TP2 : Analyse discriminante géométrique

2014-2015

On considère un jeu de données concernant $n = 101$ victimes d'infactus du myocarde réparties en $K = 2$ groupes : le groupe G_1 des victimes décédées ($n_1 = 51$) et le groupe G_2 des victimes qui survivent ($n_2 = 50$). On veut construire un score de risque de décès applicable à une nouvelle victime. Pour chaque victime, 7 variables cliniques ont été mesurées :

- **FRCAR** : fréquence cardiaque,
- **INCAR** : index cardiaque,
- **INSYS** : index systolique,
- **PRDIA** : pression diastolique,
- **PAPUL** : pression artérielle pulmonaire,
- **PVENT** : pression ventriculaire,
- **REPUL** : resistance pulmonaire.

La variable qualitative (à expliquer) est donc la variable pronostics (0=survie et 1=décès). Voici les 5 premières lignes du tableau de données.

	PRONO	FRCAR	INCAR	INSYS	PRDIA	PAPUL	PVENT	REPUL
1	SURVIE	90	1.71	19.0	16	19.5	16.0	912
2	DECES	90	1.68	18.7	24	31.0	14.0	1476
3	DECES	120	1.40	11.7	23	29.0	8.0	1657
4	SURVIE	82	1.79	21.8	14	17.5	10.0	782
5	DECES	80	1.58	19.7	21	28.0	18.5	1418

Exercice 1 Dans cet exercice, $p = 3$ variables cliniques ont été retenues pour construire le score : fréquence cardiaque, index systolique, pression diastolique.

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
---------------------------------
0.5037338
```

```
Correlations avec la variable discriminante :
------------------------------------------------------------
FRCAR  0.3258713
INSYS -0.9443729
PRDIA  0.6649957
```

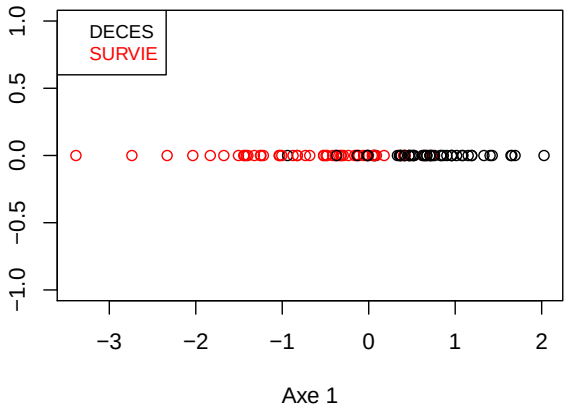

FIGURE 1 – *Plot des victimes sur le premier axe discriminant*

3. On effectue ensuite une analyse discriminante géométrique sur ces données. Quel en est le principe ?
4. On obtient les fonctions linéaires discriminantes suivantes :

```
           DECES  SURVIE
constant -39.808 -47.126
FRCAR      0.516   0.561
INSYS      1.080   1.375
PRDIA      0.636   0.478
```

   En notant $\mathbf{x}$=(FRCAR, INSYS, PRDIA), donner l'expression $L_1(\mathbf{x})$ de la fonction discriminante du groupe 1 (décès) et l'expression $L_2(\mathbf{x})$ de la fonction discriminante du groupe 2 (survie) . En déduire la valeur de ces deux fonctions (les deux scores) de la seconde victime du tableau de données.

5. Quelle prédiction proposez-vous pour cette victime ?
6. Dans le cas particulier de $K = 2$ groupes, on préfère parfois construire une seule fonction linéaire discriminante $\Delta_{1/2}(\mathbf{x})$ appelée fonction discriminante de Fisher. Donner l'expression cette fonction et sa valeur (son score) pour la seconde victime.
7. A quel seuil faut-il comparer ce score pour prédire le groupe de cette victime ? Quel est le lien avec l'AFD ?
8. On obtient ainsi une prédiction pour les $n = 101$ victimes. En notant y le vecteur des vrais groupes et *yhat* le vecteur des groupes prédits, on obtient la matrice de confusion suivante :

```
       y
yhat     DECES SURVIE
  DECES     46     10
  SURVIE     5     40
```

En déduire le taux de mauvais classement, de bon classement, le taux de vrais positifs (la sensibilité) et le taux de vrais négatifs (la spécificité) de cette règle de décision. Interprétez ces taux.

9. En tant que statisticien, que feriez-vous d'autre pour évaluer mieux cette règle de décision ?

Exercice 2 On va maintenant refaire l'analyse discriminante géométrique avec R en conservant toutes les variables.

— Les données se trouvent dans le fichier infarctus.Rdata.
— La procédure `linear_func` permettant le calcul des fonctions linéaires discriminante, a été implémentée dans le fichier "LDA_procedures_chavent.R".
— Le code R du TP se trouve dans le fichier "TP_AD_geom.R".

1. Determiner les fonctions linéaires discriminantes de la règle géométrique avec la fonction `linear_func` . En déduire la fonction linéaire discriminante de Fisher.
2. Calculer la valeur de cette fonction (le score de Fisher) pour un nouveau patient pour lequel :

```
FRCAR INCAR INSYS PRDIA PAPUL PVENT REPUL
90  1.71    19    16  19.5    16   912
```

Quel sera alors la prédiction pour ce patient ?

3. Calculer le score de Fisher des 151 victimes du jeux de données et prédire leur groupe d'appartenance dans un vecteur yhat.
4. Représenter graphiquement ce score et interprétez ce graphique à l'aide des variables initiales.
5. Determiner la matrice de confusion, le taux de mauvais classement, de bon classement, le taux de vrais positifs (la sensibilité) et le taux de vrais négatifs (la spécificité).
6. Retrouvez les résultats de la question précédente en utilisant cette fois la fonction `lda` du package MASS. Pour retrouver la règle géométrique de classement, il faut utiliser l'argument `prior=c(0.5,0.5)` qui indique qu'on fait l'hypothèse d'équiprobabilité des deux groupes. Cette hypothèse vous semble-elle raisonnable sur ces données ?
7. Calculer le taux d'erreur de classement par la méthode de l'échantillon test (avec 70 victimes dans l'échantillon d'apprentissage).
8. Calculer le taux d'erreur de classement en appliquant 100 fois la méthode de l'échantillon test (avec 70 victimes dans l'échantillon d'apprentissage).
9. Calculer le taux d'erreur de classement par la méthode de la validation croisée (leave one out).