

Projet (I2) à rendre le 3 janvier 2022

Ce devoir est à réaliser par groupe de 1, 2, ou 3 élèves. Il faudra rendre un rapport en pdf, un fichier texte pour les prédictions du premier exercice et les codes utilisés en annexe.

1 Régression binaire

Le fichier “simu.txt” contient une simulation d’un modèle binaire :

- 2000 observations.
- 2 variables explicatives X_1 et X_2 .
- La variable à expliquer Y est à valeurs dans $\{1, 2\}$.

1. Déterminer le meilleur modèle possible pour modéliser $P(Y|X_1, X_2)$.
2. Utiliser le fichier “xsimutest” pour prédire 1000 observations de la variable Y du fichier de test. Il faut sauvegarder ces prédictions dans un fichier “.txt”, les séparateurs devront être des retours chariot. Par exemple, si les 1000 prédictions du fichier de test de Y sont dans le vecteur “predictions”, on le sauvera avec la procédure R :

```
write.table(predictions, "predictions.txt", row.names=F, col.names=F)
```

Il faudra joindre ce fichier à votre devoir.

2 Analyse en composantes principales

On dispose des mesures suivantes sur plusieurs types de voitures vendues en 2015 dans le fichier “voitures”. Vous devez effectuer une analyse en composantes principales à l’aide de toutes les variables.

1. Quel est le pourcentage d’inertie expliquée par les trois premiers facteurs? Par le premier plan factoriel?
2. Interpréter les 2 axes principaux à partir des corrélations des variables avec ces axes.
3. Représentez les individus sur le premiers plan factoriel et répondez aux questions suivantes :
 - (a) Les individus sont-ils bien représentés sur le premier plan factoriel?
 - (b) Quelles sont les caractéristiques des individus en haut du graphe?
 - (c) Quelles sont les caractéristiques des individus à droite du graphe?
 - (d) Quelles sont les caractéristiques des individus en bas à gauche du graphe?
 - (e) Peut-on dire que les individus PEUGEOTRCZ et JAGUARF ont un profil semblable? Si oui quel est-il?
 - (f) Peut-on dire que les individus LANCIA et LANDROVER ont un profil semblable? Si oui quel est-il?
 - (g) Interpréter la représentation graphique des individus.

3 Classification

1. Avec les données sur les voitures, réalisez une (ou des) classifications avec la méthode des kmean et interprétez les résultats obtenus.
2. Faire une classification hiérarchique avec la méthode de “Ward” . Interpréter le dendrogramme. En combien de classes aurait-on envie de couper ce graphique?

4 Races de chiens

On dispose du fichier “chiens” qui détaille les caractéristiques des chiens en fonction de la race. Les variables sont les suivantes :

- TAI (Taille) : 3 = *grande*, 2 = *moyenne* et 1 = *petite*.
 - POI (Poids) : 3 = *lourd*, 2 = *moyen* et 1 = *lger*.
 - VEL (Vélocité) : 3 = *rapide*, 2 = *moyenne* et 1 = *lent*.
 - INT (Intelligence) : 3 = *grande*, 2 = *moyenne* et 1 = *faible*.
 - AFF (Affectuosité) : 2 = *grande* et 1 = *faible*.
 - AGR (Agressivité) : 2 = *grande* et 1 = *faible*.
 - FON (Fonction) : 3 = *garde*, 2 = *chasse* et 1 = *compagnie*.
1. En prenant la variable FON comme variable supplémentaire, faire une anayse des correspondances multiples de ces données.
 2. En déduire une description des différentes races de chiens.