Devoir surveillé de classification supervisée

Durée : 2h, tout type de document autorisé, **accès à internet interdit**

Rendu : vous répondrez aux questions posées sur votre copie papier, en fin de devoir vous enverrez votre programme R à l'adresse mail vincent.vandewalle@univ-lill2.fr.

Fichiers : les fichiers de données à analyser **ronfle.csv** et **clients.csv**, ainsi que le sujet du contrôle au format pdf se trouvent à l'url http ://vincent.vandewalle.perso.sfr.fr/DSPolytech/.

Exercice 1 (12 pts)

On dispose de données sur 100 patients :

— RONFLE : 1 si ronfle, 0 sinon
— AGE : age du patient
— POIDS : poids du patient en kg
— TAILLE : taille du patient en cm
— ALCOOL : le nombre de verres d'alcool bus par jour par le patient
— SEXE : 1 si femme, 0 si homme
— TABA : 1 si fumeur, 0 si non fumeur

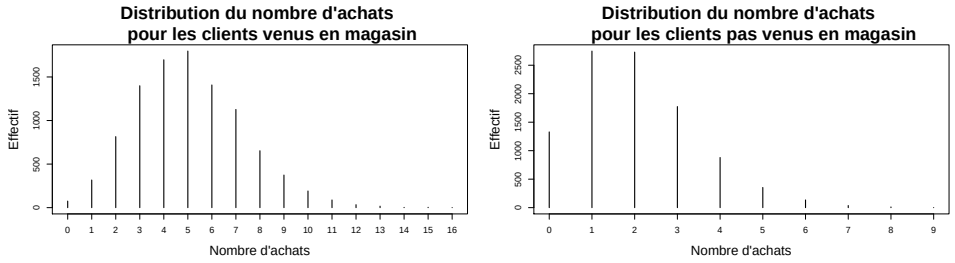
Les données sont présentées dans le fichier ronfle.csv. L'objectif est de prédire le ronflement à partir des autres variables.

1. Résumer le jeu de données
2. Quel est le pourcentage de ronfleurs parmi les 100 patients ?
3. Résumer la variable AGE ? Commenter.
4. Faire un graphique permettant d'illustrer le lien entre la variable AGE et la variable RONFLE.
5. Produire un résumé de la variable AGE en fonction des diverses valeurs possibles de RONFLE. Commenter le résultat obtenu.
6. Quel test statistique permet de tester l'effet de la variable RONFLE sur la variable AGE ?
7. Effectuer une régression logistique visant à expliquer le ronflement en fonction de l'âge.
8. Donner l'expression mathématique de $P(\text{RONFLE} = 1/\text{AGE} = x)$
9. En déduire la probabilité qu'un patient de 70 ans ronfle.
10. L'effet de l'âge sur le ronflement est-il significatif ?
11. Évaluer les performances du modèle sur les mêmes données qui ont servies à l'apprendre.
12. Quel est le risque d'une telle évaluation ? Quelle méthode préconiseriez-vous ici ?
13. Pour améliorer les capacités prédictives du modèle, ajuster le modèle avec toutes les variables. Toutes variables ont-elles un effet significatif ?
14. Réaliser une sélection de variables pas à pas. Interpréter le modèle obtenu. Par combien est multiplié le risque de ronfler quand la consommation d'alcool journalière augmente de 1 verre ?
15. Proposer et mettre en œuvre une méthode de visualisation des données sur l'axe de séparation maximale des classes (ronfleur ou non), à défaut on pourra réaliser une ACP.
16. Proposer et mettre en œuvre une autre méthode de classification supervisée étudiée en cours.

Exercice 2 (8 pts)

On souhaite prédire l'achat d'un produit en magasin par un client suite à la réception d'une offre publicitaire. Pour cela on dispose du nombre d'achats effectués par chacun des clients au cours des 12 derniers mois (noté X).

Une étude antérieure menée sur 10 000 clients dans laquelle on disposait à la fois du nombre d'achats par client (avant la réception de l'offre publicitaire !) et du fait qu'il ait répondu positivement ($Y = 1$) ou négativement ($Y = 0$) à l'offre publicitaire, a donné les résultats suivants :



1. Aux vues des données quelle hypothèse de la LDA et de la QDA n'est pas vérifiée ?
2. Par la suite on modélise les données de la façon suivante :

$$P(Y = 1) = \pi_1 \ , \ P(Y = 0) = \pi_0 \ , \ X/Y = 1 \sim \mathcal{P}(\lambda_1) \text{ et } X/Y = 0 \sim \mathcal{P}(\lambda_0).$$

avec $\mathcal{P}(\lambda)$ la loi de Poisson de paramètre λ ($\lambda > 0$). On rappelle que si $X \sim \mathcal{P}(\lambda)$ alors :

$$\forall x \in \mathbb{N},\ P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

En quoi cette modélisation vous semble-t-elle plus raisonnable aux vues des données ?

3. Donner l'expression mathématique de $P(Y = 1/X = x)$.
4. On rappelle que l'estimateur du maximum de vraisemblance du paramètre d'une loi de Poisson est la moyenne. Estimer π_1, π_0, λ_1 et λ_0 à partir des données présentes dans le fichier clients.csv.
5. En déduire, pour un client ayant effectué 3 achats au cours des 12 derniers mois, sa probabilité de venir en magasin suite à la réception de la publicité ? Vous pourrez vous aider de la fonction `dpois`.
6. Évaluer les performances du modèle ainsi produit.
7. Lien avec un autre modèle déjà étudié :
   (a) Montrer que $\ln P(Y = 1, X = x) = \ln \pi_1 - \lambda_1 - \ln(x!) + x \ln \lambda_1$
   (b) En déduire que :
$$\ln \frac{P(Y = 1/X = x)}{P(Y = 0/X = x)} = \ln \frac{\pi_1}{\pi_0} - \lambda_1 + \lambda_0 + x \ln \frac{\lambda_1}{\lambda_0}.$$
   (c) A quoi l'expression précédente vous fait-elle penser ? Dans ce cas quelle autre méthode vue en cours pourrait-on utiliser pour estimer directement $P(Y = 1/X = x)$?
   (d) Donner l'expression mathématique de $P(Y = 1/X = x)$ estimée à partir de cette seconde méthode appliquée au fichier clients.csv.