Devoir surveillé de classification supervisée

Durée : 2h, trois feuilles recto-verso manuscrites autorisées, tout type de calculatrice autorisé

Nous disposons de données sur des patients atteints d'une hépatite ($n = 143$ patients, $d = 15$ variables). Dans cette étude nous nous intéresserons à prédire l'état des patients, variable `class` (`die` ou `live`), à partir des autres variables à disposition.

age	bilirubin	sgot	class	sex	steriod	antivirals	fatigue
Min. : 7.0	Min. :0.300	Min. : 14.00	die : 29	female : 15	no :69	no : 22	no :94
1st Qu. :32.0	1st Qu. :0.700	1st Qu. : 31.00	live :114	male :128	yes :74	yes :121	yes :49
Median :39.0	Median :1.000	Median : 58.00					
Mean :40.8	Mean :1.415	Mean : 82.85					
3rd Qu. :50.0	3rd Qu. :1.500	3rd Qu. :100.50					
Max. :78.0	Max. :8.000	Max. :648.00					

TABLE 1: Résumé des variables 1 à 8

malaise	anorexia	spleen_palpable	spiders	asites	varices	histology
no :57	no : 29	no : 28	no :48	no : 20	no : 18	no :76
yes :86	yes :114	yes :115	yes :95	yes :123	yes :125	yes :67

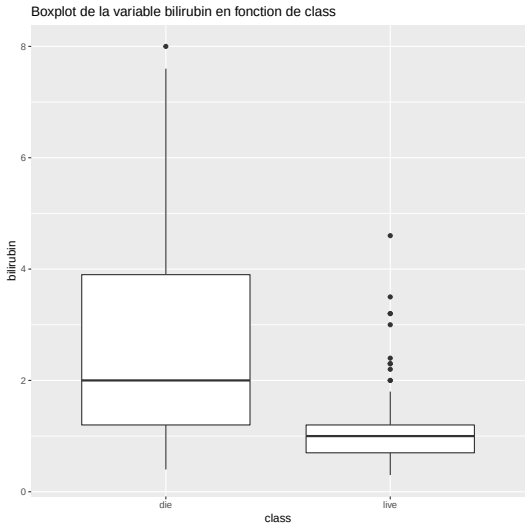
TABLE 2: Résumé des variables 9 à 15

Analyses préliminaires

1. Le jeu de données de l'étude résulte en fait d'un jeu de données plus grand duquel on a supprimé les individus avec des données manquantes. Pourquoi ne peut-on pas prendre en compte le individus avec des valeurs manquantes lors de l'ajustement d'une régression logistique ?

   On ne peut pas prendre en compte les individus avec les données manquantes dans la régression logistique puisque qu'il est impossible de calculer la probabilité d'appartenance à la classe pour ces individus (pas possible d'appliquer la formule de calcul).

On décide ici d'ajuster un modèle prédictif de la variable `class` en fonction de la variable `bilirubin`. Mais avant on réalise un graphique descriptif, ainsi qu'un test de l'ANOVA de l'effet de la variable `class` sur la variable `bilirubin` :




```
> summary(lm(bilirubin ~ class, data = d))

Call:
lm(formula = bilirubin ~ class, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1724 -0.4202 -0.1202  0.1798  5.4276

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5724     0.1963  13.105  < 2e-16 ***
classlive    -1.4522     0.2198  -6.606 7.49e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.057 on 141 degrees of freedom
Multiple R-squared:  0.2363,Adjusted R-squared:  0.2309
F-statistic: 43.64 on 1 and 141 DF,  p-value: 7.495e-10
```

2. Quel est la part de variance de la variable `bilirubin` expliquée par la variable `class` ?

   La part de variance de la variable `bilirubin` expliquée par la variable `class`, est donnée par le `Multiple R-squared`. Ainsi la part de variance expliquée est de 23,6%.

3. Ici quelle est l'hypothèse nulle testée par le test de l'ANOVA ?

   L'hypothèse nulle testée est $H_0 : \{E[bilirubin|class = die] = E[bilirubin|class = liver]\}$.

4. La variable `class` a-t-elle un effet significatif sur la variable `bilirubin` au niveau $\alpha = 0,05$?
   Oui la variable `class` a un effet significatif sur la variable `bilirubin` car la probabilité critique du test de l'ANOVA (p-value : 7.495e-10) est inférieure à $0,05$.
5. La probabilité critique pourrait cependant être erronée du fait que toutes les hypothèses de l'ANOVA ne semblent par réunies ici. Quelle est l'hypothèse de l'ANOVA qui semble sûrement violée ici ?
   L'ANOVA fait l'hypothèse d'homogénéité des variances, ce qui ne semble pas raisonnable ici puisque la variance de `bilirubin` semble différente selon les modalité de la variable `class`.
   On lance `var.test(bilirubin  class, data = d)`, et on obtient :
   ```
   F test to compare two ???
   F = 8.907, num df = 28, denom df = 113, p-value < 2.2e-16
   ```
6. Que compare le F test ? Qu'en conclure ?
   Le F test est un test de Fisher, il s'agit d'un test d'égalité des variances.

Ajustement d'un modèle de régression logistique

On décide maintenant d'ajuster la régression logistique de la variable `class` en fonction de la variable `bilirubin`. On lance `glm(formula = class  bilirubin, family = "binomial", data = d)` :

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.9953     0.4424   6.770 1.28e-11 ***
bilirubin    -1.0247     0.2337  -4.385 1.16e-05 ***

    Null deviance: 144.22  on 142  degrees of freedom
Residual deviance: 114.42  on 141  degrees of freedom
AIC: 118.42
```

7. A partir de l'ensemble des éléments à votre disposition, la modalité prédite pour la variable `class`, est-elle la modalité `dead` ou `live` ?
   Ici la modalité `die` apparait en première, elle est donc codée par 0 (modalité de référence), et la modalité `live` correspond à la modalité prédite de la variable `class` par le modèle.
   Par ailleurs cela est cohérent avec le signe du coefficient associé à la variable `bilirubin` qui est négatif, ce qui est cohérent avec les boîtes à moutâches puisque quand `bilirubin` augmente, la probabilité de la modalité `die` semble diminuer.
8. Quel est l'utilité du critère AIC ?
   Le critère AIC est un critère de choix de modèle. Il permet de choisir entre différents modèles. Le modèle retenu est celui qui minimise le critère AIC.
9. Quel est la probabilité de survie d'un patient avec une valeur de $0,75$ pour la `bilirubine` ?

$$P(class = live|bilirubin = 0,75) = \frac{\exp(2,9953 - 1,0247 \times 0,75)}{1 + \exp(2,9953 - 1,0247 \times 0,75)} = 0,90.$$

   Ainsi la probabilité de survie du patient est de 90%.
10. Comment peut-on mesurer les performances du modèle ajusté ?
    On peut mesurer les performances du modèle ajusté, en calculant un taux de de bon classement, la sensibilité, la spécificité ou en réalisant la courbe ROC.

On décide maintenant d'ajuster le modèle avec l'ensemble des variables explicatives, puis de réaliser une sélection de variables pas à pas.

Les coefficients du modèle final sont les suivants :

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   18.6848  1476.1356   0.013  0.98990
bilirubin     -1.0764     0.3136  -3.432  0.00060 ***
sexmale      -16.8908  1476.1352  -0.011  0.99087
malaiseyes     1.7556     0.7035   2.495  0.01258 *
anorexiayes   -3.0382     0.9825  -3.092  0.00199 **
spidersyes     1.7582     0.6540   2.688  0.00718 **
asitesyes      2.2301     0.7186   3.103  0.00191 **
```

```
---

    Null deviance: 144.217  on 142  degrees of freedom
Residual deviance:  73.379  on 136  degrees of freedom
AIC: 87.379
```

11. Quel critère la sélection pas à pas cherche-t-elle à minimiser ?
    Il cherche à minimiser le critère AIC.
12. Quel intérêt peut-il y avoir à effectuer une étape de sélection de variables ?
    Une sélection une étape de sélection de variables permet de limiter le nombre de variables présentes dans le modèle. Cela simplifie son interprétation d'une part, et d'autre part cela permet d'améliorer se performances en obtenant un modèle effectuant un meilleurs compromis biais variance.
13. Quel est la probabilité de survie pour le patient suivant :

    ```
       bilirubin  sex malaise anorexia spiders asites
    16         2 male      no       no     yes     no
    ```

    On calcule $s = 18,6848 - 1,0764 \times 2 - 16,8908 + 1,7582 = 1,3994$. On en déduit la probabilité de survie en calculant $\exp(s)/(1+\exp(s)) = 80,2\%$.
14. Commenter la ligne correspondant au coefficient `sexmale` ?
    Le coefficient correspondant à `sexmale` à une très grande valeurs en valeurs absolue, mais aussi une forte variance d'estimation, et une probabilité critique du test de nullité du coefficient qui ne ressort pas significative.

En effectuant le croisement `table(d$sex, d$class)`, on obtient :

```
       die live
female   0   15
male    29   99
```

15. Cela explique-t-il le résultat obtenu ?
    En fait aucune femme n'est décédée dans l'étude, ainsi on devrait avoir $P(live|female)$ estimée à 1.

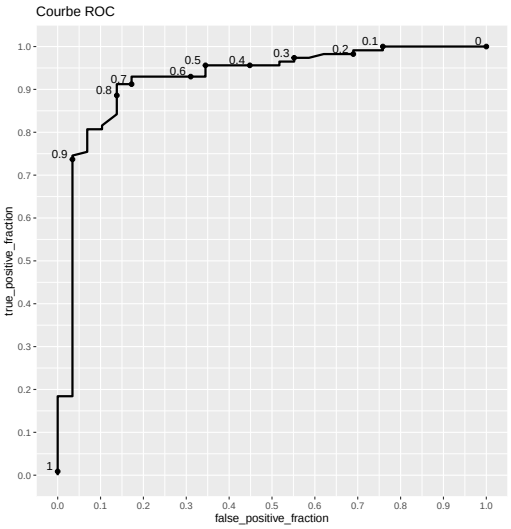
Soit β_0 et β_{male} les coefficients du modèle de la régression logistique de la variable `class` en fonction de la variable `sex`.

16. Exprimer $P(\text{class} = \text{live}|\text{sex} = \text{female})$ en fonction de β_0 et/ou β_{male}.

$$P(\text{class} = \text{live}|\text{sex} = \text{female}) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$$

17. Pour coller aux données, quelle devrait être la valeur β_0 ? Que doit-il se passer dans le logiciel à votre avis ?
    Pour coller les aux données la valeur β_0 devrait être très grande, théoriquement $+\infty$. En pratique le logiciel lui affecte une valeur assez grande 18.6848, ce qui explique la compensation pour le coefficient sexmale ici -16.8908.

On trace la courbe ROC associée au modèle final, les points affichés représentent des valeurs particulières pour le seuil sur la probabilité.



18. Quelle quantité est représentée sur l'axe de x ? Sur l'axe des y ?

    Sur l'axe des x, on a le taux de faux positif (1-S_p). Sur l'axe des y, on a le taux de vrais positif (S_e).

19. A combien faut-il fixer le seuil si on veut retrouver 95% des patients survivants ? Quel est le taux de faux positifs qui en découle ?

    Il faut fixer le seuil à 0,5 si on veut retrouver 95% des patients survivants. Le taux de faux positifs qui en résulte est de 35%.

20. Approximativement quelle valeur optimale en terme de compromis sensibilité/spécificité suggère le graphique ? Quelles sont dans ce cas les valeurs de la sensibilité et de la spécificité qui en découlent ?

    La valeur optimale du seuil est pour laquelle la courbre est la plus proche du coin supérieur gauche, c'est à dire choisir 0,7. La sensibilité qui en résulte est d'environ 90%. La spécificité qui en résulte est d'environ 82,5% par le calcul (1 - 17,5%).

Questions sur l'analyse discriminante probabiliste

18. Rappeler le principe de l'analyse discriminante probabiliste.

    L'analyse discriminante probabiliste consiste à modèliser la distribution de $X|Y$, la distribution de Y et d'en déduire la distribution de $Y|X$.

19. Quelle est la différence entre analyse discriminante linéaire (LDA) et quadratique (QDA) ?

    Dans la LDA on suppose l'égalité des matrices de variance covariance sachant la classe, ce que l'on ne suppose pas dans le cadre de la QDA.

20. Est-il possible ici d'ajuster une LDA ou QDA à partir de l'ensemble des variables à votre disposition ? Pour quelle autre solution d'analyse discriminante probabiliste pourriez-vous opter ?

    Non ce n'est pas possible puisqu'on dispose aussi de variables qualitatives. Dans ce cas on pourrait utiliser le classifieur de Bayes naïf qui fait l'hypothèse d'indépendance des variables sachant la classe, ce qui permet d'éviter de modèliser les corrélations entre les variables explicative de différentes nature à l'intérieur du modèle de $X|Y$.