Introduction à la Science des données

Travail pratique 04 – Apprentissage supervisé

Professeurs : Andres Perez-Uribe & Carlos Peña
Assistants : Axel Fahy, Shabnam Ataee, Xavier Brochet
Emails: prenom.nom@heig-vd.ch

Objectifs :

- Réaliser l'exploration d'une base de données réelle en calculant des caractéristiques statistiques des attributs et en générant des représentations graphiques de ces statistiques.
- Programmer un modèle de classification à base de règles en utilisant des attributs pertinents.
- Utiliser l'algorithme des K plus proches voisins dans un problème de classification et évaluer sa performance. Comparer les résultats avec l'approche à base de règles.

1. Wine database

Nous allons utiliser la base de données «Wine Data Set » disponible sur le dépôt de bases de donnes maintenu par l'Université de Californie à Irvine (UCI). Le lien direct est celui-ci : http://archive.ics.uci.edu/ml/datasets/Wine?Quality

1) Créez un notebook pour lire la base de données et générez un dataframe.
2) Se familiariser avec la base de données (p.ex., nombre d'observations, des classes, d'attributs, statistiques des attributs, données manquantes).
3) Réaliser une analyse exploratoire de la base de données en utilisant des box plots et des scatter plots. Analysez s'il y a des variables avec très peu de variabilité et cherchez à identifier des variables qui ont des valeurs différentes pour les différentes classes.

2. Modèle à base de règles

Utilisez les box-plots des variables qui ont une majorité de valeurs différentes pour les différentes classes afin de programmer des règles (if-then-else) permettant la classification de chaque observation. a) Essayez au moins trois variables de manière individuelle et b) deux classificateurs à multiples variables (p.ex., deux ou trois) et évaluez le nombre d'observations qui sont correctement classées (accuracy) par ses modèles.

3. L'algorithme des k plus proches voisins (k-NN)

Adaptez le code de k-NN vu en classe pour traiter le problème de classification des vins et testez la performance (accuracy) des modèles pour K=1,2,3,5,7,10. Utilisez la méthode de validation hold-out (calculez la moyenne de performances en répétant le « split » 10 fois) et la validation croisée avec n_folds=5.
Présentez un résumé de vos résultats (c.a.d., un tableau avec les résultats ainsi qu'un plot ou bar chart) et commentez ceux-ci.

4. Évaluation des modèles

Utilisez la bibliothèque scikit-learn (sklearn.metrics) pour calculer la matrice de confusion du meilleur modèle trouvé précédemment (c.a.d., pour un K donné), sur l'ensemble de validation pour chaque « fold » de la validation croisée. Y a-t-il des classes pour lesquelles nous avons plus de peine à faire la bonne classification ?

5. L'algorithme LVQ

Utilisez l'algorithme LVQ pour traiter le problème de classification des vins et testez la performance (accuracy) des modèles et calculez la matrice de confusion du meilleur modèle trouvé en explorant différentes valeurs de hyper-paramètres (c.a.d., nombre de prototypes, learning rate et nombre d'epochs), sur l'ensemble de validation pour chaque « fold » de la validation croisée. Commentez vos résultats et comparez l'accuracy de ce modèle avec celle obtenu avec k-NN.

Rapport

Préparer un fichier compressé nommé nom1_nom2_ISD_TP4.zip intégrant UN SEUL notebook et le télécharger sur Cyberlearn avant la date limite indiquée. Veuillez bien séparer les différentes parties dans le notebook et veuillez également intégrer vos réponses aux questions posées dans les points 1 à 5.