Perceptron et K-Nearest Neighboors

Travaux Pratiques

On se place dans ce TP dans un cadre de classification binaire, où $Y_i \in \{-1, 1\}$ est expliqué par p régresseurs $X_i^1, \ldots, X_i^p$. Deux exemples seront traités :

- un exemple simulé, pour lequel on prendra $p = 2$, X_i^1 et X_i^2 uniformément répartis dans le carré $[0, 1] \times [0, 1]$, et la loi conditionnelle de Y_i sachant (X_i^1, X_i^2) est donnée par :

$$\begin{cases} \mathbb{P}(Y_i = 1) = \alpha & \text{si} \quad 2X_i^1 + X_i^2 < 1.5, \\ \mathbb{P}(Y_i = 1) = \beta & \text{si} \quad 2X_i^1 + X_i^2 > 1.5, \end{cases}$$

  où α et β sont deux paramètres sur lesquels on jouera un peu pour tester les méthodes ;
- un exemple de données réelles issu du livre [3] "The elements of statistical learning : data mining, inference and prediction" (section 4.4.2) concernant l'explication du risque d'attaques cardiaques par des facteurs comme l'âge, la consommation de tabac, etc. dont on avait récupéré les données dans le TP précédent.

On cherchera dans ce TP à comparer les mérites de trois approches simples et populaires du problème : la régression logistique, l'algorithme du perceptron de Rosenblatt et la méthodes des K plus proches voisins. Lors du TP précédent, on a construit une fonction R appelée `rexemple` prenant pour arguments n, α, et β, et renvoyant un n-échantillon $\left(X_i^1, X_i^2, Y_i\right)_{1 \leq i \leq n}$ de l'exemple simulé sous la forme de liste de trois éléments : `list(x1, x2, y)`, où `x1, x2` et `y` sont des tableaux. On avait également étudié la régression logistique sur l'exemple simulé et sur les données réelles.

- Perceptron -

1. Programmer et essayer l'algorithme du perceptron tel qu'il est décrit dans la section 4.5 de [3].
2. Visualiser l'évolution de la droite de séparation avec au fil des itérations.
3. Montrer sa convergence quand les données sont séparables c'est-à-dire par exemple si $\alpha = 1$ et $\beta = 0$. Que se passe-t-il dans les autres cas ?
4. Que donne l'algorithme du perceptron sur les données réelles ?

- "K-Nearest Neighbors" -

5. Appliquer la méthode $K - NN$ aux données simulées, et visualiser le classifieur obtenu pour différentes valeurs de K. Comment choisir K ?
6. A partir de quelle taille d'échantillon obtient-on un classifieur de bonne qualité ?
7. Traiter ensuite le cas réel de l'explication du risque d'attaques cardiaques.
8. Récupérer sur le site `http://www-stat.stanford.edu/~tibs/ElemStatLearn/` les données de la base zipcode, et comprendre ce qu'elles contiennent.
9. Appliquer la méthode $K-NN$ (version multi-classe) aux données issues de la base zipcode avec différents choix de $K \geq 1$. Estimer la matrice de confusion $(\mathbb{P}\{C_K(X) = i, Y = j\})_{i,\, j}$ associée au classifieur C_K ainsi obtenu. Proposer une méthode pour choisir K et la mettre en oeuvre.

Références

[1] Pierre-André Cornillon, Arnaud Guyader, François Husson, Nicolas Jégou, Julie Josse, Maela Kloareg, Eric Matzner-Lober, and Laurent Rouviere. *Statistiques avec R.* Didact Statistiques. Presses Universitaires de Rennes, 2nd edition, 2010.

[2] Pierre André Cornillon and Eric Matzner-Lober. *Régression avec R.* Springer, Collection Pratique R, 1st edition, 2011.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning : data mining, inference and prediction.* Springer, 2 edition, 2009.