Scoring

Décembre 2013

Exercice 1.

On se place dans le cas d'une population $\mathcal{P}$ sur laquelle est définie une partition en $K = 2$ groupes, notés G_1 et G_2. Les proportions théoriques de ces deux groupes sont respectivement p_1 et p_2. On suppose que la distribution de probabilité d'une observation $X = (X_1, \dots, X_p)$ est donnée pour chaque groupe k par une densité de probabilité notée $f_k(x)$. On suppose que

$$f_k(x) = \frac{1}{(2\pi)^{p/2}(det\Sigma_k)^{1/2}} e^{-\frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)}$$

avec $\mu_k \in \Re^p$ le vecteur des moyennes théoriques et Σ_k la matrice de variance-covariance de dimension $p \times p$.

On considère ici *la règle de Bayes de classement* qui consiste à affecter une nouvelle observation x au groupe G_1 si $P(G_1|x) > P(G_2|x)$ où $P(G_k|x)$ est la probabilité a posteriori de G_k.

1. Montrer que dans le cas où l'on fait l'hypothèse $\Sigma_1 = \Sigma_2 = \Sigma$ d'égalité des matrices de variance-covariance, cela revient à affecter une nouvelle observation x au groupe G_1 si

$$x'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2) + ln(\frac{p_1}{p_2}) > 0$$

   On donne le résultat suivant : $(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2) = \mu_1'\Sigma^{-1}\mu_1 - \mu_2'\Sigma^{-1}\mu_2$

   Quel est le nom de cette première règle ?

2. A quelle hypothèse supplémentaire correspond la règle qui consiste à affecter une nouvelle observation x au groupe G_1 si

$$x'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2) > 0$$

   Quel est le nom de cette seconde règle ?

3. On a appliqué ces deux règles sur un jeu de données de taille $n = 136$ avec $n_1 = 37$ observations dans G_1 et $n_2 = 99$ observations dans G_2. Pour la première règle, les probabilités à priori p_1 et p_2 ont été estimées par les proportions des groupes. On obtient les tableaux de confusion suivant :

Table 1: Matrice de confusion obtenue avec la règle de la question 1.

		Groupes d'affectation	
		G_1	G_2
Groupes réel	G_1	26	11
	G_2	1	98

Table 2: Matrice de confusion obtenue avec la règle de la question 2.

		Groupes d'affectation G_1	G_2
Groupes réel	G_1	30	7
	G_2	4	95

Indiquez dans le tableau ci-dessous le pourcentage d'observations du groupe 1 et le pourcentage d'observations du groupe 2 qui sont bien classées en fonction du choix de probabilités a priori. Quel est l'influence du choix des probabilités a priori sur ces pourcentages de bien classés ?

Table 3: Pourcentage d'observations bien classées

	G_1	G_2
p_k proportionnelles		
p_k égales		

Exercice 2.

Les données concernent $n = 1260$ exploitations agricoles réparties en $K = 2$ groupes : le groupe G_1 exploitations saines et le groupe G_2 des exploitations défaillantes. On veut construire un score de détection du risque financier applicable aux exploitations agricoles. Pour chaque exploitation agricole on a mesuré une batterie de critères économiques et financiers et finalement $p = 4$ ratios financiers on étés retenus pour construire le score :

- **R2** : capitaux propres / capitaux permanents,
- **R7** : dette à long et moyen terme / produit brut,
- **R17** : frais financiers / dette totale,
- **R32** : (excédent brut d'exploitation - frais financiers) / produit brut.

La variable qualitative (à expliquer) est donc la variable difficulté de paiement (0=sain et 1=défaillant). Voici les 5 premières lignes du tableau de données.

```
   DIFF    R2    R14    R17    R32
1 saine 0.622 0.2320 0.0884 0.4313
2 saine 0.617 0.1497 0.0671 0.3989
3 saine 0.819 0.4847 0.0445 0.3187
4 saine 0.733 0.3735 0.0621 0.4313
5 saine 0.650 0.2563 0.0489 0.4313
```

1. Combien d'axes discriminants peut-on construire en effectuant une Analyse Factorielle Discriminante (AFD) sur ces données ? Comment est construit cet axe ? (expliquer rapidement sans formules).
2. Interprétez rapidement les résultats ci-dessous de l'AFD.

```
Pouvoir discriminant
----------------------------------
0.5530
```

```
Correlations avec la variable discriminante :
------------------------------------------------------------
R2   0.850
R14 -0.849
R17 -0.444
R32  0.774
```

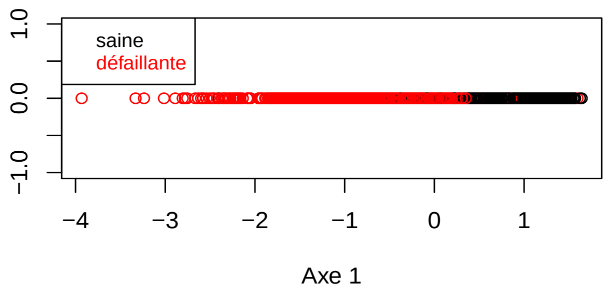

Figure 1: *Plot des exploitations agricoles sur le premier axe discriminant*

3. On effectue maintenant une Analyse Discriminante Linéaire (LDA) sur ces données. Avec l'hypothèse des probabilités a priori égales, on obtient les fonctions linéaires de classement suivantes :

```
              saine defaillante
constant -14.456723  -12.353960
R2        15.037436   10.210386
R14        4.594013    6.603902
R17      110.821176  128.723956
R32       34.013566   25.133656
```

   En déduire la fonction de score linéaire de Fisher. Que va permettre de mesurer ce score ?

4. Calculer le score de la première exploitation agricole du tableau de données. Estimer la probabilité à posteriori que cette exploitation agricole soit saine.

5. Quelle prediction proposez-vous pour cette exploitation ? Cette prédiction elle correcte ?

6. On obtient ainsi une prédiction pour les $n = 1260$ exploitations agricoles. En notant y le vecteur des vrais groupes et *yhat* le vecteur des groupes prédits, on obtient la matrice de confusion suivante :

```
             yhat
y             saine defaillante
  saine         614          39
  defaillante   129         478
```

   En déduire le taux de bon classement, le taux de vrais positifs (TVP) et le taux de vrais négatifs (TVN) de cette règle de décision. Interprétez et critiquez ces taux. Que proposeriez-vous de faire pour modifier les TVN et les TVP ?

7. En tant que statisticien, que feriez-vous d'autre pour évaluer la qualité de ce score et/ou de cette règle de décision ?