Université d'Angers
Département de Mathématiques

Année 2018-2019
M2 Data Science

Statistique en Grande Dimension et Apprentissage - TP 1

Ce TP est consacré aux k-plus proches voisins. A travers la mise en oeuvre de cet algorithme, les objectifs sont de se familiariser avec le calcul empirique du risque, la validation simple, la validation croisée ainsi que le choix de paramètres (ou hyper-paramètres). On abordera aussi la notion de surapprentissage.

Exercice 1 (Exemple simulé)**.** On se place ici dans le cadre suivant. On suppose que $\mathbf{X}$ suit la loi uniforme sur $[0,1]^2$ et que Y est une variable aléatoire à valeurs dans $\{0,1\}$ telle que

$$\mathbb{P}(Y=1|\mathbf{X}=x) = \begin{cases} \alpha & \text{si } |X_1+2X_2| \leq 1 \\ \beta & \text{si } |X_1+2X_2| > 1. \end{cases}$$

1. Ecrire une fonction pour générer un échantillon $\mathcal{D}_n$ de taille n.
2. Simulez un échantillon d'entraînement (`Xtrain`,`Ytrain`) de taille 100 et un échantillon test de taille 100 également que vous noterez (`Xtest`,`Ytest`).
3. Faire une représentation graphique de l'échantillon d'entraînement en attribuant une couleur différente et un symbole différent selon la valeur de Y (paramètres, `col` et `pch` de la fonction `plot`).
4. Appliquez la fonction `knn` du package `class` (écrire `library(class)`), pour prédire les points de coordonnées $(1/2, 1/2)$ et $(1/4, 3/4)$.
5. Appliquez la fonction `knn`, pour prédire les labels de l'échantillon test avec $k = 1$, $k = 10$, $k = 20$.
6. A l'aide de la fonction `table`, afficher la matrice de confusion (qui permet de visualiser les classements prédits et les classements réels) pour chacun des cas proposés.
7. En déduire l'erreur de classification pour chacun des cas proposés.
8. Calculez l'erreur par validation simple pour $k = 1 : 20$ et tracez l'évolution sur un graphe. Que constatez-vous ?
9. Calculez l'erreur par validation croisée sur tout l'échantillon en utilisant la fonction `knn.cv` (validation croisée de type Leave-One-Out) puis en programmant une validation croisée 5-fold. Que constatez-vous ? Quel est le choix optimal de k ?

Exercice 2. Récupérer les jeux de données `synth_train.txt` et `synth_test.txt`. On a $Y \in \{1,2\}$ et $X \in \mathbb{R}^2$. On dispose de 100 données d'apprentissage et 200 données test.

1. Charger le jeu de données d'apprentissage dans `R` à l'aide de la commande `read.table`. Afficher les données d'apprentissage.
2. Reproduire les questions 5, 6, 7 et 9 de l'exercice précédent avec $k = 30$.

3. Représentez la *frontière de décision* pour $k = 30$ puis $k = 15$, puis $k = 1$. Dans quels cas peut-on parler de sur ou sous-apprentissage ?
   Pour fabriquer ce graphe, calculez la prédiction sur une grille de points adaptée aux données, puis représentez la prédiction des points de cette grille en leur attribuant une couleur.

Exercice 3. Dans ce denier exercice, on s'intéresse à une base de données célèbre : la base de données `MNIST` qui contient 70000 images de chiffres en noir et blanc et en 28×28 pixels.

1. Chargez les jeux de données `mnist_train.csv` et `mnist_test.csv`.
2. En remarquant que la première colonne est celle des réponses, fabriquez des `data.frame`(s) `Xtrain`, `Ytrain`, `Xtest` et `Ytest`.
3. Dans ce jeu de données, on atteint un peu les limites algorithmiques des k-plus proches voisins dont la complexité est en $O(nkd)$ où d est la dimension du problème (ici, $d = 784$).
4. Représentez l'une des observations sous la forme d'un tableau 28×28 afin de vous familiariser avec l'objet.
5. Tentez la mise en oeuvre de l'algorithme avec seulement 10 données test. Précédez la commande de `t1=Sys.time()` et écrivez ensuite `t1=Sys.time(); difftime(t2,t1)` afin d'évaluer le temps de calcul. Que pouvez-vous en déduire sur le temps de calcul de la base de données test complète ?
6. On propose donc dans la suite de travailler seulement avec un sous-échantillon d'entraînement de 12000 observations. Fabriquez ce sous-échantillon en utilisant la fonction `sample`.
7. Mettez en oeuvre l'algorithme sur ce sous-échantillon d'entraînement avec une base test de 500 observations (pour limiter le temps de calcul) et $k = 10$ voisins.
8. Représentez la matrice de confusion et calculez l'erreur de classification.