SY19
TP 3: Classifieur de Bayes
Classification linéaire et quadratique

1 Participation au marché du travail

Les données `Participation` dans le package `Ecdat` portent sur les choix de femmes mariées de participer ou non au marché du travail, en fonction de variables socio-économiques. Analyser ces données en utilisant la régression logistique. Interpréter les résultats. Quels facteurs semblent jouer un rôle significatif ?

2 Analyse des données `spam`

1. Charger les données `spam`. Partitionner les données en un ensemble d'apprentissage (environ 2/3 des observations) et un ensemble de test.
2. Construire un classifieur en utilisant la régression logistique. Afficher la matrice de confusion sur les données de test. Quel est le taux d'erreur de test de ce classifieur ?
3. Mêmes questions pour l'analyse discriminante linéaire (fonction `lda` de la librairie `MASS`).
4. Les probabilités d'erreur de ces deux classifieurs sont-elles significativement différentes ?
5. En utilisant la fonction `roc` de la librairie pROC, afficher les courbes COR des deux classifieurs précédents. Si on veut détecter 80% des spams, quel percentage de courriels désirables seront classés à tort comme spams ?

3 Estimation de la probabilité d'erreur de Bayes

On considère un problème de classification avec $c = 3$ classes et $p = 2$ prédicteurs. La distribution marginale de Y est définie par les probabilités a priori :

$$\pi_1 = 0.3, \quad \pi_2 = 0.3, \quad \pi_3 = 0.4,$$

et les densités conditionnelles de $\mathbf{X}$ sachant $Y = k$, $k = 1, 2, 3$ sont des lois normales multidimensionnelles $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ avec

$$\boldsymbol{\mu}_1 = (0, 0)^T, \quad \boldsymbol{\mu}_2 = (0, 2)^T, \quad \boldsymbol{\mu}_3 = (2, 0)^T,$$

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$

1. Estimer la probabilité d'erreur de Bayes pour ce problème (utiliser la fonction `dmvnorm` de la librairie `mvtnorm` pour calculer la densité de la loi normale multidimensionnelle).
2. Générer des ensembles d'apprentissage de différentes tailles, et comparer les probabilités d'erreur des classifieurs ADL et ADQ entraînés sur ces données à la probabilité d'erreur de Bayes.