Data Mining
TP1

Julien Blanchard

Mars 2020

1. Le jeu de données *german.data* décrit des prêts accordés par une banque allemande. A l'aide de la librairie *pandas*, chargez-le dans un dataframe `df`. Faites les vérifications d'usage à l'aides des fonctions `df.dtypes`, `df.shape`, `df.count()`, `df.describe()`.
2. Tracez l'histogramme et la boîte à moustaches de la variable *duration*. Tracez le diagramme en secteur de la variable *purpose*.
3. Vous allez utiliser *scikit-learn* pour prédire la variable *class* à l'aide d'algorithmes d'apprentissage. Etant donné que *scikit-learn* supporte peu les variables catégoriques, commencez par appliquer une analyse des correspondances multiples (module *mca.py*) pour transformer vos variables catégoriques en variables numériques (laissez la classe à part). Il faut d'abord effectuer un codage disjonctif complet à l'aide de `pandas.get_dummies()`.
4. Construisez des modèles pour prédire la variable *class* à l'aide de différents algorithmes d'apprentissage : arbre de décision, k plus proches voisins, régression logistique, forêts aléatoires, SVM...
5. Evaluez vos algorithmes à l'aide d'une validation simple puis d'une validation croisée (utilisez comme mesure l'*accuracy*, càd le taux de réussite).
6. En faisant varier la complexité du modèle, représentez la qualité des modèles en fonction de la complexité. Interprétez les courbes obtenues. Est-ce conforme au cours ?