Python et Machine Learning
Travaux Dirigés
Sixième partie

Salim Lardjane

Université de Bretagne Sud

1. Lire les données de températures moyennes annuelles de la station météorologique autrichienne de Sonnblick à partir du fichier AvgTemp.xls, disponible sur le forum. Calculer le coefficient de corrélation de Pearson et de Spearman, puis le tau de Kendall, entre la variable de température et l'année.
Calculer l'accroissement annuel de la température, en supposant que la dépendance de la température par rapport à l'année est linéaire. Cet accroissement est-il significatif ? Vérifier que le modèle de régression linéaire ajusté est satisfaisant en testant la normalité des résidus (par exemple, à l'aide du test de Kolmogorov-Smirnov).

2. Récupérer le programme logistic.py sur le forum, le faire tourner et essayer de comprendre son fonctionnement.

3. Récupérer le programme acp.py sur le forum, le faire tourner et essayer de comprendre son fonctionnement.

4. Importer sous Python les données bankloan.sav du répertoire Programmes sur le Forum. Il s'agit de données SPSS. On pourra éventuellement passer par R. Il s'agit de données concernant des prêts accordés par une banque. La variable Default indique si le prêt a été remboursé ou pas. Les autres variables fournissent des informations sur le titulaire du prêt.

5. On n'utilisera pas la variable "Ed" et les trois dernières variables du tableau. Les supprimer des données.

6. Sous Python, scinder le tableau de données en deux parties : les prêts pour lesquels la variable Default est renseignée – on appelera la table obtenue *data* – et les prêts pour laquelle elle ne l'est pas et pour lesquels on veut faire une prévision – on appellera la table obtenue *prev* –.

7. Sous Python, scinder les données *data* en un ensemble d'apprentissage et un ensemble de test.

8. On va utiliser la méthode des plus proches voisins pour prévoir la variable Default. Pour choisir le nombre de voisins, faire une boucle ou celui-ci varie entre 1 et 12 et obtenir le score de prévision sur l'ensemble test à chaque fois. Quel nombre de voisins retenir ? Quel est le score de test correspondant ?

9. Faire les prévisions pour les nouvelles données du tableau *prev* avec le nombre de voisins déterminé à la question précédente.

10. Reproduire les analyses précédentes en utilisant la régression logistique et la classification par machine à vecteurs supports linéaire (linear SVC).

11. Quelle méthode retenir à ce stade ?