Université de Montpellier Année 2018-2019
HMMA304 - Modèles linéaires généralisés

Examen final

Durée 2h - Sans document - 5 novembre 2018

Exercice 1 (10 pts)

1 (4 pts) On rappelle que la loi Beta de paramètre $\alpha > 0$ et $\beta > 0$ admet pour densité de probabilité

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbf{1}_{]0,1[}(x)\,.$$

Décrire les procédures mises en oeuvre par l'intermédiaire du code R suivant.

```
f <- function(x)
{
  sqrt((1-x)*x)/(exp(x)+log(1+x))
}

x <- runif(10000)
mean(f(x))
sqrt(var(f(x)))

x <- rbeta(10000,3/2,3/2)
h <- function(x)
{
  1/(exp(x)+log(1+x))
}
gamma(3/2)^2/gamma(3)*mean(h(x))
gamma(3/2)^2/gamma(3)*sqrt(var(h(x)))
```

2 (2 pts) Donner le code R permettant de générer une réalisation suivant une variable aléatoire X telle que $\mathbb{P}(X = 1) = 0.6$, $\mathbb{P}(X = 2) = 0.3$ et $\mathbb{P}(X = 3) = 0.1$.

2 (4 pts) Pour quel type de données et répondre à quelle question les modèles log-linéaires graphiques sont-ils utilisés ? Donner un exemeple. Quelle procédure peut-on mettre en oeuvre pour sélectionner le modèle log-linéaire graphique le plus adapté à un ensemble de données ?

Exercice 2 (4 pts)

On considère n variables aléatoires indépendantes $y_1, \ldots, y_n$ telles que y_i est distribué suivant une loi binomiale négative de paramètres (h, π_i). Nous supposons que h est fixé. De manière générique, la loi négative binomiale modélise, dans le contexte d'une suite d'épreuves de Bernoulli indépendantes, le nombre d'échecs nécessaires pour obtenir h succès, π représentant la probabilité de succès :

$$f(y; \pi) = C_{h+y-1}^{y} \pi^h (1-\pi)^y \mathbf{1}_{\mathbb{N}}(y) \,.$$

1 (1 pt) Montrer que la loi binomiale négative appartient à la famille exponentielle.

2 (1 pt) Calculer $\mathbb{E}(y)$ et $\mathbb{V}(y)$.

Pour tout $i = 1, \ldots, n$, on suppose que $\log(1 - \pi_i) = \beta_1 + \beta_2 x_i$.

3 (1 pt) Montrer qu'il s'agit d'un modèle linéaire généralisé. La fonction de lien canonique a-t-elle été utilisée ?

4 (1 pt) Donner la log-vraisemblance et les équations de vraisemblance.

Exercice 3 (6 pts)

Expliquer en détails les procédures mises en oeuvre par l'intermédiaire du code R suivant.

```
library(glmnet)
library(nnet)
library(caret)

set.seed(8750)
n <- 1000
p <- 200
x <- matrix(rnorm(n*p), nrow=n, ncol=p)
colnames(x) <- paste("v", 1:p, sep="")

y <- rep(0,n)
for (i in 1:n)
{
  p <- exp(sum(x[i,]))/(1+exp(sum(x[i,])))
  y[i] <- sample(c(0,1),1,prob=c(1-p,p))
}

train_rows <- sample(1:n,0.66*n)
x.train <- x[train_rows, ]
x.test <- x[-train_rows, ]
```

```
y.train <- y[train_rows]
y.test <- y[-train_rows]

train <- data.frame(y=y.train,x.train)
test <- data.frame(y=y.test,x.test)

model1 <- multinom(y~.,family="binomial",data=train)
# weights:  202 (201 variable)
# initial  value 457.477139
# iter  10 value 3.602629
# iter  20 value 1.458515
# iter  30 value 0.921504
# iter  40 value 0.191383
# iter  50 value 0.009916
# iter  60 value 0.001003
# iter  70 value 0.000214
# final  value 0.000086
# converged

mean(predict(model1,test)!=as.factor(y.test))
# [1] 0.1647059

model <- glmnet(x.train, as.factor(y.train), family="binomial", alpha=1)
model2 <- train(x.train, as.factor(y.train), method="glmnet", family="binomial",
                metric="Accuracy", trControl=
                  trainControl(method="repeatedcv", number=5, repeats=10),
                tuneGrid=data.frame(alpha=1, lambda=model$lambda))
mean(predict(model2,x.test)!=as.factor(y.test))
# [1] 0.1735294

y.star <- rep(0,340)
for (i in 1:340)
{
  p <- exp(sum(x.test[i,]))/(1+exp(sum(x.test[i,])))
  if (p>=1/2) y.star[i] <- 1
}
mean(y.star!=y.test)
# [1] 0.02647059
```

1

Examen final
HMMA 304
Correction
5/11/2018

Exercice 1

Nous souhaitons approcher

$$\int_0^1 \left[\frac{\sqrt{x(1-x)}}{e^x + \log(1+x)}\right] dx = I$$

Dans un premier temps, une méthode de Monte Carlo standard est mise en oeuvre. En effet,

$$I = \mathbb{E}_{U_{[0,1]}}\left[\frac{\sqrt{X(1-X)}}{e^X + \log(1+X)}\right]$$

On génère 10000 réalisations iid suivant la loi uniforme : $U_1, \ldots, U_{10000}$.

On estime I par

$$\hat{I}^1 = \frac{1}{10000} \sum_{i=1}^{10000} \left\{ \frac{\sqrt{U_i(1-U_i)}}{e^{U_i} + \log(1+U_i)} \right\}$$

L'écart-type de $\hat{I}^1$ est également estimé.

Dans un second temps, on remarque que

$$I = \mathbb{E}_{Beta\left(\frac{3}{2},\frac{3}{2}\right)} \left[\frac{\left(\Gamma\left(\frac{3}{2}\right)\right)^2}{\Gamma(3)} \frac{1}{e^X + \log(1+X)} \right]$$

Puis, on simule 10000 réalisations suivant une loi $Beta\left(\frac{3}{2},\frac{3}{2}\right)$ et on estime I par

$$\hat{I}^2 = \left(\frac{1}{10000} \sum_{i=1}^{10000} \frac{1}{e^{Y_i} + \log(1+Y_i)} \right) \frac{\Gamma\left(\frac{3}{2}\right)^2}{\Gamma(3)}$$

L'écart-type de $\hat{I}^2$ est également estimé.

2)

sample(c(1,2,3), 1, prob = c(0.6, 0.3, 0.1))

3) Voir cours.

Exercice 2

1)

$$f(y;\pi) = \exp\{n \log(\pi) + y \log(1-\pi)\} C_{n+y-1}^{y} \mathbb{1}_{\mathbb{N}}(y)$$

$$\theta = \log(1-\pi)$$

$$\Rightarrow \pi = 1 - e^{\theta}$$

$$b(\theta) = -n \log(1-e^{\theta})$$

2) $$\mathbb{E}[y] = b'(\theta) = n \frac{e^{\theta}}{1-e^{\theta}} = n\left(\frac{1-\pi}{\pi}\right)$$

$$\mathbb{V}[y] = b''(\theta) = n \frac{e^{\theta}}{(1-e^{\theta})^2} = n\left(\frac{1-\pi}{\pi^2}\right)$$

4

3] $\log(1-\pi) = \beta_1 + \beta_2 x$

$$\Rightarrow 1-\pi = e^{\beta_1+\beta_2 x}$$

$$\Rightarrow \pi = 1 - e^{\beta_1+\beta_2 x}$$

$$\Rightarrow \mathbb{E}[y] = \ln\left[\frac{e^{\beta_1+\beta_2 x}}{1-e^{\beta_1+\beta_2 x}}\right]$$

$$= g(\beta_1+\beta_2 x)$$

$$g(u) = \ln\left[\frac{e^{u}}{1-e^{u}}\right] = b'(u)$$

C'est un modèle linéaire généralisé qui a été utilisé ainsi que le lien canonique.

4] $LV(\beta_1, \beta_2)$

$$= \ln \sum_{i=1}^{n} \log\left(1-e^{\beta_1+\beta_2 x_i}\right) + \sum_{i=1}^{n} y_i\left(\beta_1+\beta_2 x_i\right)$$

$$+ \text{cst}$$

$$\frac{dLV}{d\beta_1}(\beta_1, \beta_2) = 0$$

$$\Longleftrightarrow \sum_{i=1}^{n} y_i - n \sum_{i=1}^{n} \left[\frac{e^{\beta_1 + \beta_2 x_i}}{1 - e^{\beta_1 + \beta_2 x_i}} \right] = 0$$

$$\frac{dLV}{d\beta_2}(\beta_1, \beta_2) = 0$$

$$\Longleftrightarrow \sum_{i=1}^{n} y_i x_i - n \sum_{i=1}^{n} \left[\frac{x_i e^{\beta_1 + \beta_2 x_i}}{1 - e^{\beta_1 + \beta_2 x_i}} \right] = 0$$

Pas de solution analytique.

Exercice 3

On génère un jeu de données contenant 1000 individus, 200 variables prédictives et une variable binaire à expliquer suivant un modèle de régression logistique dont tous les coefficients de variables sont égaux à 1.

On isole 33% des données comme données de test.
On compare les taux d'erreur des classifieurs oracle, par régression logistique classique et régression logistique avec régularisation L1 dont le paramètre de régularisation est calibré par validation croisée à 5 ensembles répétés 10 fois.

6