apGIS 4 : Corrigé du devoir surveillé de classification supervisée

Vincent Vandewalle

30/03/2018

Décription générale des données

L'objectif de ce devoir est d'étudier les facteurs qui influencent la mortalité d'un accident.

Les données sont issues de l'observatoire national interministériel de sécurité routière (extraction du 3 juillet 2012). Il s'agit de la base de données des accidents sur 6 années avec informations de géolocalisation. Le descriptif détaillé de ces données est présent dans le fichier *descriptif.pdf*.

Les données sont issues d'une version simplifiée du fichier ETALAB ACCIDENTS. Ce fichier simplifié est constitué initialement de 454.372 lignes correspondant à 454.372 accidents. Parmis ces accidents les effectifs des différentes valeurs possibles de la variable *ttue* (qui représente le nombre de morts) sont les suivants :

Valeur	0	1	2	3	4	5	7	26
Effectif	429511	23142	1465	188	48	12	5	1

Ici il y a beaucoup plus d'accidents non-mortels que d'accidents mortels (heureusement ...).

1. Réflexion préliminaire

Question 1 : Les méthodes étudiées en classification supervisées vous permettent-elles de prédire le nombre de morts (justifier) ? Pourquoi une régression linéaire classique ne serait-elle pas non plus très adaptée (justifier) ? (1,5 points)

Réponse 1 : Ici la variable nombre de tués est une variable qui prend les valeurs 0, 1, ..., 26. Il s'agit donc d'une variable à valeurs entières, ce qui n'est pas adapté au cadre de la classification supervisée, sauf si on considère chaque nombre possible comme une modalité, mais alors pour certaines modalités on n'aurait pas d'effectif assez grand pour bien conduire l'esimation. Dans ce cas une solution pourrait être de considérer plutôt un modèle parcimonieux de régression logistique sur variables ordinales. La régression linéaire classique n'est pas non plus adaptée ; la variable à expliquer est bien numérique, mais l'hypothèse de normalité des résidus ne peut pas être envisagée du fait du faible nombre de valeurs différentes de la variable numérique. En pratique elle semble cependant l'approche la plus adaptée pour prédire le nombre d'accidents. En fait l'approche la plus adaptée ici serait de réaliser une régression de Poisson :

$$Y \sim \mathcal{P}(\lambda(x))$$

où le paramètre d'intensité de la loi de Poisson, $\lambda(x)$, dépendrait des variables explicatives. L'hypothèse faite dans le cadre du modèle linéaire généralisé est

$$\log(\lambda(x)) = \beta_0 + \sum_{j=1}^{d} \beta_j x_j$$

Il peut être ajusté à l'aide de la fonction *glm* en précisant l'option *family = "poisson"*.

Pour simplifier le problème on décide de créer une nouvelle variable nommée *mort* qui prend la valeur 1 si l'accident est mortel et 0 sinon.

Question 2 : En l'état actuel la proportion d'accidents non mortels est très élevée. Quel serait le taux de bon classement obtenu si on classait tous les individus dans la classe majoritaire ? Quelle serait alors la sensibilité et la spécifité de la règle de classement ? (1,5 points)

Réponse 2 :

```
n0 = 429511  # Effectif accidents non mortels
n1 = 23142 + 1465 +  188 + 48 +  12 + 5 +1  # Effectif accident mortels
p0 = n0/(n1+n0)
p0 # proportion classe majoritaire
```

```
## [1] 0.9452849
```

```
TBC = p0
TBC # simplement la proportion de la classe majoritaire
```

```
## [1] 0.9452849
```

```
Se = 0
Se # On ne prédit que des 0, donc aucun 1 bien prédit
```

```
## [1] 0
```

```
Sp = 1
Sp # On ne prédit que des 0, donc tous les 0 bien prédits
```

```
## [1] 1
```

Enfin pour limiter la quantité de données à traiter, on tire au hasard 24.861 lignes correspondant à des accidents non mortels, et on conserve toutes les lignes correspondant à des accidents mortels. Il s'agit en fait d'une méthode d'échantillonnage appelée échantillonnage retrospectif. Son impact sur l'estimation des paramètres du modèle de régression logistique fera l'objet d'une question ultérieure.

Vous pouvez maintenant charger le fichier *accidents.Rda* resultant de ces prétraitements.

```
load("accidents.Rda")
```

Question 3 : Maintenant que vous disposez des données vous pouvez avancer davantage sur l'analyse. A l'aide d'un simple résumé des données et d'une lecture du descriptif des variables, lister les variables qui vous semblent pertinentes pour l'analyse. Donner aussi la liste des variables qu'il faut aboslument exclure de l'analyse. (2 pts)

Réponse 3 : On commence par faire le résumé des données

```
summary(accidents)
```

```
##                              org
##  Gendarmerie                   :24330
##  Préfecture de Police de Paris: 3023
##  C.R.S.                        : 2039
##  P.A.F                         :   16
##  Sécurité publique             :20314
##
##
##                                    lum
##  Plein jour                          :31250
##  Crépuscule ou aube                  : 3146
##  Nuit sans éclairage public          : 7816
```

```
##  Nuit avec éclairage public non allumé:  457
##  Nuit avec éclairage public allumé    : 7053
##
##
##                                                        agg
##  Hors agglomération                                      :24550
##  Agglomération de plus de 300 000 habitants              : 5179
##  Agglomération entre 20 000 habitants et 50 000 habitants  : 4953
##  Agglomération entre 10 000 habitants et 20 000 habitants  : 4371
##  Agglomération entre 100 000 habitants et 300 000 habitants: 3251
##  Agglomération entre 50 000 habitants et 100 000 habitants : 3061
##  (Other)                                                 : 4357
##                 int                         atm
##  Hors intersection :39315   Normale          :39649
##  Intersection en X : 4502   Pluie légère     : 4614
##  Intersection en T : 2957   Temps couvert    : 2010
##  Giratoire         :  977   Pluie forte      : 1165
##  Autre intersection:  848   Temps éblouissant:  589
##  (Other)           : 1122   (Other)          : 1693
##  NA's              :    1   NA's             :    2
##                             col             com             dep
##  Autre collision              :17643   Min.   :  1.0   Min.   : 10.0
##  Deux véhicules - par le coté  :11243   1st Qu.: 63.0   1st Qu.:300.0
##  Deux véhicules - frontale     : 7431   Median :146.0   Median :570.0
##  Sans collision                : 6055   Mean   :201.5   Mean   :533.2
##  Deux véhicules - par l'arrière: 4101   3rd Qu.:299.0   3rd Qu.:760.0
##  (Other)                       : 3248   Max.   :987.0   Max.   :974.0
##  NA's                          :    1
##                    catr
##  Route Départementale:21982
##  Voie Communale      :17209
##  Route Nationale     : 4466
##  Autoroute           : 3152
##  autre               :  812
##  (Other)             :  332
##  NA's                : 1769
##                                      infra            voie
##  Hors infrastructure spécifique         :42750   Min.   :    0.0
##  Carrefour aménagé                      : 2855   1st Qu.:    0.0
##  Pont - autopont                        :  830   Median :    7.0
##  Bretelle d'échangeur ou de raccordement:  688   Mean   :  211.3
##  Souterrain - tunnel                    :  288   3rd Qu.:  123.0
##  (Other)                                :  523   Max.   :38300.0
##  NA's                                   : 1788   NA's   :3123
##        v1              v2                                          circ
##  Min.   :0.000          :47347   A sens unique                    : 5772
##  1st Qu.:0.000   A      : 1428   Bidirectionnelle                 :34306
##  Median :0.000   D      :  196   A chaussées séparées             : 5616
##  Mean   :0.095   E      :  172   Avec voies d'affectation variable:  227
##  3rd Qu.:0.000   B      :  101   NA's                             : 3801
##  Max.   :9.000   N      :   91
##  NA's   :24102   (Other):  387
##       nbv               pr             pr1
##  Min.   : 0.000          :17453   Min.   :   0.0
```

```
##  1st Qu.: 2.000   0000   : 6245   1st Qu.:   0.0
##  Median : 2.000   0      : 3891   Median : 174.5
##  Mean   : 1.943   0001   :  971   Mean   : 298.4
##  3rd Qu.: 2.000   1      :  590   3rd Qu.: 544.0
##  Max.   :90.000   0002   :  571   Max.   :8000.0
##  NA's   :1825     (Other):20001   NA's   :17492
##                   vosp                    prof
##  Hors voies spéciale:45380   Plat          :34136
##  Piste cyclable     :  761   Pente         : 7505
##  Bande cyclable     :  593   Sommet de côte: 1195
##  Voie réservée      : 1173   Bas de côte   :  974
##  NA's               : 1815   NA's          : 5912
##
##
##                 plan                                 situ
##  Partie rectiligne :34187   Sur chaussée               :38257
##  En courbe à gauche: 5363   Sur bande d'arrêt d'urgence:  340
##  En courbe à droite: 5201   Sur accotement             : 5937
##  En « S »          :  870   Sur trottoir               :  996
##  NA's              : 4101   Sur piste cyclable         :  200
##                             NA's                       : 3992
##
##       ttue             tbg               tbl              tindm
##  Min.   : 0.000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.500   Median : 0.0000   Median : 0.0000   Median : 1.0000
##  Mean   : 0.542   Mean   : 0.4305   Mean   : 0.5351   Mean   : 0.8193
##  3rd Qu.: 1.000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000
##  Max.   :26.000   Max.   :34.0000   Max.   :53.0000   Max.   :52.0000
##
##                typenumero        numero           distancemetre
##  Numéro non renseigné: 6128   Min.   : 0.000e+00   Min.   :-159.000
##  Adresse postale     :15125   1st Qu.: 0.000e+00   1st Qu.:   0.000
##  Candélabre          : 2158   Median : 9.000e+00   Median :   0.000
##  Autre               : 1459   Mean   :3.555e+233   Mean   :   0.866
##  NA's                :24852   3rd Qu.: 6.400e+01   3rd Qu.:   0.000
##                               Max.   :5.410e+237   Max.   : 900.000
##                               NA's   :34506        NA's   :25048
##         libellevoie      coderivoli         grav                gps
##               :24842          :49708   Min.   :   0.43   Métropole:48735
##  AUTOROUTE A1 :  121   0000   :    1   1st Qu.:   0.86   Antilles :  513
##  AUTOROUTE A6 :   79   0067   :    1   Median :  88.02   Guyane   :  167
##  AUTOROUTE A86:   73   0130   :    1   Mean   :  59.08   Réunion  :  307
##  A13          :   71   0350   :    1   3rd Qu.: 100.00
##  A4           :   71   0390   :    1   Max.   :2859.20
##  (Other)      :24465   (Other):    9
##       lat                 long                                adr
##  Min.   :0.000e+00   Min.   :-4.760e+05                        :48577
##  1st Qu.:0.000e+00   1st Qu.: 0.000e+00   AUTOROUTE A1         :   24
##  Median :4.350e+06   Median : 0.000e+00   ROUTE DEPARTEMENTALE 3 :   12
##  Mean   :6.937e+75   Mean   : 1.231e+08   MADELEINE (ROUTE DE LA):   11
##  3rd Qu.:4.752e+06   3rd Qu.: 2.680e+05   BADUEL (ROUTE DE)      :   10
##  Max.   :1.930e+80   Max.   : 6.112e+11   ROUTE DEPARTEMENTALE 14:    8
##  NA's   :21900       NA's   :21906        (Other)                : 1080
```

```
##      numac              mort
##  Min.   :    15   Min.   :0.0
##  1st Qu.:214206   1st Qu.:0.0
##  Median :429495   Median :0.5
##  Mean   :327769   Mean   :0.5
##  3rd Qu.:441935   3rd Qu.:1.0
##  Max.   :454372   Max.   :1.0
##
```

Les variables à ne surtout pas utiliser sont les variables dont le calcul fait intervenir la variable *ttue* qui sert à définir la variable *mort* qui est justement celle que l'on cherche à prédire ici ... Ainsi il faut absolument exclure les variables *ttue* et *grav*.

Les variables reliées au nombres de blessés (*tbg*, *tbl* et *tindm*) ne sont pas formellement à exlucre, mais cependant les inclure dans le modèle serait un peu hors sujet si le but principal est de trouver les diverses conditions de la voirie, de la météo, qui ont une influence sur la mortalité de l'accident.

Toutes les autres variables peuvent être intéressantes, cependant on veillera par exemple à exclure les variables avec trop de valeurs manquantes, où alors avec trop de modalités différentes (par exemple adresses). Après, inclure toutes les autres variables ne serait potentiellement pas pertinent dans un but explicatif, et il vaut mieux réfléchir à un focus particulier dans l'analyse. On peut par exemple inclure dans le modèle les variable *lum*, *agg*, *int*, *atm*, *col*, *catr* (mais attention au grand nombre de valeurs manquantes).

Question 4 : Enfin pour conclure sur les questions d'ordre général. Les données disponibles vous suffisent-elles à prédire les zones les plus dangereuses ? Pourquoi ? (1 point)

Réponse 4 : Non elles ne sont pas suffisantes puisque les données ne contiennent ici que des informations sur les accidents corporels. Il faudrait aussi des données globales de circulation pour prédire la dangerosité des différentes routes.

2. Etudes bivariées préliminaires

Question 5 : Y-a-t'il une liaison significative entre la variable *lum* et la variable *mort* ? Quelle est la condition de luminosité où les accidents corporels sont les plus mortels ? (2 points)

Réponse 5 : Ici on va faire un tableau de contingence croisant les variables *lum* et *mort* puis effectuer un test du chi-2 d'indépendance

```
tab <- xtabs(~mort + lum,data = accidents)
tab # tableau de contingence
```

```
##     lum
## mort Plein jour Crépuscule ou aube Nuit sans éclairage public
##    0      17358               1425                       1663
##    1      13892               1721                       6153
##     lum
## mort Nuit avec éclairage public non allumé
##    0                                   211
##    1                                   246
##     lum
## mort Nuit avec éclairage public allumé
##    0                              4204
##    1                              2849
```

```
chisq.test(tab) # p-value < 2.2e-16
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 3254.6, df = 4, p-value < 2.2e-16
```

On rejette donc l'hypothèse d'indépendence entre les variables *lum* et *mort*. La lumière a un impact sur la mortalité de l'accident.

Pour trouver les conditions météorologiques où les accidents sont les plus mortels on calcule simplement un tableau des profils colonne à partir de *tab* :

```
prop.table(tab,2)
```

```
##     lum
## mort Plein jour Crépuscule ou aube Nuit sans éclairage public
##    0  0.5554560          0.4529561                   0.2127687
##    1  0.4445440          0.5470439                   0.7872313
##     lum
## mort Nuit avec éclairage public non allumé
##    0                             0.4617068
##    1                             0.5382932
##     lum
## mort Nuit avec éclairage public allumé
##    0                          0.5960584
##    1                          0.4039416
```

On voit que les conditions météorologiques où les accidents sont le plus mortels sont la nuit sans éclairage public, avec une proportion de mort de 78% dans cette condition.

Question 6 : Parmis les variables explicatives *lum*, *agg*, *int*, *atm*, *col*, quelle est la plus corrélée à la variable *mort* ? (2 points)

Réponse 6 : Ici on peut par exemple ici calculer le V de Cramer entre *mort* et toutes les autres variables puis les ordonner de celle qui a le V de Cramer le plus grand à celle qui a le V de Cramer le plus petit.

```
var = c("lum","agg","int","atm","col")
Vcramer = rep(NA,length(var))
names(Vcramer) = var
for (j in var){
  tab <- table(accidents$mort,accidents[,j])
  Vcramer[j] <- chisq.test(tab)$statistic/(sum(tab)*prod(dim(tab)-1))
}
Vcramer
```

```
##         lum         agg         int         atm         col
## 0.016364019 0.027733256 0.006531098 0.001212541 0.013377023
```

Ainsi la variable la plus corrélée à la variable *mort* est la variable *agg* (localisation par rapport à l'agglomération), avec un V de Cramer de $2,7\%$.

On peut analyser davantage le croisement entre les variables *mort* et *agg*

```
profils = prop.table(with(accidents,table(mort,agg)),2)
profils[,order(profils[2,],decreasing = T)]
```

```
##     agg
## mort Hors agglomération
##    0          0.2914868
##    1          0.7085132
##     agg
```

```
## mort Agglomération entre 2 000 habitants et 5 000 habitants
##    0                                                0.3755514
##    1                                                0.6244486
##     agg
## mort Agglomération de moins de 2 000 habitants
##    0                                  0.3888314
##    1                                  0.6111686
##     agg
## mort Agglomération entre 5 000 habitants et 10 000 habitants
##    0                                                 0.4431503
##    1                                                 0.5568497
##     agg
## mort Agglomération entre 10 000 habitants et 20 000 habitants
##    0                                                  0.6064974
##    1                                                  0.3935026
##     agg
## mort Agglomération entre 20 000 habitants et 50 000 habitants
##    0                                                 0.7522714
##    1                                                 0.2477286
##     agg
## mort Agglomération entre 50 000 habitants et 100 000 habitants
##    0                                                  0.7883045
##    1                                                  0.2116955
##     agg
## mort Agglomération entre 100 000 habitants et 300 000 habitants
##    0                                                   0.8012919
##    1                                                   0.1987081
##     agg
## mort Agglomération de plus de 300 000 habitants
##    0                                   0.8764240
##    1                                   0.1235760
```

Ainsi les accidents les plus mortels sont les accidents hors agglomération.

3. Ajustement du modèle de régression logistique

Question 7 : Ajuster un modèle de régression logistique permettant de prédire la variable *mort* en fonction de *lum*. Quelle est la modalité de la variable *lum* qui sert de modalité de référence ? Par combien est multiplié le risque d'accident mortel quand la lumière passe de "plein jour" à "crépuscule ou aube". (2 points)

Réponse 7 : On ajuste le modèle et on fait le résumé

```
model1 <- glm(mort ~ lum, family = "binomial", data = accidents)
summary(model1)
```

```
##
## Call:
## glm(formula = mort ~ lum, family = "binomial", data = accidents)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7593  -1.0844  -0.1628   1.2733   1.3465
##
## Coefficients:
```

```
##                                       Estimate Std. Error z value
## (Intercept)                           -0.22274    0.01138 -19.566
## lumCrépuscule ou aube                  0.41147    0.03758  10.949
## lumNuit sans éclairage public          1.53106    0.02989  51.222
## lumNuit avec éclairage public non allumé  0.37621    0.09452   3.980
## lumNuit avec éclairage public allumé     -0.16633    0.02680  -6.205
##                                       Pr(>|z|)
## (Intercept)                            < 2e-16 ***
## lumCrépuscule ou aube                  < 2e-16 ***
## lumNuit sans éclairage public          < 2e-16 ***
## lumNuit avec éclairage public non allumé 6.88e-05 ***
## lumNuit avec éclairage public allumé     5.46e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68929  on 49721  degrees of freedom
## Residual deviance: 65507  on 49717  degrees of freedom
## AIC: 65517
##
## Number of Fisher Scoring iterations: 4
```

La modalité que sert de référence est la première modalité de la variable *lum*

```
levels(accidents$lum)
```

```
## [1] "Plein jour"
## [2] "Crépuscule ou aube"
## [3] "Nuit sans éclairage public"
## [4] "Nuit avec éclairage public non allumé"
## [5] "Nuit avec éclairage public allumé"
```

Il s'agit donc de la modalité "Plein jour", donc tous les coefficients du modèle doivent être calculés relativement à cette modalité.

Donc pour obtenir l'odds-ratio de "Crépuscule ou aube" vs "Plien jour", il suffit de prendre l'exponentielle du coefficient associé à cette modalité

```
exp(coef(model1)["lumCrépuscule ou aube"])
```

```
## lumCrépuscule ou aube
##              1.509041
```

Question 8 : Ajuster un modèle de régression logistique permettant de prédire la *mort* en fonction de *lum*, *agg*, *int*, *atm* et *col*. Commenter le résultat. (1 point)

Réponse 8 :

```
model2 <- glm(mort ~ lum + agg + int + atm + col, family = "binomial",
              data = accidents)
# summary(model2)
```

Ici on remarque que presque toutes les modalités ont un effet significatif par rapport à la modalité de référence. Si on veut tester l'effet d'une variable dans son intégralité, alors il faut alors faire une test de modèles emboîtés. Par exemple si on veut tester l'effet de la variable *int* (type d'intersection), alors on fait :

```
model2bis <- glm(mort ~ lum + agg + atm + col, family = "binomial",
              data = accidents)
```

```r
anova(model2bis,model2, test = "LRT")
```

```
## Error in anova.glmlist(c(list(object), dotargs), dispersion = dispersion, : models were not all fitte
```

Ici, celà ne fonctionne pas car la variable *int* comporte des valeurs manquantes, du coup quand on ajuste le modèle 2 bis les cas avec les valeurs qui ne sont manquantes que pour la variable *int* sont réintégrés au modèle, et par conséquent les tests deviennent inapropriés ... Une solution serait de se limiter ici au ligne complètes pour les variables *mort*, *lum*, *agg*, *int*, *atm* et *col* (lignes qui ont étés utilisées pour le modèle 2). On procède donc comme suit :

```r
idx = apply(!is.na(accidents[,c("mort","lum","agg","int","atm","col")]),1,all)
# idx est vecteur de booléens qui prend la valeur vraie si toutes variables sont
# observées et 0 sinon
model2bis <- glm(mort ~ lum + agg + atm + col, family = "binomial",
              data = accidents[idx,])
anova(model2bis,model2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: mort ~ lum + agg + atm + col
## Model 2: mort ~ lum + agg + int + atm + col
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     49691      53984
## 2     49683      53824  8   160.02 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On en déduit donc ici que l'effet de la variable *int* est significatif.

Question 9 : Faire une sélection de variables pas à pas sur le modèle précédent. Qu'en concluez-vous ? (1 point)

Réponse 9 :

```r
step(model2,trace = FALSE)
```

```
##
## Call:  glm(formula = mort ~ lum + agg + int + atm + col, family = "binomial",
##     data = accidents)
##
## Coefficients:
##                                                    (Intercept)
##                                                        1.13159
##                                          lumCrépuscule ou aube
##                                                        0.23723
##                                       lumNuit sans éclairage public
##                                                        0.71363
##                          lumNuit avec éclairage public non allumé
##                                                        0.36429
##                              lumNuit avec éclairage public allumé
##                                                        0.44413
##                     aggAgglomération de moins de 2 000 habitants
##                                                       -0.50852
##        aggAgglomération entre 2 000 habitants et 5 000 habitants
##                                                       -0.45787
##       aggAgglomération entre 5 000 habitants et 10 000 habitants
##                                                       -0.66337
```

```
##   aggAgglomération entre 10 000 habitants et 20 000 habitants
##                                                       -1.22512
##   aggAgglomération entre 20 000 habitants et 50 000 habitants
##                                                       -1.83378
##  aggAgglomération entre 50 000 habitants et 100 000 habitants
##                                                       -2.03399
## aggAgglomération entre 100 000 habitants et 300 000 habitants
##                                                       -2.11665
##                 aggAgglomération de plus de 300 000 habitants
##                                                       -2.67876
##                                             intIntersection en X
##                                                       -0.14949
##                                             intIntersection en T
##                                                       -0.30649
##                                             intIntersection en Y
##                                                       -0.21050
##                          intIntersection à plus de 4 branches
##                                                       -0.49159
##                                                  intGiratoire
##                                                       -0.61484
##                                                     intPlace
##                                                       -0.31893
##                                               intPassage à niveau
##                                                        1.58790
##                                            intAutre intersection
##                                                        0.07357
##                                                 atmPluie légère
##                                                       -0.31253
##                                                  atmPluie forte
##                                                       -0.18819
##                                                atmNeige - grêle
##                                                       -0.32024
##                                             atmBrouillard - fumée
##                                                        0.37159
##                                            atmVent fort - tempête
##                                                        0.39962
##                                              atmTemps éblouissant
##                                                        0.62853
##                                                  atmTemps couvert
##                                                        0.28605
##                                                       atmAutre
##                                                        0.07599
##                             colDeux véhicules - par l'arrière
##                                                       -1.36767
##                               colDeux véhicules - par le coté
##                                                       -0.88244
##                         colTrois véhicules et plus - en chaîne
##                                                       -1.66308
##             colTrois véhicules et plus - collisions multiples
##                                                       -0.12951
##                                               colAutre collision
##                                                       -0.06159
##                                                colSans collision
##                                                       -0.53284
```

```
##
## Degrees of Freedom: 49717 Total (i.e. Null);  49683 Residual
##   (4 observations deleted due to missingness)
## Null Deviance:       68920
## Residual Deviance: 53820     AIC: 53890
```

Ici la sélection pas à pas ne conduit pas à supprimer des variables. On en déduit donc qu'on a intérêt à conserver toutes les variables dans le modèle.

Question 10 : Evaluer les performances prédictives du modèle précédent. (2 points)

Réponse 10 : On va réaliser la courbe ROC, caculer l'AUC, trouver le meilleur seuil et retourner les S_e, S_p, et TBC correspondants :

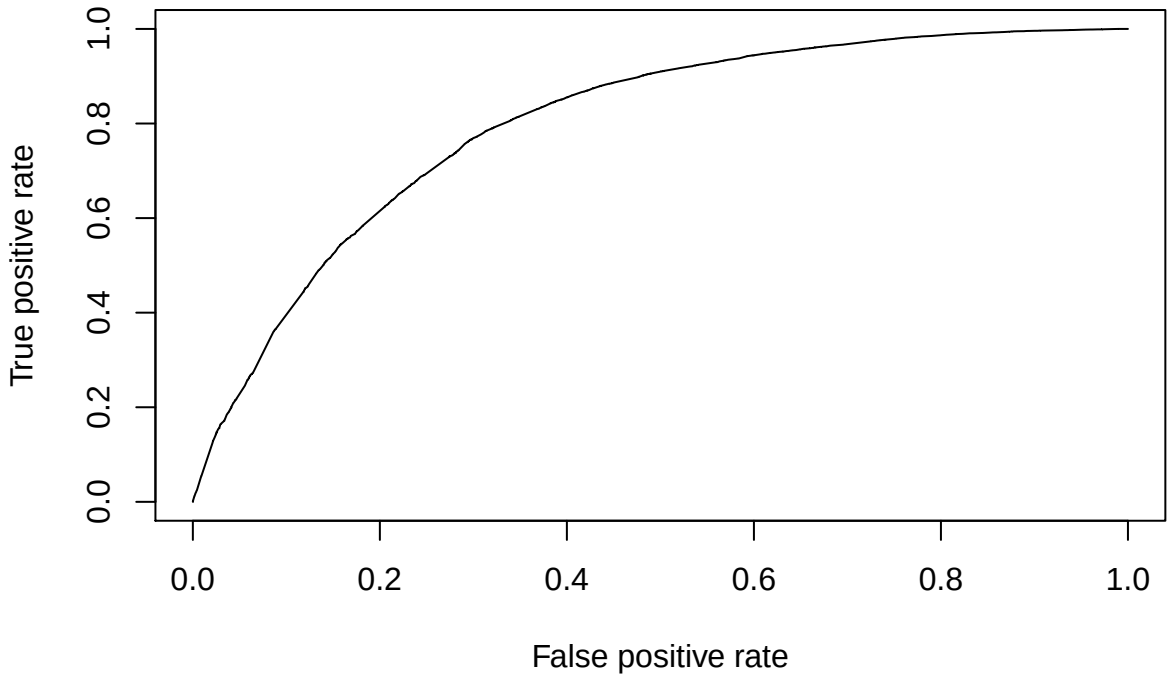
```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(caTools)
# On se limite directement aux ligne complètes pour évacuer le pb des manquants
model2 <- glm(mort ~ lum + agg + int + atm + col, family = "binomial",
              data = accidents[idx,])
S <- predict(model2, type="response")
pred <- prediction(S, accidents[idx,"mort"])
#Coordonnées de courbe ROC
perf <- performance(pred, "tpr", "fpr")
# Plot courbe ROC
plot(perf)
```



```r
# Calcul de l'AUC
colAUC(S, accidents[idx,"mort"])
```

```
##              [,1]
## 0 vs. 1 0.7985582
```

```r
id_best <- which.min((perf@x.values[[1]])^2 + (1 - perf@y.values[[1]])^2)
alpha_best = perf@alpha.values[[1]][id_best]
# Le seuil optimal est proche de 0,5 ce qui colle avec le seuil optimal au niveau de l'erreur de Bayes.
Se = perf@y.values[[1]][id_best]
Se
```

```
## [1] 0.7653566
```

```r
Sp = 1 - perf@x.values[[1]][id_best]
Sp
```

```
## [1] 0.7036084
```

```r
# Enfin, le TBC peut facilement est déduite de Se et Sp en faisant :
# TBC = Se * P(Y = 1) + Sp * P(Y = 0)
# Or dans l'échantillon à disposition P(Y = 1) = P(Y = 0) = 0,5
TBC = 0.5*Se + 0.5*Sp
TBC
```

```
## [1] 0.7344825
```

Ici en toute rigueur, on aurait dû partitionner les données en un échantillon d'apprentissage et un échantillon test pour évaluer correctement les performances du modèle.

Question 11 : Maintenant à vous de jouer, et de proposer le modèle de régression logistique le plus pertinent possible, commenter les variables retenues au final, et évaluer les performances de ce modèle. (4 points)

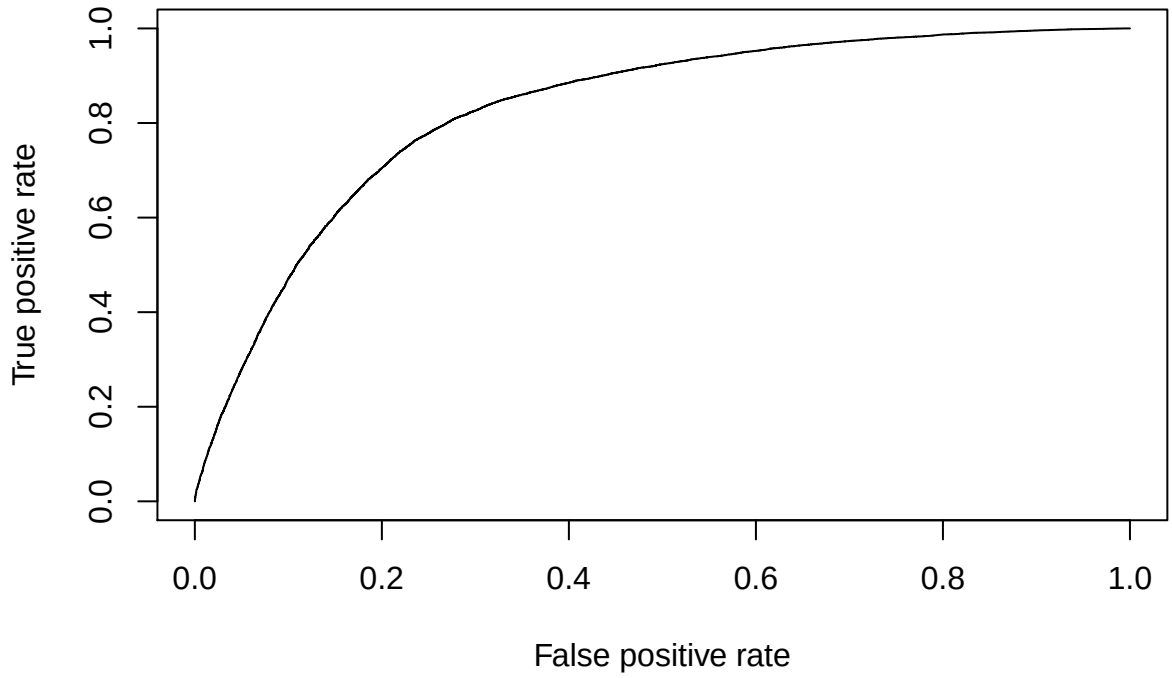
Réponse 11 : Ici on va essayer d'inclure d'autres variables dans le modèle. Cependant cette question n'est pas aisée si on veut réaliser des selections de variables pas à pas futures, puisqu'il faudra veuiller à toujours travailler sur le même ensemble d'indivdus, et que l'on ne veut pas se priver de trop de données dans l'analyse. On se propose donc de rajouter les variable *col* et *dep* au modèle précédent. La variable *dep* nécessaitant un recodage préalbale sous forme de facteur.

```r
table(accidents$dep,useNA = "always")
```

```
##
##   10   20   30   40   50   60   70   80   90  100  110  120  130  140  150
##  449  387  299  161  112 1283  243  164  108  229  394  201 2292  371   93
##  160  170  180  190  201  202  210  220  230  240  250  260  270  280  290
##  259  715  300  236  186  262  394  406   90  331  388  408  447  408  520
##  300  310  320  330  340  350  360  370  380  390  400  410  420  430  440
##  734  860  197 1149 1046  650  231  409  736  206  313  352  514  191  818
##  450  460  470  480  490  500  510  520  530  540  550  560  570  580  590
##  478  155  310   76  589  401  374  190  184  510  144  521  585  195 1329
##  600  610  620  630  640  650  660  670  680  690  700  710  720  730  740
##  562  215  770  511  560  200  262  655  391 1100  196  463  414  233  472
##  750  760  770  780  790  800  810  820  830  840  850  860  870  880  890
## 3032  714  853  738  240  537  310  238  844  464  435  292  337  285  347
##  900  910  920  930  940  950  971  972  973  974 <NA>
##  114  733 1048 1193 1065  547  487  404  311  567    0
```

```r
accidents$dep = factor(accidents$dep)
idx = apply(!is.na(accidents[,c("mort","lum","agg","int","atm","col","org","dep")]),1,all)
```

```r
library(ROCR)
# On se limite directement aux ligne complètes pour évacuer le pb des manquants
model3 <- glm(mort ~ lum + agg + int + atm + col + org + dep, family = "binomial",
              data = accidents[idx,])
S <- predict(model3, type="response")
pred <- prediction(S, accidents[idx,"mort"])
#Coordonnées de courbe ROC
perf <- performance(pred, "tpr", "fpr")
# Plot courbe ROC
plot(perf)
```




```r
# Calcul de l'AUC
colAUC(S, accidents[idx,"mort"])
```

```
##              [,1]
## 0 vs. 1 0.8279865
```

```r
# Optimisation du seuil
id_best <- which.min((perf@x.values[[1]])^2 + (1 - perf@y.values[[1]])^2)
alpha_best = perf@alpha.values[[1]][id_best]

Se = perf@y.values[[1]][id_best]
Se
```

```
## [1] 0.7833783
```

```r
Sp = 1 - perf@x.values[[1]][id_best]
Sp
```

```
## [1] 0.74645
```

```r
# Enfin, le TBC peut facilement est déduite de Se et Sp en faisant :
# TBC = Se * P(Y = 1) + Sp * P(Y = 0)
```

```r
# Or dans l'échantillon à disposition P(Y = 1) = P(Y = 0) = 0,5
TBC = 0.5*Se + 0.5*Sp
TBC
```

```
## [1] 0.7649141
```

Les résultats sont donc sensiblement meilleurs que ceux du modèle précédent, mais attention, nous n'avons pas considéré d'échantillon test ici, donc risque de biais d'optimisme ...

4. Redressement des paramètres de la régression logistique en échantillonnage rétrospectif

Faisons maintenant un peu de mathématiques !!! Comme dit au début les proportions dans les données qui vous ont étées fournies ne sont pas représentatives des vraies proportions de mort et de non-mort. Vous allez maintenant montrer que cette modification perturbe assez peu les résultats de la régression logistique.

Question 12 En notant $f_0(x)$ et $f_1(x)$ les densités de probabilité de X dans les classes 0 et 1, puis π_0 et π_1 les proportions des classes 0 et 1, montrer que (2 points)

$$\ln \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \ln \frac{\pi_1 f_1(x)}{\pi_0 f_0(x)}$$

Réponse 12 : On a

$$\ln \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \ln \frac{\frac{P(Y=1,X=x)}{P(X=x)}}{\frac{P(Y=0,X=x)}{P(X=x)}} = \ln \frac{P(Y=1,X=x)}{P(Y=0,X=x)} = \ln \frac{P(Y=1)P(X=x|Y=1)}{P(Y=0)P(X=x|Y=0)} = \ln \frac{\pi_1 f_1(x)}{\pi_0 f_0(x)}$$

Il s'agit ici d'un abus de langage dans les notations car $P(X=x)$ n'a pas de sens ici pour les variables aléatoires continues.

Question 13 : Supposons maintenant qu'on ait ajusté le un modèle de régression logistique mais sur des données issues d'un échantillonnage rétrospectif en imposant $\tilde{\pi}_0$ et $\tilde{\pi}_1$ les proportions des différentes classes. On notera $\tilde{P}(Y=1|X=x)$ les probabilités estimées par ce modèle. En passant par l'équation utilisée dans la question 12, quel est alors le lien entre $\ln \frac{P(Y=1|X=x)}{P(Y=0|X=x)}$ et $\ln \frac{\tilde{P}(Y=1|X=x)}{\tilde{P}(Y=0|X=x)}$. (1 point)

Réponse 13 : En reprenant les résultats de la question 12, on simplement

$$\ln \frac{\tilde{P}(Y=1|X=x)}{\tilde{P}(Y=0|X=x)} = \ln \frac{\tilde{\pi}_1 f_1(x)}{\tilde{\pi}_0 f_0(x)}$$

Ici $f_1(x)$ et $f_0(x)$ restent inchangés, on en déduit donc que

$$\ln \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \ln \frac{\tilde{P}(Y=1|X=x)}{\tilde{P}(Y=0|X=x)} + \ln \frac{\pi_1 \tilde{\pi}_0}{\pi_0 \tilde{\pi}_1}$$

Question 14 : Enfin supposons qu'on ait ajusté un modèle de régression logistique à l'aide de l'échantillonage rétrospectif sur $\ln \frac{\tilde{P}(Y=1|X=x)}{\tilde{P}(Y=0|X=x)}$ sous la forme :

$$\ln \frac{\tilde{P}(Y=1|X=x)}{\tilde{P}(Y=0|X=x)} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \cdots + \tilde{\beta}_d x_d$$

alors comment peut-on en déduire le modèle de régression logistique sur $\ln \frac{P(Y=1|X=x)}{P(Y=0|X=x)}$ si les proportions π_0 et π_1 sont connues. Quels sont alors les liens entre les β_0, β_1, ..., β_d (inconnus) et les $\tilde{\beta}_0$, $\tilde{\beta}_1$, ..., $\tilde{\beta}_d$ (connus). (1 point)

Réponse 14 : Ici on voit que seule l'ordonnée à l'origine est modifiée : $\beta_0 = \tilde{\beta}_0 + \ln \frac{\pi_1 \tilde{\pi}_0}{\pi_0 \tilde{\pi}_1}$. Pour les autres paramètres on conserve $\beta_1 = \tilde{\beta}_1, \ldots, \beta_d = \tilde{\beta}_d$.

Question 15 : En déduire dans le cas des données accident la correction qu'il faut appliquer aux résultats du modèle précédemment appris pour obtenir les "vrais" $P(\text{mort} = 1|X = x)$. (1 point)

Réponse 15 : Ici, on a $\pi_0 = 0.9452849$, $\pi_1 = 0.05471508$, et $\tilde{\pi}_1 = \tilde{\pi}_0 = 0,5$. Ainsi $\beta_0 = \tilde{\beta}_0 - 2,84$, ce qui conduit ici à une dimuntion de la probabilité $P(mort = 1|X = x)$ (conforme à l'intuition), cependant cela ne change rien à l'ordre des probabilités calculées, ainsi la courbe ROC et l'AUC resteraient inchangés.

5. Réflexion autour de l'utilisation de l'analyse discrimante probabiliste

Question 16 : Ici pourquoi sur les données dont vous disposez ne pouvez-vous pas appliquer la LDA ou la QDA ? (1 point)

Réponse 16 : Ici on ne peut pas appliquer la LDA et la QDA car les données considérées sont qualitatives, on ne peut donc pas modéliser leur distribution sachant le groupe par une loi normale.

Question 17 : Utiliser la fonction *naiveBayes* du package *e1071* pour apprendre le modèle prédictif. A l'aide de la documentation de la fonction dire à quelle méthode vue en cours la méthode utilisée est à relier, et expliquer l'hypothèse qui est faite ici. (1 point)

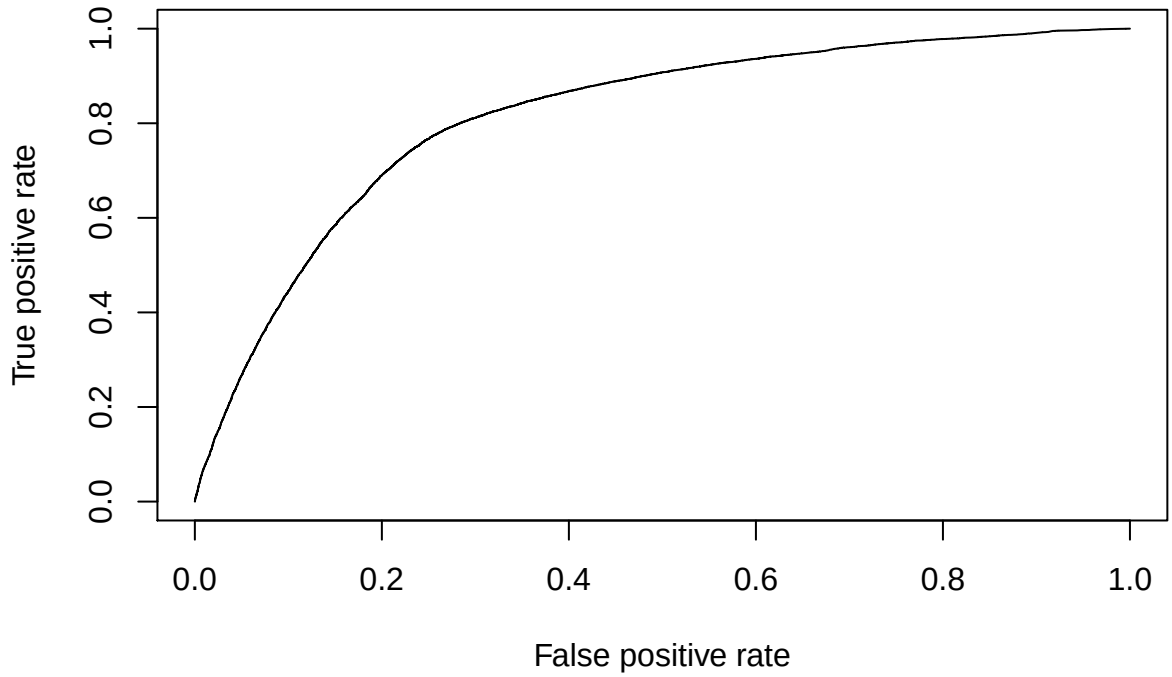
Réponse 7 : L'approche utilisée consiste à supposer l'idépendance de chacune des variables explicative sachant la variable de classe. Ainsi, il est aisé de modéliser la distribution de chacune des variables qualitatives sachant la variable de classe par une simple distribution multinomiale, puis d'en déduire ensuite la probabilité de la classe sachant les covariables à l'aide de l'application du théorème de Bayes. Cette méthode est donc à relier à l'analyse discriminante probabiliste. Le modèle peut donc être estimé comme suit :

```
library(e1071)
nb_accidents <- naiveBayes(mort ~ lum + agg + int + atm + col + org + dep,data = accidents)
```

Remarquons par ailleurs que la modélisation permet de prendre en compte sans aucun mal des données avec des valeurs manquantes, puis que seule les lois du type $X_j|Y = k$ doivent être estimées, ainsi même si un individu comporte quelques valeurs manquantes alors il peut quand même intervenir dans l'estimation de certaines lois $X_j|Y = k$.

Enfin on peut évaluer les performances de ce modèle on peut procéder comme suit :

```
S <- predict(nb_accidents,newdata = accidents,type = "raw")[,2]
pred <- prediction(S, accidents$mort)
#Coordonnées de courbe ROC
perf <- performance(pred, "tpr", "fpr")
# Plot courbe ROC
plot(perf)
```




```r
# Calcul de l'AUC
colAUC(S, accidents$mort)
```

```
##              [,1]
## 0 vs. 1 0.8144915
```

```r
# Optimisation du seuil
id_best <- which.min((perf@x.values[[1]])^2 + (1 - perf@y.values[[1]])^2)
alpha_best = perf@alpha.values[[1]][id_best]
alpha_best
```

```
## [1] 0.5915636
```

```r
Se = perf@y.values[[1]][id_best]
Se
```

```
## [1] 0.7736616
```

```r
Sp = 1 - perf@x.values[[1]][id_best]
Sp
```

```
## [1] 0.7454648
```

```r
# Enfin, le TBC peut facilement est déduite de Se et Sp en faisant :
# TBC = Se * P(Y = 1) + Sp * P(Y = 0)
# Or dans l'échantillon à disposition P(Y = 1) = P(Y = 0) = 0,5
TBC = 0.5*Se + 0.5*Sp
TBC
```

```
## [1] 0.7595632
```

Les résultats obtenus ici sont assez comparables à ceux obtenus par la régression logistique.