

## Examen de Scoring Appliqué à la Détection du Risque

Durée : 2 heures.

L'utilisation d'un poste informatique, de logiciels de statistique, du cours et d'internet sont autorisés.

## Exercice 1.

Table 1 indique pour dix clients bancaires : (i) le nombre de jours de découvert du compte courant dans l'année écoulée (ii) l'existence d'une épargne : O si le client détient un livret et N sinon (iii) la classe : 2 si le client est interdit bancaire et 1 dans le cas contraire.

| client | découvert | épargne | classe |
|--------|-----------|---------|--------|
| 1      | 10        | O       | 1      |
| 2      | 56        | N       | 1      |
| 3      | 0         | O       | 1      |
| 4      | 18        | N       | 1      |
| 5      | 76        | O       | 2      |
| 6      | 28        | O       | 1      |
| 7      | 128       | N       | 2      |
| 8      | 67        | O       | 2      |
| 9      | 25        | O       | 1      |
| 10     | 9         | N       | 1      |

Table 1: Echantillon d'apprentissage : dix clients bancaires décrits par découvert & épargne (prédicteurs) et répartis selon la classe : interdit bancaire ou non (variable à prédire)

*Notations concernant l'échantillon.*  $z_{i,k} \in \{0, 1\}$  indique si le client  $i$  appartient ( $z_{i,k} = 1$ ) ou non ( $z_{i,k} = 0$ ) à la classe  $k$ ,  $y_i \in \{0, 1\}$  indique si ce client possède ( $y_i = 1$ ) ou non ( $y_i = 0$ ) un livret d'épargne et  $x_i \in \mathbb{N}$  désigne le nombre de jours de découvert. On note  $n$  la taille de l'échantillon ( $n=10$ ).

*Notations concernant la population.* Pour  $k \in \{1, 2\}$ ,  $\pi_k$  désigne la proportion de clients de la banque appartenant à la classe  $k$  et  $\alpha_k$  la proportion de clients de cette classe possédant un livret d'épargne.

1. Interprétez les statistiques suivantes :

- (a)  $n_k = \sum_{i=1}^{10} z_{i,k}$
- (b)  $\bar{x}_k = \sum_{i=1}^{10} z_{i,k} x_i / n_k$
- (c)  $\nu_k = \sum_{i=1}^{10} z_{i,k} y_i$
- (d)  $\nu = \sum_{k=1}^2 \nu_k$

*Modèle de Poisson conditionnel et hypothèse d'indépendance locale.* Quel que soit  $k \in \{1, 2\}$ , on suppose que le nombre de jours de découvert d'un client de la classe  $k$  est distribué selon une loi de Poisson  $\mathcal{P}(\lambda_k)$  et qu'il est indépendant de la détention d'un livret d'épargne.

On note  $\theta = (\pi_1, \pi_2, \alpha_1, \alpha_2, \lambda_1, \lambda_2)$  le paramètre du modèle.

2. Montrez que la log-vraisemblance de  $\theta$  s'écrit :  $\ell(\theta) = \ell_1(\pi_1, \pi_2) + \ell_2(\alpha_1, \alpha_2) + \ell_3(\lambda_1, \lambda_2)$  où :

$$\begin{aligned} \ell_1(\pi_1, \pi_2) &= \sum_{i=1}^{10} \sum_{k=1}^2 z_{i,k} \log \pi_k \\ \ell_2(\alpha_1, \alpha_2) &= \sum_{i=1}^{10} \sum_{k=1}^2 z_{i,k} \{y_i \log \alpha_k + (1 - y_i) \log(1 - \alpha_k)\} \\ \ell_3(\lambda_1, \lambda_2) &= \sum_{i=1}^{10} \sum_{k=1}^2 z_{i,k} \{-\lambda_k + x_i \log \lambda_k - \log(x_i!)\}. \end{aligned}$$

3. Montrez que l'estimation de  $\theta$  par Maximum de Vraisemblance conduit à :  $\hat{\pi}_k = n_k/n$ ,  $\hat{\alpha}_k = \nu_k/n_k$ ,  $\hat{\lambda}_k = \bar{x}_k$ .

4. Déterminez la valeur de  $BIC$  du modèle estimé.

Pour  $k \in \{1, 2\}$ , on considère le score :  $S_k(x, y) = \log \hat{\pi}_k + y \log \hat{\alpha}_k + (1 - y) \log(1 - \hat{\alpha}_k) - \hat{\lambda}_k + x_i \log \hat{\lambda}_k$  où  $x$  représente le nombre de jours de découvert d'un nouveau client et où  $y \in \{0, 1\}$  indique l'existence d'une épargne :  $y = 1$  s'il possède un livret et  $y = 0$  sinon.

- Montrez que la règle du MAP revient à affecter  $(x, y)$  à la classe  $k$  correspondant au score  $S_k(x, y)$  le plus élevé.
- A quelle classe affecteriez vous M. Phong qui présente cinquante jours de découvert et qui détient un livret ?

*Modèle Gaussien conditionnel et hypothèse d'indépendance locale.* On suppose désormais que le nombre de jours de découvert d'un client de la classe  $k$  ( $k \in \{1, 2\}$ ) est distribué selon une loi normale  $\mathcal{N}(\mu_k, \sigma_k^2)$  et qu'il est indépendant de la détention d'un livret d'épargne.

On note  $\psi = (\pi_1, \pi_2, \alpha_1, \alpha_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$  le paramètre du modèle.

- Estimez  $\psi$  par Maximum de Vraisemblance.
- Le modèle Gaussien conditionnel est-il meilleur selon  $BIC$  que le modèle de Poisson conditionnel ?
- A quelle classe affecteriez vous M. Phong sous ce nouveau modèle ?

### Exercice 2.

Table 2 donne pour dix clients de Lloyds (i) le score d'attrition (probabilité de passer à la concurrence) déterminé en Juin 2015 (ii) le statut observé en Janvier 2016 : 1 si le client est passé à la concurrence et 0 sinon.

| <i>identifiant client</i>            | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   |
|--------------------------------------|------|------|------|------|------|------|------|------|------|------|
| <i>score d'attrition (Juin 2015)</i> | 0,18 | 0,29 | 0,71 | 0,53 | 0,43 | 0,34 | 0,90 | 0,18 | 0,47 | 0,64 |
| <i>statut observé (Janvier 2016)</i> | 0    | 1    | 0    | 1    | 0    | 0    | 1    | 0    | 1    | 1    |

Table 2: - Score d'attrition en juin 2015 & statut en juin 2016 de dix clients de Lloyds

Pour un seuil  $s$  donné, on affecte à la classe à risque (classe 1) tout client dont le score d'attrition est supérieur ou égal à  $s$  et à la classe de non risque (classe 0) tout client dont le score d'attrition est strictement inférieur à  $s$ .

- On fixe  $s = 0,5$ .
  - Dressez le tableau de contingence : statut observé  $\times$  classe estimée.
  - Déterminez le taux de faux positifs et le taux de vrais positifs.
  - Estimez l'erreur de classement par resubstitution.
- Représentez la courbe ROC obtenue en considérant  $s = 0,1; 0,2; \dots; 0,9$ .
- Déterminez la valeur d' $AUC$ .

### Exercice 3.

On considère le jeu de données finance du package `Rmixmod`. Les variables quantitatives : `EBITDA.Total.Assets`, `Value.Added.Total.Sales`, `Quick.Ratio`, `Accounts.Payable.Total.Sales` sont prises pour prédicteurs et `Health` pour variable à prédire.

On se donne pour objectifs : (i) estimer un modèle de régression logistique sur les firmes de 2002 (ii) prédire grâce au modèle estimé la santé des firmes de 2003 (iii) comparer la santé prédite des firmes de 2003 à la santé observée.

- Décrivez le modèle de régression logistique dans ce contexte.
- Estimez le paramètre de ce modèle en prenant les entreprises de 2002 pour échantillon d'apprentissage.
- Déterminez une fonction de score issue de ce modèle permettant de prédire la faillite.
- A quelle classe : `Healthy/Bankruptcy` cette fonction affecte-t-elle la première firme de 2003 ?
- A quel taux d'erreur de classement des firmes de 2003 la fonction de score considérée conduit-elle ?

### Exercice 4.

On considère les données `banknote` du package `mclust` de R. On modélise les données de chaque classe de billet : `genuine/counterfeit` par une loi normale multidimensionnelle.

- Quel est le meilleur modèle selon le critère  $BIC$  : homoscédastique/hétéroscédastique ?
- On considère un modèle homoscédastique.
  - Estimez<sup>a</sup> le poids, le centre et la matrice des covariances de chaque classe de billet.

<sup>a</sup>par Maximum de Vraisemblance

- (b) Quelle fonction de score permet de décider de l'authenticité d'un billet ?
- (c) A quelle classe cette fonction affecte-t-elle un billet présentant les caractéristiques suivantes : Length=214, 5, Left=130, 1, Right=130, 0, Bottom=8, 3, Top=10, 5, Diagonal=140, 1 ?
- (d) Interprétez la sortie de la suite de commandes suivantes.

```
library(mclust, rmixmod)
data(banknote)

learn_homo <- mixmodLearn(banknote[, -1], knownLabels=banknote[, 1],
models = mixmodGaussianModel(listModels=c("Gaussian_pk_L_C")),
criterion = c("BIC", "CV"))

prediction <- mixmodPredict(data = data.frame(Length=215.0, Left=129.1, Right=131.2,
Bottom=8.2, Top=10.4, Diagonal=139.9), classificationRule = learn_homo["bestResult"])
print(prediction)
```

---