SY19 – A16
TP 3: classification linéaire

1 Analyse des données `spam`

Chargez les données `spam`. Séparez-les en un ensemble d'apprentissage (2/3 des exemples) et un ensemble de test. Appliquez sur ces données l'analyse discriminate linéaire et la régression logistique. Calculer la matrice de confusion et le taux d'erreur pour chacune des méthodes. Tracer sur le même graphique les courbes COR. Quels prédicteurs ont un coefficient significativement non nuls (régression logistique) ?

2 Comparaison ADL-ADQ sur des données simulées

Dans cet exercice, on génère des données gaussiennes avec $K = 2$, $p = 3$, et les paramètres suivants :

$$\pi_1 = \pi_2 = 0.5, \quad \mu_1 = (0, 0, 0)^T, \quad \mu_2 = (1, 1, 1)^T$$

$$\Sigma_1 = I_3, \quad \Sigma_2 = 0.8 I_3.$$

Générer des échantillons d'apprentissage de taille $n \in \{30, 100, 1000, 10000\}$ et un échantillon de test de taille 10000. Etudiez le taux d'erreur de test pour l'ADL et l'ADQ, en fonction de n. Commentez les résultats.