INITIATION À LA FOUILLE DE DONNÉES
TP IV : régression linéaire et sélection de modèle

Créez un projet sur Enterprise Miner et utilisez le jeu de données « USCrime.csv ». L'objectif de cet exercice est d'observer les différentes méthodes et les différents critères de sélection de variables pour la méthode de régression linéaire.

1. Définition du problème

À l'aide du fichier de description « USCrime.txt », définissez le problème associé à ce jeu de données : quelle est la variable cible ?

2. Régression « brute »

Importez les données du fichier « USCrime.csv ». Construisez un diagramme vous permettant d'effectuer une régression linéaire. Dans un premier temps, vous utiliserez les répartitions par défaut des ensembles d'apprentissage, de validation, et de test (40%, 30%, 30%).

Vous répondrez aux questions suivantes :
- Lorsque que vous visualisez vos résultats dans l'onglet « plot », que signifie diagramme ? Si la prédiction était parfaite (sans aucune erreur), que devrait-on voir ?
- Quelles sont les erreurs d'apprentissage, de validation, et de test ?
- À partir des onglets « estimates » et « table », identifiez les 5 variables dont les coefficients sont en valeur absolue les plus importants. Pensez-vous qu'elles peuvent expliquer de façon cohérente la réponse ? Qu'est-ce que l' « intercept » ?

Réitérez cette démarche avec 100% des données en apprentissage.
- Que constatez vous sur les variables sélectionnées ?
- Quelle est l'erreur d'apprentissage ? Comparez la à celle obtenue précédemment. La différence vous semble-t-elle importante ?

3. Sélection de variables

Redécoupez vos ensembles d'apprentissage, de validation et de test selon la répartition usuelle (40%, 30%, 30%). Pour observer les étapes de sélection de modèles, cochez les options adéquates dans le menu Output/Printed_Output du noeud « Regression ». A partir du menu « Selection Method » du noeud « Regression », lancez et observez les résultats obtenus pour :
- différentes méthodes de sélection de modèles : backward, forward (et stepwize si vous avez le temps),
- différents critères de sélection : AIC, erreur de validation (et validation croisée si vous avez le temps).

Pour chacun des paramétrages, utiliser le menu output pour observer les sélections successives de variables. Donnez une description des processus de sélection forward / backward :
- À l'étape t, quelle variable est ajoutée / éliminée ?
- Pour le critère AIC, quel est le modèle sélectionné ?

Pour chacun des paramétrages, explorez les différents menus de résultats de la méthode. Vous devrez à chaque fois :
- Trouver quelles variables ont été sélectionnées.
- Identifier les variables les plus significatives dans le modèle final.
- Donner les coefficients des paramètres du modèle pour ces variables.
- Utiliser le menu plot et observer les valeurs prédites en fonction de différents paramètres.

Comparez globalement vos différents résultats, en terme d'erreurs de généralisation, et d'interprétation.