Analyses discriminantes quadratique et linéaire

# Exercice I : Règle de Bayes

On suppose que la population est répartie en deux classes, en proportions *π*1 et *π*2 = 1 *− π*1, issues des distributions gaussiennes bivariées *N*(***μ***1*,*Σ1) et *N*(***μ***2*,*Σ2).

## Cas (a), (b), (c), (d), (e) : équation des frontières de décision de la règle de Bayes

En règle générale, pour la règle de Bayes, les fonctions sont de la forme :

gk(x) = – (x-μk)’∑-1(x-μk) – ln(det∑k) + ln лk – ln (2л)

gk(x) = – (x)’∑-1(x) + (μk)’ ∑-1(x) – (μk)’ ∑-1 (μk) – ln(det∑k) + ln лk – ln (2л)

L’équation de la frontière de décision est obtenue en résolvant l’équation gk(x) = gl(x)

a. л1= 0.5, μ1= (0,0)’, μ2= (1,1)’, ∑1= ∑2=

 la frontière est une droite

b. л1= 0.1, μ1= (0,0)’, μ2= (1,1)’, ∑1= ∑2=

la frontière est une droite

c. л1= 0.5, μ1= (0,0)’, μ2= (1,1)’, ∑1= ∑2=

la frontière est une droite

d. л1= 0.6, μ1=μ2= (1,1)’, ∑1= ∑2=

 la frontière est circulaire

e. л1= 0.6, μ1= (0,0)’, μ2= (1,1)’, ∑1= ∑2=

 la frontière est ellipsoïdale

## Tracé des différents cas

## Estimation de la probabilité d’erreur

Dans le cadre de la règle de Bayes, la probabilité d’erreur théorique est donnée par :



La borne de Bhattacharyya est donnée par :

 avec 

A l’aide de ces formules, nous obtenons les valeurs suivantes :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cas** | **Erreur estimée** | **Erreur théorique de Bayes** | **Cas** | **Erreur estimée** | **Borne de Batthacharyya** |
| **A** | **0.24** | **0.240** | **D** | **0.24** | **0.365** |
| **B** | **0.41** | **0.031** | **E** | **0.27** | **0.38** |
| **C** | **0.18** | **0.199** |  |  |  |

# Exercice II : Analyse discriminante sur les données Crabes

## 1.Les fonctions lda, qda, contour et sample. Différence entre predict et predict.lda

**Contour** : trace des lignes de niveaux afin de représenter le troisième paramètre sur un plan déterminé par le premier et le deuxième paramètre.

**Sample :** représente un tirage aléatoire à partir des différents paramètres (taille du tirage, la probabilité du tirage des nombres, les nombres à tirer).

**Lda** : analyse discriminante linéaire. La fonction retourne la probabilité d’obtenir une des variables, le coefficient de linéarité et la moyenne des variables par groupe.

**Qda** : analyse discriminante quadratique. La fonction retourne la probabilité d’obtenir une des variables dans l’un des groupes et la moyenne des variables par groupe.

**Predict** est la fonction générale utilisée pour faire des prédictions à partir de résultats de différentes fonctions d’analyse. **Predict.lda** est une méthode de **predict** destinée particulièrement aux objets de type lda.

## 2.Analyses discriminantes linéaire et quadratique

C:\Users\Pierre\Desktop\SY09\qda_crabes.emfC:\Users\Pierre\Desktop\SY09\lda_crabes.emfNous obtenons les graphiques suivants (lda à droite, qda à gauche) :

On trouve une erreur de 0.095 pour la lda et 0.085 pour la qda. L’aspect similaire des frontières et le peu de différences dans les erreurs s’explique par le fait que les données utilisées sont linéairement séparables. L’analyse quadratique n’est donc pas nécessaire ici pour obtenir des résultats pertinents.

## C:\Users\Pierre\Desktop\SY09\qda_crabes_sample.emfC:\Users\Pierre\Desktop\SY09\lda_crabes_sample.emf3.Analyses discriminantes avec échantillon d’apprentissage

Avec les proportions 2/3,1/3 pour les données d’apprentissage et de test, on retrouve des résultats similaires. Cependant, les valeurs d’erreur sont désormais à 0.025 et 0.035, respectivement pour les analyses linéaire et quadratique. En répétant les calculs 200 fois, on finit par obtenir une erreur moyenne de 0.031925 pour l’analyse linéaire et 0.0288 pour l’analyse quadratique. Outre le fait que ces valeurs sont de nouveaux proches, on remarque que l’erreur est moins importante que pour la question précédente. On peut en déduire que l’utilisation d’un échantillon d’apprentissage distinct de l’échantillon de test augmente la précision du résultat.

## 4.Analyses discriminantes avec échantillon d’apprentissage, différentes proportions

On répète les analyses avec différentes proportions :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LDA** | **Proportions** | **1/5 – 4/5** | **2/5 – 3/5** | **3/5 – 2/5** | **4/5 – 1/5** |
| **Erreur estimée** | 0.081725 | 0.057650 | 0.0392 | 0.0207 |
| **QDA** | **Proportions** | 1/5 – 4/5 | 2/5 – 3/5 | 3/5 – 2/5 | 4/5 – 1/5 |
| **Erreur estimée** | 0.07475 | 0.05305 | 0.035425 | 0.014325 |

On note que plus l’échantillon d’apprentissage est grand, plus l’erreur sera faible. Ce résultat est logique. Concernant le choix d’une méthode par rapport à l’autre, on peut remarquer que qda est en général un peu plus précise. Cependant, la différence n’est pas bien grande, ce qui s’explique par le jeu de données utilisé qui est linéairement séparable. Dans ce cas précis, les deux méthodes peuvent se valoir, néanmoins en règle générale, et surtout si on ne sait pas comment sont organisées, il est préférable d’utiliser l’analyse discriminante quadratique pour s’approcher au plus de la réalité.