SY09 Printemps 2009
TP 4
Analyses discriminantes quadratique et linéaire

Exercice 1. Règle de Bayes : formes linéaire et quadratique

On considère $g = 2$ classes gaussiennes, en dimension $p = 2$.

1. Donner une équation de la frontière de décision de la règle de Bayes (avec des coûts 0-1) dans chacun des cas suivants :
   (a) $\pi_1 = 0.5$, $\mu_1 = (0,0)'$, $\mu_2 = (1,1)'$, $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$;
   (b) $\pi_1 = 0.1$, $\mu_1 = (0,0)'$, $\mu_2 = (1,1)'$, $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$;
   (c) $\pi_1 = 0.5$, $\mu_1 = (0,0)'$, $\mu_2 = (1,1)'$, $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & -0.3 \\ -0.3 & 1 \end{pmatrix}$;
   (d) $\pi_1 = 0.6$, $\mu_1 = \mu_2 = (1,1)'$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$;
   (e) $\pi_1 = 0.6$, $\mu_1 = (0,0)'$, $\mu_2 = (1,1)'$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.
2. ☐ Suivant chacun des cas précédents, générer des vecteurs $Vect1X_1$, $Vect1X_2$, $Vect2X_1$ et $Vect2X_2$ de taille $n = 500$.
3. ☐ Pour chacun des cas précédents, dessiner $Vect1X_1$ en fonction de $Vect1X_2$ puis $Vect2X_1$ en fonction de $Vect2X_2$, ainsi que la frontière de décision calculée pour les trois premiers cas.
4. ☐ Donner l'expression d'un estimateur de la probabilité d'erreur et calculer les réalisations pour les différents cas. Comparer avec la probabilité d'erreur théorique pour les cas a, b et c et avec la borne de Bhattacharyya pour les cas d et e.

Exercice 2. LDA, QDA sur les données Crabes, et évaluation des performances

On désire utiliser les méthodes d'analyse discriminante linéaire et l'analyse discriminante quadratique pour connaître l'espèce des crabes à partir du jeu de données utilisé dans les TPs précédents. On rappelle que les données des crabes sont disponibles dans la librairie MASS. On peut appliquer l'analyse discriminante linéaire et l'analyse discriminante quadratique en se servant des fonctions LDA et QDA du language R. Les deux variables utilisées sont les variables FL et RW. Utiliser le code R donné sur le site pour effectuer le travail demandé par la suite.

1. Expliquer en deux lignes maximum ce que fait chacune des instructions suivantes du langage R : contour, sample, lda, qda.
2. Exécuter le code suivant et dites quelle est la différence entre predict et predict.lda :

```
> help(predict)
> help(predict.lda)
```

3. ☐ Pour les deux méthodes LDA et QDA, donner un estimateur de la probabilité d'erreur et calculer sa réalisation en se servant des données d'apprentissage utilisées pour effectuer la LDA et la QDA. Que constatez-vous ?
4. ☐ Appliquer la méthode de l'échantillon test avec 2/3 des observations prises pour l'ensemble d'apprentissage. En se basant sur le résultat, choisir le meilleur modèle entre LDA et QDA.
5. ☐ Répéter la méthode de l'échantillon test en choisissant différents échantillons d'apprentissage et de test (par exemple, vous pouvez modifier la taille de l'ensemble d'apprentissage et de test en modifiant les probabilités de tirage aléatoire dans la fonction sample). Que constatez-vous ?

Exercice 3. LDA, QDA, ...

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 avec les paramètres μ_k et 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 aux quatre cas ci-dessus, et donner pour chacun l'expression d'une fonction discriminante g de $\mathbb{R}^2$ dans $\mathbb{R}$, telle que la règle de décision s'écrive sous la forme
$$d(\mathbf{x}) = \begin{cases} a_1, & g(\mathbf{x}) \leq 0, \\ a_2, & g(\mathbf{x}) > 0. \end{cases}$$
3. ☐ Suivant chacun des cas précédents, générer des vecteurs $Vect1X_1$, $Vect1X_2$, $Vect2X_1$ et $Vect2X_2$ de taille $n = 500$.
4. ☐ Estimer les probabilités d'erreur.