TP3 : Analyse discriminante probabiliste

2014-2015

Les données concernent $n = 1260$ des exploitations agricoles réparties en $K = 2$ groupes : le groupe G_1 des exploitations saines ($n_1 = 653$) et le groupe G_2 des exploitations défaillantes ($n_2 = 607$). 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. La variable qualitative (à expliquer) est la variable difficulté de paiement (0=sain et 1=défaillant). Ces données sont étudiées en détail dans l'article de Dominique Desbois, publié dans la revue CSBIGS :

`http://www.bentley.edu/centers/sites/www.bentley.edu.centers/files/csbigs/Desbois.pdf`

Exercice 1. Dans cet exercice $p = 4$ 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.

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. On veut effectuer une analyse discriminante probabiliste sur ces données. Quel en est le principe ?
2. On choisit d'effectuer une analyse discriminante linéaire (LDA). Quelles sont les hypothèses faites sur les données lorsqu'on applique cette méthode ? Quelle hypothèse faut-il ajouter pour retrouver l'analyse discriminante géométrique ?
3. En notant $\mathbf{x}$=(R2,R7,R17,R32) on obtient deux fonctions linéaires discriminante $L_1(\mathbf{x})$ (pour le groupe G1 des exploitations saines) et $L_2(\mathbf{x})$ (pour le groupe G2 des exploitations défaillantes) dont les coefficients sont :

	saine	défaillante
constant	-21.04500	-15.98344
R2	22.64520	16.09855
R7	11.23535	10.81185
R17	135.18344	154.53170
R32	37.59187	26.57469

En déduire les fonctions linéaires discriminantes de l'approche géométrique.

4. En déduire la fonction linéaire discriminante $\Delta_{2/1}(\mathbf{x}) = L_2(\mathbf{x}) - L_1(\mathbf{x})$. Que va permettre de mesurer cette fonction score ?
5. Calculer avec cette fonction le score de la première exploitation agricole. A quelle seuil doit-on comparer ce score si on veut prédire le groupe de cette exploitation ?
6. En déduire une estimation la probabilité (à posteriori) que cette exploitation agricole soit défaillante. A quelle seuil doit-on comparer cette probabilité si on veut prédire le groupe de cette exploitation ?
7. Quelle prediction proposez-vous pour cette exploitation ? Cette prédiction est-elle correcte ?
8. 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 :

	y	
yhat	saine	défaillante
saine	589	132
défaillante	64	475

En déduire le taux de bon classement, le taux de vrais positifs (sensibilité) et le taux de vrais négatifs (spécificité). Que proposeriez-vous de faire pour augmenter la spécificité ? Quelle serait la conséquence sur la sensibilité ?

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

Exercice 2 On va maintenant refaire l'analyse discriminante probabiliste avec R.

— Les données se trouvent dans le fichier farms.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_proba.R".

On conserve comme variables explicatives les ratios financiers :

```
stockholders' equity / invested capital [r2]
short-term debt / total debt [r3]
long and medium-term debt / gross product [r7]
short-term debt / circulating asset [r14]
financial expenses / total debt [r17]
```

```
financial expenses / gross product [r18]
financial expenses / EBITDA [r21]
(EBITDA - financial expenses) / gross product [r32]
immobilized assets / gross product [r36]
```

1. On applique d'abord la méthode LDA. Determiner la fonction linéaire discriminante $\Delta_{2/1}(\mathbf{x}) = \beta_0 + \beta'\mathbf{x}$ et calculer les scores des 1260 exploitations agricoles. Proposer une représentation graphique et interpréter ce graphique en fonction des variables initiales.
2. Quelle représentation graphique équivalente peut-être obtenue en utilisant uniquement les fonctions `lda` et `predict.lda` du package `MASS`.
3. Dans le cas particulier de deux groupes, comment calculer les probabilités à posteriori à partir des scores $\Delta_{2/1}(\mathbf{x})$? Estimer les probabilités (à posteriori) des exploitations agricoles à partir des scores $\Delta_{2/1}(\mathbf{x})$. Retrouver ces probabilités en utilisant uniquement avec les fonctions `lda` et `predict.lda` du package `MASS`.
4. A l'inverse, comment calculer les scores $\Delta_{2/1}(\mathbf{x})$ à partir des probabilités à posteriori ?
5. Déterminer le taux de mauvais classement, de vrais positifs (la sensibilité), de vrais négatifs (la spécificité) de la méthode LDA par la méthode de validation croisée (leave one out).
6. Déterminer le taux de mauvais classement, de vrais positifs (la sensibilité), de vrais négatifs (la spécificité) de la méthode QDA par la méthode de validation croisée (leave one out).
7. Appliquer ensuite la méthode de régression logistique avec la fonction `glm` et déterminer le taux d'erreur apparent de la règle du maximum à posteriori.
8. En utilisant le package `ROCR`, tracer la courbe des taux d'erreurs (apparents) de la régression logistique en fonction des seuils en prenant comme score la probabilité à posteriori de défaillance (défaillance=1) puis en prenant comme score cette probabilité sur l'échelle logit (le score $\Delta_{2/1}(\mathbf{x})$). Le choix d'un seuil de 0.5 dans le premier cas et d'un seuil de 0 dans le second, pour définir la règle de décision vous semble-t-il approprié ?
9. En prenant comme score la probabilité à posteriori de défaillance, tracez la courbe ROC (en faisant apparaître une couleur indiquant le seuil), puis calculer le critère AUC de ce score.
10. Constuire aléatoirement un échantillon d'apprentissage et un échantillon test (en mettant 900 exploitations dans l'échantillon d'apprentissage). Déterminer alors par la méthode de l'échantillon test : le taux d'erreur, la courbe ROC, le crière AUC, pour les méthodes LDA, QDA et régression logistique.

Exercice 3 : analyse discriminante avec SAS.

1. Lire le code SAS suivant et déterminer quelle méthode d'analyse discriminante est appliquée :

```
proc discrim data=don.infarctus pool=no;
   class PRONO;
run;

proc discrim data=don.infarctus pool=no;
   class PRONO;
   priors proportional;
run;

proc discrim data=don.infarctus  pool=YES;
   class PRONO;
run;

proc discrim data=don.infarctus pool=YES;
   class PRONO;
   priors proportional;
   run;
```

2. On exécute maintenant dans SAS le code suivant :

```
proc discrim data=don.infarctus pool=YES list crossvalidate
 distance out=sortie scores;
   class PRONO;
   priors proportional;
run;
```

   Les résultats sont dans le fichier "Resultats SAS exercice 2.pdf". Interpétez ces résultats.