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

```
> summary(ronfle)
```

```
     RONFLE          AGE            POIDS            TAILLE
 Min.   :0,00   Min.   :23,00   Min.   : 42,00   Min.   :158,0
 1st Qu.:0,00   1st Qu.:43,00   1st Qu.: 77,00   1st Qu.:166,0
 Median :0,00   Median :52,00   Median : 95,00   Median :186,0
 Mean   :0,35   Mean   :52,27   Mean   : 90,41   Mean   :181,1
 3rd Qu.:1,00   3rd Qu.:62,25   3rd Qu.:107,00   3rd Qu.:194,0
 Max.   :1,00   Max.   :74,00   Max.   :120,00   Max.   :208,0
     ALCOOL           SEXE           TABA
 Min.   : 0,00   Min.   :0,00   Min.   :0,00
 1st Qu.: 0,00   1st Qu.:0,00   1st Qu.:0,00
 Median : 2,00   Median :0,00   Median :1,00
 Mean   : 2,95   Mean   :0,25   Mean   :0,64
 3rd Qu.: 4,25   3rd Qu.:0,25   3rd Qu.:1,00
 Max.   :15,00   Max.   :1,00   Max.   :1,00
```

2. Quel est le pourcentage de ronfleurs parmi les 100 patients ?

```
> effectifs = table(ronfle$RONFLE)
> frequences = effectifs/sum(effectifs)
> frequences[2]
```

```
   1
0,35
```

Les poncentage de ronfleur parmi les 100 patients est égal à 35%.

3. Résumer la variable AGE ? Commenter.

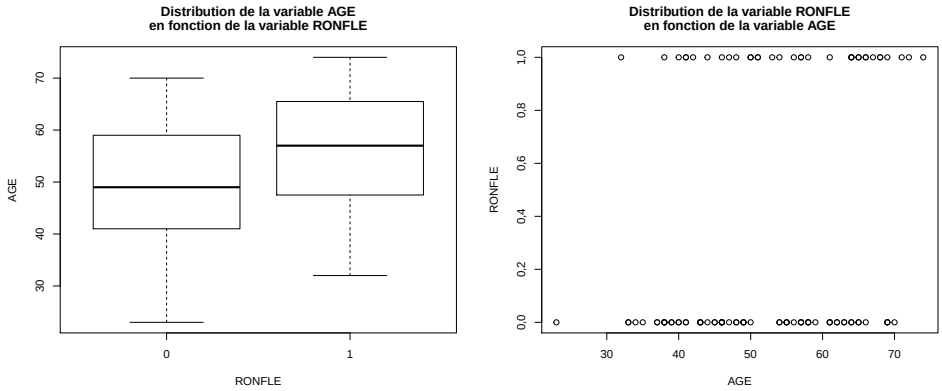
```
> summary(ronfle$AGE)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23,00   43,00   52,00   52,27   62,25   74,00
```

L'âge des patients varie entre 23 et 74 ans. L'âge moyen est de 52,27 ans et l'âge médian est de 52 ans.

4. Faire un graphique permettant d'illustrer le lien entre la variable AGE et la variable RONFLE.

```
> par(mfcol = c(1,2))
> boxplot(AGE ~ RONFLE, data = ronfle, xlab = "RONFLE", ylab = "AGE",
+         main = c("Distribution de la variable AGE","en fonction de la variable RONFLE"))
> plot(RONFLE ~ AGE,data = ronfle,
+      main = c("Distribution de la variable RONFLE","en fonction de la variable AGE"))
```



5. Produire un résumé de la variable AGE en fonction des diverses valeurs possibles de RONFLE. Commenter le résultat obtenu.

```
> by(ronfle$AGE,ronfle$RONFLE,summary)

ronfle$RONFLE: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23,00   41,00   49,00   50,26   59,00   70,00
------------------------------------------------------------
ronfle$RONFLE: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   32,0    47,5    57,0    56,0    65,5    74,0
```

On remarque que les ronfleurs sont légèrement plus vieux (en moyenne 56 ans) que le non ronfleurs (en moyenne 50,26 ans).

6. Quel test statistique permet de tester l'effet de la variable RONFLE sur la variable AGE ?
   Une analyse de la variance à un facteur (ANOVA) permet de tester l'effet de la variable RONFLE sur la variable AGE.
   Résultat de l'ANOVA (non demandé)

```
> anova = lm(AGE ~ as.factor(RONFLE), data = ronfle)
> summary(anova)

Call:
lm(formula = AGE ~ as.factor(RONFLE), data = ronfle)

Residuals:
    Min      1Q  Median      3Q     Max
-27,262  -9,065  -1,262   9,250  19,738

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          50,262      1,379  36,452   <2e-16 ***
as.factor(RONFLE)1    5,738      2,331   2,462   0,0156 *
---
Signif. codes:  0
```

   Ici la variable RONFLE à un effet significatif sur la variable AGE.

7. Effectuer une régression logistique visant à expliquer le ronflement en fonction de l'âge.

```
> logist1 = glm(RONFLE ~ AGE, family = "binomial", data = ronfle)
> logist1
```

```
Call:  glm(formula = RONFLE ~ AGE, family = "binomial", data = ronfle)

Coefficients:
(Intercept)          AGE
   -3,11574      0,04696

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:            129,5
Residual Deviance: 123,5         AIC: 127,5
```

8. Donner l'expression mathématique de $P(\text{RONFLE} = 1/\text{AGE} = x)$

$$P(\text{RONFLE} = 1/\text{AGE} = x) = \frac{e^{-3,116+0,047x}}{1 + e^{-3,116+0,047x}}$$

9. En déduire la probabilité qu'un patient de 70 ans ronfle.
   La probabilité qu'un patient de 70 ans ronfle est de :

$$P(\text{RONFLE} = 1/\text{AGE} = 70) = \frac{e^{-3,116+0,047\times 70}}{1 + e^{-3,116+0,047\times 70}} = 0,543$$

10. L'effet de l'âge sur le ronflement est-il significatif ?

```
> summary(logist1)

Call:
glm(formula = RONFLE ~ AGE, family = "binomial", data = ronfle)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1,2510  -0,9610  -0,7212   1,2045   1,8946

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3,11574    1,09564  -2,844  0,00446 **
AGE          0,04696    0,01988   2,362  0,01816 *
---
Signif. codes:  0
```

   La probabilité critique associée au coefficient de la variable AGE est de 0,0182, ce qui est inférieur à $0,05$ donc on rejette donc l'hypothèse d'absence d'effet de la variable AGE sur la présence de ronflement.

11. Évaluer les performances du modèle sur les mêmes données qui ont servies à l'apprendre.
    Pour mesurer les performance du modèle on réalise la table de confusion en classant les individus dans la classe de plus grande probabilité. On en déduit la table de confusion, la sensibilité, la spécificité, le taux de bon classement et le taux de mauvais classement :

```
> # Valeurs prédites
> Ypred = (logist1$fitted.values > 0.5)*1
> # Table de confusion
> tableConfusion = table(Y = ronfle$RONFLE, Ypred)
> tableConfusion

   Ypred
Y    0  1
  0 61  4
  1 28  7

> # Sensibilité
> tableConfusion[2,2]/sum(tableConfusion[2,])

[1] 0,2

> # Spécificité
> tableConfusion[1,1]/sum(tableConfusion[1,])

[1] 0,9384615
```

```
> # Taux de bon de classement
> mean(Ypred == ronfle$RONFLE)

[1] 0,68

> # Taux de mauvais classement
> mean(Ypred != ronfle$RONFLE)

[1] 0,32
```

La sensibilité modèle est très faible tandis que la spécificité est très bonne. L'erreur de classement est égale à $0,32$, ce qui est assez décevant puisque le pourcentage de ronfleur parmi les patients est égal à $0,35$.

12. Quel est le risque d'une telle évaluation ? Quelle méthode préconiseriez-vous ici ?
    Le risque d'une telle évaluation est le biais d'optimisme, car les mêmes données servent à la fois à apprendre le modèle et à le tester. Ici on préconniserait la validation croisée leave-one-out.
13. Pour améliorer les capacités prédictives du modèle, ajuster le modèle avec toutes les variables. Toutes variables ont-elles un effet significatif ?

```
> logist2 = glm(RONFLE ~ ., family = "binomial", data = ronfle)
> summary(logist2)

Call:
glm(formula = RONFLE ~ ., family = "binomial", data = ronfle)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1,5911  -0,8516  -0,5317   1,0415   2,3542

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5,14959    5,96562  -0,863  0,38802
AGE          0,06213    0,02330   2,666  0,00767 **
POIDS       -0,01543    0,03319  -0,465  0,64195
TAILLE       0,01510    0,04754   0,318  0,75079
ALCOOL       0,23654    0,08611   2,747  0,00601 **
SEXE        -0,65218    0,67369  -0,968  0,33301
TABA        -1,20057    0,55798  -2,152  0,03143 *
---
Signif. codes:  0
```

Les variables AGE, ALCOOL et TABA ont un effet significatif sur la variable RONFLE.

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 ?

```
> logistbest = step(logist2)

Start:  AIC=123,42
RONFLE ~ AGE + POIDS + TAILLE + ALCOOL + SEXE + TABA

         Df Deviance    AIC
- TAILLE  1   109,52 121,52
- POIDS   1   109,63 121,63
- SEXE    1   110,39 122,39
<none>        109,42 123,42
- TABA    1   114,37 126,37
- AGE     1   117,52 129,52
- ALCOOL  1   117,82 129,82

Step:  AIC=121,52
RONFLE ~ AGE + POIDS + ALCOOL + SEXE + TABA

         Df Deviance    AIC
```

```
- POIDS   1   109,72 119,72
- SEXE    1   110,45 120,45
<none>        109,52 121,52
- TABA    1   114,42 124,42
- AGE     1   117,85 127,85
- ALCOOL  1   118,12 128,12

Step:  AIC=119,72
RONFLE ~ AGE + ALCOOL + SEXE + TABA

         Df Deviance    AIC
- SEXE    1   110,66 118,66
<none>        109,72 119,72
- TABA    1   114,52 122,52
- AGE     1   118,02 126,02
- ALCOOL  1   118,13 126,13

Step:  AIC=118,66
RONFLE ~ AGE + ALCOOL + TABA

         Df Deviance    AIC
<none>        110,66 118,66
- TABA    1   114,80 120,80
- AGE     1   119,76 125,76
- ALCOOL  1   122,86 128,86
```

```
> logistbest

Call:  glm(formula = RONFLE ~ AGE + ALCOOL + TABA, family = "binomial",
    data = ronfle)

Coefficients:
(Intercept)          AGE       ALCOOL         TABA
   -4,28973      0,06526      0,26204     -1,05540

Degrees of Freedom: 99 Total (i.e. Null);  96 Residual
Null Deviance:            129,5
Residual Deviance: 110,7         AIC: 118,7
```

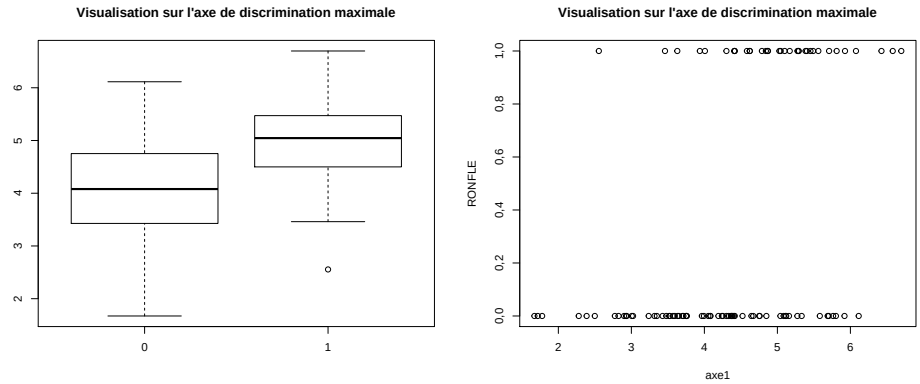
Les variables retenues sont les variables AGE, ALCOOL et TABA.

Le risque de ronfler quand la consommation d'alcool journalière augmente de 1 verre est mutipliée par 1,3.

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.

    Pour visualiser les données on pourrait réaliser une analyse factorielle discriminante :

```
> library("MASS")
> lda = lda(RONFLE ~ .,data = ronfle)
> axe1 = as.matrix(ronfle[,-1]) %*% lda$scaling
> par(mfcol = c(1,2))
> boxplot(axe1 ~ ronfle$RONFLE, main = "Visualisation sur l'axe de discrimination maximale")
> plot(axe1,ronfle$RONFLE,ylab = "RONFLE",main = "Visualisation sur l'axe de discrimination maximale")
```



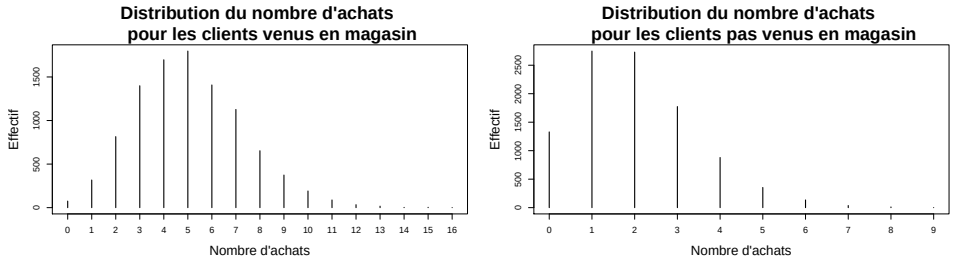
16. Proposer et mettre en œuvre une autre méthode de classification supervisée étudiée en cours.
On pourrait utiliser l'analyse discriminante linéaire :

```
> lda = lda(RONFLE ~ .,data = ronfle)
> lda

Call:
lda(RONFLE ~ ., data = ronfle)

Prior probabilities of groups:
   0    1
0,65 0,35

Group means:
       AGE    POIDS   TAILLE   ALCOOL      SEXE      TABA
0 50,26154 90,47692 180,9538 2,369231 0,3076923 0,6769231
1 56,00000 90,28571 181,3714 4,028571 0,1428571 0,5714286

Coefficients of linear discriminants:
               LD1
AGE     0,05973655
POIDS  -0,01620579
TAILLE  0,01590170
ALCOOL  0,24058822
SEXE   -0,55413371
TABA   -1,14621434
```

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 ?
   Ici c'est l'hypothèse de normalité qui n'est pas vérifiée car les données sont discrètes et les faibles effectifs ne permettent pas d'envisager l'approximation d'une loi discrète par une loi continue.
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 ?
   Ici cette modélisation est plus raisonnable car elle respecte la nature des données, par ailleurs on retrouve un distribution de probabilités par classe similaire à la loi de Poisson.
3. Donner l'expression mathématique de $P(Y = 1/X = x)$.

$$P(Y = 1/X = x) = \frac{P(X = x/Y = 1)P(Y = 1)}{P(X = x/Y = 1)P(Y = 1) + P(X = x/Y = 0)P(Y = 0)} = \frac{\pi_1 e^{-\lambda_1}\frac{\lambda_1^x}{x!}}{\pi_1 e^{-\lambda_1}\frac{\lambda_1^x}{x!} + \pi_0 e^{-\lambda_0}\frac{\lambda_0^x}{x!}}$$

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.

```
> clients = read.csv("clients.csv")
> effectifs = table(clients$y)
> pi1 = effectifs[2]/sum(effectifs)
> pi1

  1
0,5

> pi0 = effectifs[1]/sum(effectifs)
> pi0

  0
0,5

> # Moyenne des clients ayant effectué un achat
> lambda1 = mean(clients$x[clients$y == 1])
> lambda1

[1] 5,0469

> # Moyenne des clients n'ayant pas effectué d'achat
> lambda0 = mean(clients$x[clients$y == 0])
> lambda0
```

```
[1] 1,9994
```

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`.
   En appliquant la formule

```
> p = pi1 * dpois(3,lambda1) / (pi1 * dpois(3,lambda1) + pi0 * dpois(3,lambda0))
> p
```

```
        1
0,4329761
```

   La probabilité que le client vienne en magasin est de 0,433.

6. Évaluer les performances du modèle ainsi produit.
   Compte-tenu du grand nombre de données et du faible nombre de paramètres estimés le risque de sur-ajustement est très limité.
   On décide donc ici d'évaluer le modèle sur les données d'apprentissage.
   On commence par calculer les probabilité a posteriori pour l'ensemble des individus :

```
> p = pi1 * dpois(clients$x,lambda1) / (pi1 * dpois(clients$x,lambda1) + pi0 * dpois(clients$x,lambda0))
```

   Affectation des individus à la classe de plus forte probabilité puis calcul des indicateurs usuels :

```
> Ypred = (p > 0.5)*1
> # Table de confusion
> tableConfusion = table(Y = clients$y, Ypred)
> tableConfusion
```

```
   Ypred
Y      0    1
  0 8581 1419
  1 2606 7394
```

```
> # Sensibilité
> tableConfusion[2,2]/sum(tableConfusion[2,])
```

```
[1] 0,7394
```

```
> # Spécificité
> tableConfusion[1,1]/sum(tableConfusion[1,])
```

```
[1] 0,8581
```

```
> # Taux de bon de classement
> mean(Ypred == clients$y)
```

```
[1] 0,79875
```

```
> # Taux de mauvais classement
> mean(Ypred != clients$y)
```

```
[1] 0,20125
```

7. Lien avec un autre modèle déjà étudié :
   (a) Montrer que $\ln P(Y=1, X=x) = \ln \pi_1 - \lambda_1 - \ln(x!) + x \ln \lambda_1$

   $$\ln P(Y=1, X=x) = \ln P(X=x/Y=1)P(Y=1) = \ln \pi_1 + \ln P(X=x/Y=1) = \ln \pi_1 + \ln e^{-\lambda_1} + \ln \lambda_1^x - \ln(x!)$$

   On en déduit que

   $$\ln P(Y=1, X=x) = \ln \pi_1 - \lambda_1 - \ln(x!) + x \ln \lambda_1$$

   (b) En déduire que :

   $$\ln \frac{P(Y=1/X=x)}{P(Y=0/X=x)} = \ln \frac{\pi_1}{\pi_0} - \lambda_1 + \lambda_0 + x \ln \frac{\lambda_1}{\lambda_0}.$$

   $$\ln \frac{P(Y=1/X=x)}{P(Y=0/X=x)} = \ln \frac{P(Y=1, X=x)}{P(Y=0, X=x)} = \ln P(Y=1, X=x) - \ln P(Y=0, X=x)$$

   et finalement :

   $$\ln \frac{P(Y=1/X=x)}{P(Y=0/X=x)} = \ln \pi_1 - \lambda_1 - \ln(x!) + x \ln \lambda_1 - (\ln \pi_0 - \lambda_0 - \ln(x!) + x \ln \lambda_0) = \ln \frac{\pi_1}{\pi_0} - \lambda_1 + \lambda_0 + x \ln \frac{\lambda_1}{\lambda_0}.$$

(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 $P(Y = 1/X = x)$?
Cette méthode fait penser à la régression logistique avec $\beta_0 = \ln \frac{\pi_1}{\pi_0} - \lambda_1 + \lambda_0$ et $\beta_1 = \ln \frac{\lambda_1}{\lambda_0}$.
On peut donc estimer β_0 et β_1 en appliquant directement la régression logistique.

(d) Donner l'expression mathématique de $P(Y = 1/X = x)$ estimée à partir de cette seconde méthode appliquée au fichier clients.csv.

```
> logistclients = glm(y ~ x, data = clients, family = "binomial")
> logistclients

Call:  glm(formula = y ~ x, family = "binomial", data = clients)

Coefficients:
(Intercept)            x
     -3,020        0,918

Degrees of Freedom: 19999 Total (i.e. Null);  19998 Residual
Null Deviance:            27730
Residual Deviance: 17420         AIC: 17420
```

L'expression mathématique de $P(Y = 1/X = x)$ est ici :

$$P(Y = 1/X = x) = \frac{e^{-3,02+0,918x}}{1 + e^{-3,02+0,918x}}$$