
Jeu de données "Spambase"

Spambase est un jeu de 4601 données, chacune décrivant un mail et sa catégorie : spam ou non-spam. Les attributs descriptifs de ces mails sont les fréquences d'apparitions de certains mots ou caractères ainsi que des informations sur la quantité de caractères mis en capitales. Ce TP a pour objectif de vous familiariser avec la méthode de régression logistique vue en cours. Vous profiterez également de cette séance pour tester les autres méthodes proposées par Enterprise Miner comme les réseaux de neurones. Vous utiliserez ces méthodes comme des "boîtes noires" et vous choisirez les configurations par défaut de ces méthodes. Attention, tout comme la régression linéaire et logistique, certaines méthodes ne s'appliquent que pour des classes continues ou que pour des classes binaires.

La régression logistique est utilisée pour obtenir un classifieur de données à classes binaires. Cette méthode est fondée sur une hypothèse importante : la fonction $\text{Logit}(P(Y|X))$ est approximable par un hyperplan. La méthode de régression logistique calcule une estimation des coefficients de cet hyperplan à partir des données.

Etape 1 : La régression logistique.

Chargez le jeu de données spambase. Aucune modification des données n'est demandée ou nécessaire. Partitionnez le jeu de données ainsi : 40% apprentissage, 30% validation, 30% test. Utilisez le noeud de régression et observez les différents paramètres proposés. Faites le lien entre la régression logistique et la régression linéaire vue au TP précédent.

De manière analogue au TP précédent, lancez la méthode de régression avec différentes configurations : backward, forward, stepwise, et différents critères de sélection des attributs. Observez à chaque fois dans le menu output des résultats le déroulement de la phase de sélection d'attributs.

Etape 2 : Autres méthodes.

Comparez maintenant les résultats obtenus avec les autres méthodes proposées par Enterprise Miner : les arbres de décision, les réseaux de neurones, etc...