SY09 Printemps 2013
TP 4
Analyses discriminantes quadratique et linéaire

Exercice 1. Règle de Bayes

On suppose que la population est répartie en deux classes, en proportions π_1 et $\pi_2 = 1 - \pi_1$, issues des distributions gaussiennes bivariées $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ et $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$.

1. Donner une équation de la frontière de décision de la règle de Bayes dans chacun des cas suivants :
   (a) $\pi_1 = 0.5$, $\boldsymbol{\mu}_1 = (0,0)'$, $\boldsymbol{\mu}_2 = (1,1)'$, $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$;
   (b) $\pi_1 = 0.1$, $\boldsymbol{\mu}_1 = (0,0)'$, $\boldsymbol{\mu}_2 = (1,1)'$, $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$;
   (c) $\pi_1 = 0.5$, $\boldsymbol{\mu}_1 = (0,0)'$, $\boldsymbol{\mu}_2 = (1,1)'$, $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & -0.3 \\ -0.3 & 1 \end{pmatrix}$;
   (d) $\pi_1 = 0.6$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = (1,1)'$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$;
   (e) $\pi_1 = 0.6$, $\boldsymbol{\mu}_1 = (0,0)'$, $\boldsymbol{\mu}_2 = (1,1)'$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.
2. Pour chacune des cinq populations précédentes, en utilisant la fonction `simul`, générer un échantillon de taille $n = 1000$, et tracer le nuage associé en représentant les deux classes de manière distincte ; pour les trois premières situations, on ajoutera le tracé de la frontière de décision. Pour chaque cas de figure, donner l'expression d'un estimateur de la probabilité d'erreur, ainsi que sa réalisation sur l'échantillon correspondant. On comparera avec la probabilité d'erreur théorique lorsqu'on sait la calculer.

Exercice 2. Analyse discriminante sur les données Crabes

On désire utiliser l'analyse discriminante linéaire et l'analyse discriminante quadratique sur les données « crabes » (voir les TPS précédents) afin de déterminer une fonction permettant de distinguer le sexe à partir des mesures FL et RW. Un exemple de code R permettant d'effectuer le travail demandé dans cet exercice est disponible sur le site de l'UV.

1. Expliquer en deux lignes maximum ce que fait chacune des fonctions : `lda`, `qda`, `contour` et `sample` et comparer les fonctions `predict` et `predict.lda`.
2. On effectue l'analyse discriminante linéaire et l'analyse discriminante quadratique en prenant comme échantillon d'apprentissage l'ensemble des données. Tracer les frontières de décision ainsi obtenues. Déterminer les estimations d'erreur en les calculant sur l'ensemble d'apprentissage. Quelle constatation peut-on faire ?
3. À présent, dans chaque classe, on sélectionne *au hasard* 2/3 des exemples pour constituer un ensemble d'apprentissage, les exemples restants formant un ensemble de test. On estimera l'erreur de classification sur l'ensemble d'apprentissage d'une part, et sur l'ensemble de test d'autre part. Répéter plusieurs fois ce processus de sélection d'exemples et d'estimation d'erreur. Que peut-on conclure ?
4. Répéter le calcul précédent en modifiant les proportions du découpage. Quelle constatation peut-on faire ?

Exercice 3.

On dispose de l'ensemble d'apprentissage suivant en dimension $p = 2$:

classe ω_1		classe ω_2	
-0.6	1.5	2.3	2.9
1.3	1.2	0.7	2.4
-0.1	0.1	2.7	1.0
2.4	-1.2	3.6	1.6
0.2	0.9	1.3	2.1

Les données sont supposées suivre dans chaque classe une loi normale de moyenne $\boldsymbol{\mu}_k$ et de matrice de variance Σ_k, $k = 1, 2$. On note π_k la probabilité a priori de la classe ω_k.

1. Estimer les paramètres du modèle sous chacune des hypothèses suivantes :
   (a) $\pi_1 = \pi_2$, $\Sigma_1 = \Sigma_2 = \sigma^2 I$ avec $\sigma^2 \in \mathbb{R}^+$;
   (b) $\Sigma_1 = \Sigma_2$;
   (c) $\Sigma_1 = \begin{pmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{21}^2 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} \sigma_{12}^2 & 0 \\ 0 & \sigma_{22}^2 \end{pmatrix}$;
   (d) Σ_1 et Σ_2 quelconques.
2. Rappeler les noms des classifieurs correspondant à la règle de Bayes dans les quatre situations précédentes.
3. Dans le cas présent, une règle de décision pourra s'exprimer sous la forme

$$d(\mathbf{x}) = \begin{cases} a_1 & \text{si} \quad g(\mathbf{x}) \leq 0, \\ a_2 & \text{si} \quad g(\mathbf{x}) > 0. \end{cases}$$

   où g est une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$. Donner pour chacune des quatre situations l'expression de la fonction g obtenue avec l'ensemble d'apprentissage.