Devoir surveillé de data mining

2 ème année du cycle de formation d'ingénieurs

Durée de l'épreuve : 1h30 - Documents non autorisés
Nombre de pages : 2 - Date de l'épreuve : 2 mars 2024

Exercice 1 : On considère le tableau ci-dessous qui décrit un ensemble $\mathcal{E} = \{i_1, i_2, \ldots, i_8\}$ à l'aide de deux variables quantitatives X^1 et X^2 et d'une variable qualitative Y à deux modalités, notées A et B.

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
X^1	1	2	-2	3	0	0	1	3
X^2	3	0	0	3	1	-1	-1	-1
Y	B	A	A	B	A	A	B	B

Chaque individu est muni du poids $1/8$. On note g, g_A et g_B les centres de gravité respectifs de l'ensemble $\mathcal{E}$ et des classes I_A et I_B définies par les modalités A et B, respectivement. Dans ce qui suit, les analyses discriminantes de ces données sont réalisées à l'aide du logiciel R.

1 – Sachant que les variances non corrigées de X^1 et de X^2 sont égales et valent 5/2, déterminer la matrice de variance non corrigée[1] du couple de variables X^1 et X^2. On notera V cette matrice.

2 – Calculer la matrice variance interclasses (non corrigée) du couple de variables X^1 et X^2.

3 – On note W (resp. W^c) la matrice variance intraclasses non corrigée (resp. corrigée). On rappelle que dans le cas de 2 classes, on a la formule $W^c = \frac{n}{n-2}W$ où n désigne la taille de l'ensemble $\mathcal{E}$ étudié. Déduire des résultats précédents que $W^c = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$.

4 – Expliquer pourquoi il n'existe qu'un seul axe facteur discriminant (non trivial) pour l'Analyse Factorielle Discriminante (AFD) des données. On notera b cet unique facteur.

5 – On rappelle que R utilise la métrique $(W^c)^{-1}$ pour effectuer l'AFD. En choisissant son orientation de façon appropriée, montrer que le vecteur b est égal à $\frac{1}{\sqrt{21}}\begin{pmatrix} 3 \\ 1 \end{pmatrix}$.

1. On rappelle que la variance empirique corrigée (resp. non corrigée) d'une variable X est égale $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (resp. $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$), où n désigne la taille de l'échantillon. On définit de même la covariance non-corrigée (resp. corrigée) entre deux variables.

Par la suite, P désigne un point arbitraire du plan où l'axe des abscisses (resp. des ordonnées) indique les valeurs de X^1 (resp. X^2). On note $x = (x_1 \quad x_2)'$ le vecteur des coordonnées de P.

6 – Pour tout point P du plan, montrer que $z(P) = \frac{1}{\sqrt{21}}(3x_1 + x_2 - 3.5)$.

7 – En déduire les affectations des individus $i_1, ..., i_8$ aux classes A et B ainsi que le taux d'erreur de ce modèle.

8 – Dans cette question, la méthode de classification bayésienne est appliquée en supposant que les individus de chacune des deux classes suivent la même loi gaussienne dont la matrice variance est estimée par W^c. De plus, on suppose que la probabilité a priori de la classe A est égale à α avec $\alpha \in]0, 1[$. Ecrire la commande de R qui permet d'appliquer cette méthode aux données. On notera "don" le data.frame dans lequel sont enregistrées les données.

Exercice 2 : On consdère les données `Canines` décrivant les caractéristiques de 27 races de chiens au moyens de 6 variables qualitatives. Ci-dessous les données des 6 premières races :

```
> head(Base_canines)
          taille poids velocite intellig affect agress
beauceron    T++    P+      V++       I+    Af+    Ag+
basset       T-     P-       V-       I-    Af-    Ag+
ber_allem    T++    P+      V++      I++    Af+    Ag+
boxer        T+     P+       V+       I+    Af+    Ag+
bull-dog     T-     P-       V-       I+    Af+    Ag-
bull-mass    T++    P++      V-      I++    Af-    Ag+
```

1 – Evaluer la proximité entre les 2 premiers individus en utilisant une mesure adaptée à des données qualitatives.

2 – Quelles sont les différentes approches possibles de classification automatique dans le cas de données qualitatives.