composées d'une matrice X contenant 180 lignes ($N = 180$) et 2 colonnes, et d'une matrice Y contenant 180 lignes et 1 colonnes, contenant 1 ou 2 selon la classe à laquelle appartient la donnée correspondante. Visualiser les données dans un graphique.
2. Tirer des points aléatoirement dans le domaine d'intérêt des données, et chercher les r plus proches voisins pour plusieurs valeurs de r. Les visualiser sur le graphique. En déduire une manière simple de prédire la classe des nouveaux exemples.
3. Dessiner la *cartographie* du domaine : choisir un pas de discrétisation, et prédire la classe de chaque point de cette grille, puis représenter le tout sur le graphique.
4. Ajouter des *données aberrantes* (données d'une classe très éloignées de leur classe réelle en plus ou moins grand nombre) et observer les variations de la cartographie précédente.
5. À r fixé, calculer la performance de votre algorithme à l'aide de validation croisée. Utiliser ce score afin de choisir le *meilleur* paramètre r possible.
6. Modulariser votre code, puis le tester sur l'exemple de la classification des iris disponible sur la page `http://archive.ics.uci.edu/ml/datasets/Iris`.

Exercice 2 (Régression)**.** Supposons désormais que les données $(x_n)_{1 \leq n \leq N}$ sont associées à des valeurs cibles continues $(y_n)_{1 \leq n \leq N}$ avec $y_n \in \mathbb{R}$. On cherche pour une donnée x nouvelle à prédire la meilleure valeur y associée.

1. En utilisant la même idée des plus proches voisins, écrire un script MATLAB réalisant une prédiction.
2. Tester l'algorithme sur des données générées à partir d'une fonction $f : \mathbb{R} \to \mathbb{R}$ (une droite, une fonction sinusoïdale, une parabole...) à laquelle on ajoutera un bruit Gaussien de moyenne nulle : faire varier la variance du bruit, le nombre de voisins considérés...