Université de Lille
M2 ISN- Méthodes d'Apprentissage obouaziz@univ-lille.fr

TP. 4

Méthode des k-plus proches voisins et régression logistique

Dans cet exercice on étudie la base de donnée **heart_cleveland**. Cette base de données contient des informations sur des patients d'un hôpital de Cleveland, aux Etats-Unis. Le but est de classifier si les patients ont développé une maladie cardiaque ou non en fonction de 10 attributs. La base de données contient les variables suivantes :

- age : l'âge en années
- trestbps : taux de pression sanguine du patient au repos en mm/HG
- chol : le cholestérol en mg/dl
- thalach : vitesse cardiaque maximum
- oldpeak : ST dépression due à un exercice
- cp : douleur à la poitrine (0 angine typique, 1 angine atypique, 2 douleur autre, 3 asymptomatique)
- exang : angine provoquée par de l'exercice, 1 oui, 0 non
- slope : la pente du pic de l'exercice ST (0 croissant, 1 plat, 2 décroissante)
- ca : nombre de vaisseaux majeurs coloriés par fluoroscopie
- thal : 0 normal, 1 anomalie réparée, 2 anomalie réversible
- condition : vaut 1 si la personne a une maladie cardiaque, 0 sinon

1. Sélectionner uniquement les variables quantitatives de la base de données, avec la variable condition. Séparer les données en échantillon d'apprentissage et de test. Entraîner l'algorithme des $k = 3$ plus proches voisin et évaluer sa qualité de classification sur les données test. Tester d'autres valeurs de k et commenter.
2. Faire une boucle sur les valeurs de k allant de 1 à 80 : entraîner à chaque fois l'algorithme des k plus proches voisins et calculer l'erreur de classification. Tracer la courbe de l'erreur en fonction de k.
3. Utiliser la fonction GridSearchCV pour trouver par cross-validation la valeur de k donnant la plus petite erreur (pour k allant de 1 à 80). Tracer de nouveau l'erreur de classification en fonction de k.
4. A l'aide de la fonction StandardScaler, centrer et réduire les variables quantitatives. Appliquer de nouveau toutes les étapes précédentes avec ces nouvelles covariables.
5. On considère à présent la base de données complète avec les variables qualitatives et les variables quantitatives centrées réduites. Appliquer de nouveau une méthode de choix de k sur cette base de données. Comparer les différents résultats en fonctions des différentes traitement de la base de données et conclure.

6. Ajuster un modèle de régression logistique, interpréter les résultats et décrire les facteurs de risque pour la maladie cardiaque.
7. Comparer le taux de mauvaise classification entre le modèle de régression logistique et les k-plus proches voisins. Conclure.
8. Question supplémentaire : sous R, ajouter une étape de sélection de variables basée sur l'AIC ou le BIC et implémenter le modèle de régression logistique utilisant ce nombre réduit de variables. Calculer son erreur de classification.