Examen de contrôle continu

Durée 1h - Sans document - 15 octobre 2018

Exercice 1 (5 pts)

1 (2 pts) Quelle est la différence entre les modèles logit et probit ?

2 (3 pts) Quelle est l'objectif des méthodes de régularisation en regression ?

Exercice 2 (3 pts)

On considère le modèle de régression suivant, pour tout $i = 1, \ldots, n$

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

avec $\mathbb{E}(u_i) = 0$, $\mathbb{V}(u_i) = 1$ si i est pair et $\mathbb{V}(u_i) = 2$ si i est impair, $\mathbb{C}(u_i, u_j) = 0$ si $i \neq j$. Donner l'expression de l'estimateur des moindres carrés généralisés de β_1, commenter.

Exercice 3 (2 pts)

On considère n variables aléatoires indépendantes $y_1, \ldots, y_n$ telles que y_i est distribuée suivant une loi de Poisson de paramètre λ_i.

1 (0.5 pt) Montrer que la loi de Poisson appartient à la famille exponentielle.

2 (0.5 pt) Calculer $\mathbb{E}(y_i)$ et $\mathbb{V}(y_i)$.

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

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

Exercice 4 (10 pts)
Expliquer en détails les procédures mises en oeuvre par l'intermédiaire du code R suivant.

```
library(glmnet)
library(caret)

set.seed(875)
n <- 1000
p <- 500
pm <- 50
x <- matrix(rnorm(n*p), nrow=n, ncol=p)
colnames(x) <- paste("v", 1:p, sep="")
y <- apply(x[,1:pm], 1, sum) + rnorm(n)


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]

model1 <- glmnet(x.train, y.train, family="gaussian", alpha=1, nlambda=50)
model2 <- train(x.train, y.train ,method="glmnet", metric="RMSE", trControl=
trainControl(method="repeatedcv", number=5, repeats=10),
       tuneGrid=data.frame(alpha=1, lambda=model1$lambda))
model3 <- glmnet(x.train, y.train, family="gaussian", alpha=0)
model4 <- train(x.train, y.train, method="glmnet", metric="RMSE", trControl=
trainControl(method="repeatedcv", number=5, repeats=10),
       tuneGrid=data.frame(alpha=0, lambda=model3$lambda))

dtrain <- data.frame(y=y.train, x.train)
model5 <- lm(y~., data=dtrain)
dtest <- data.frame(x.test)

mean((predict(model2, x.test) - y.test)^2)
mean((predict(model4, x.test) - y.test)^2)
mean((predict(model5, dtest) - y.test)^2)
mean((apply(x.test[,1:pm], 1, sum)-y.test)^2)

## [1] 1.475314
## [1] 3.141372
## [1] 3.729522
## [1] 1.032703
```

Correction examen contrôle continu HMMA304

15/10/2018

Exercice 1

1) Il s'agit de deux modèles linéaires généralisés associés à la loi de Bernoulli, la variable à expliquer y est binaire. Ces deux modèles ont des fonctions de lien différentes. Soit $p = \mathbb{E}(y) = \mathbb{P}(y = 1)$

Pour le modèle logit, on suppose que

$$p = e^{x\beta} / (1 + e^{x\beta})$$

Pour le modèle probit, on suppose que $p = F_{N(0,1)}(X\beta)$.

2) Les méthodes de régularisation sont utiles lorsque le nombre de régresseurs (variables explicatives) est proche du nombre d'observations.
Objectif : diminuer la variance des estimateurs des paramètres.

Exercice 2

$$\hat{\beta}_1^{GLS} \in \arg\min_{\beta_1 \in \mathbb{R}} \sum_{i=1}^{n} \left[\frac{(y_i - \beta_1 - \beta_2 x_i)^2}{\alpha_i} \right]$$

avec $\alpha_i = 1$ si i est pair et
$\alpha_i = 2$ si i est impair.

3

Soit $h(\beta_1, \beta_2) = \sum_{i=1}^{n} \left[\frac{(y_i - \beta_1 - \beta_2 x_i)^2}{\alpha_i} \right]$

Nous avons

$$\frac{\partial h}{\partial \beta_1}(\beta_1, \beta_2) = -2\left[\sum_{i=1}^{n} \left(\frac{y_i}{\alpha_i}\right) - \beta_1 \sum_{i=1}^{n} \left(\frac{1}{\alpha_i}\right) - \beta_2 \sum_{i=1}^{n} \frac{x_i}{\alpha_i} \right]$$

$$\frac{\partial h}{\partial \beta_2}(\beta_1, \beta_2) = -2\left[\sum_{i=1}^{n} \frac{y_i x_i}{\alpha_i} - \beta_1 \sum_{i=1}^{n} \frac{x_i}{\alpha_i} - \beta_2 \sum_{i=1}^{n} \frac{x_i^2}{\alpha_i} \right]$$

Ainsi, $\frac{\partial h}{\partial \beta_1}(\beta_1^*, \beta_2^*) = 0$ et $\frac{\partial h}{\partial \beta_2}(\beta_1^*, \beta_2^*) = 0$

$$\Longleftrightarrow \begin{cases} \beta_1^* \sum \left(\frac{1}{\alpha_i}\right) + \beta_2^* \sum \frac{x_i}{\alpha_i} = \sum \frac{y_i}{\alpha_i} \\ \beta_1^* \sum \frac{x_i}{\alpha_i} + \beta_2^* \sum \frac{x_i^2}{\alpha_i} = \sum \frac{y_i x_i}{\alpha_i} \end{cases}$$

$$\Longleftrightarrow \begin{cases} \beta_1^* = \dfrac{\sum \frac{y_i}{d_i} - \beta_2^* \sum \frac{x_i}{d_i}}{\sum \frac{1}{d_i}} \\[2ex] \beta_2^* = \dfrac{\sum \frac{y_i x_i}{d_i} - \beta_1^* \sum \frac{x_i}{d_i}}{\sum \frac{x_i^2}{d_i}} \end{cases}$$

$$\Longrightarrow \sum \frac{1}{d_i} \beta_1^* = \sum \frac{y_i}{d_i} - \frac{\left[\sum \frac{y_i x_i}{d_i} - \beta_1^* \sum \frac{x_i}{d_i}\right] \sum \frac{x_i}{d_i}}{\sum \frac{x_i^2}{d_i}}$$

$$\Longrightarrow \beta_1^* \left[\sum \frac{1}{d_i} \sum \frac{x_i^2}{d_i} - \left(\sum \frac{x_i}{d_i}\right)^2\right] = \sum \frac{y_i}{d_i} \sum \frac{x_i^2}{d_i} - \left(\sum \frac{x_i}{d_i}\right)^2$$

$$\Longrightarrow \beta_1^* = \left[\frac{\sum \frac{y_i}{d_i} \sum \frac{y_i x_i}{d_i} - \left(\sum \frac{x_i}{d_i}\right)^2}{\sum \frac{1}{d_i} \sum \frac{x_i^2}{d_i} - \left(\sum \frac{x_i}{d_i}\right)^2}\right]$$

Par ailleurs,

$$\frac{\partial^2 h}{(\partial \beta_1)^2}(\beta_1,\beta_2) = 2\sum\left(\frac{1}{d_i}\right)$$

$$\frac{\partial^2 h}{(\partial \beta_2)^2}(\beta_1,\beta_2) = 2\sum \frac{x_i^2}{d_i}$$

$$\frac{\partial^2 h}{\partial \beta_1 \partial \beta_2}(\beta_1,\beta_2) = 2\sum \frac{x_i}{d_i}$$

Comme $2\sum\left(\frac{1}{d_i}\right) > 0$ et

$$4\sum\left(\frac{1}{d_i}\right)\sum \frac{x_i^2}{d_i} - 4\left(\sum \frac{x_i}{d_i}\right)^2 > 0$$

(en effet, soit $u_i = \frac{1}{\sqrt{d_i}}$ et $v_i = \frac{x_i}{\sqrt{d_i}}$ d'après Cauchy-Schwarz

$$\left(\sum u_i v_i\right)^2 \leq \sum u_i^2 \sum v_i^2 \;)$$

alors la matrice hessienne de h est définie positive et donc h strictement convexe.

6

$$\hat{\beta}_1^{GLS} = \beta_1^* .$$

Exercice 3

1] $f(y;\lambda) = e^{-\lambda} \frac{\lambda^y}{y!} \mathbb{1}_{\mathbb{N}}(y)$

$= \exp\{-\lambda + y\log(\lambda)\} \frac{1}{y!} \mathbb{1}_{\mathbb{N}}(y)$

$\theta = \log(\lambda) \Rightarrow \lambda = e^{\theta}$

$b(\theta) = e^{\theta}$

$\nu(dy) = \frac{1}{y!} \mathbb{1}_{\mathbb{N}}(y)\, \delta(dy)$

où δ est la mesure de comptage

2] $\mathbb{E}(y) = b'(\theta) = e^{\theta} = \lambda$

$\mathbb{V}(y) = b''(\theta) = e^{\theta} = \lambda$

3] $\mathbb{E}(y) = \lambda = e^{\beta_1 + \beta_2 x}$

$\Rightarrow \mathbb{E}(y) = g(\beta_1 + \beta_2 x)$

Il s'agit bien d'un modèle linéaire généralisé.

Par ailleurs, $g(u) = e^u = b'(u)$

Ainsi, c'est bien la fonction de lien canonique qui a été utilisée.

Exercice 4

Le vrai modèle contient 50 régresseurs générés à partir d'une loi gaussienne centrée réduite.

Nous avons

$$y = x_1 + \cdots + x_{50} + u$$

où $u \sim \mathcal{N}(0,1)$

La première partie du code consiste à générer des données suivant ce modèle.
Par ailleurs, 450 variables de bruit sont ajoutées au modèle. On compare alors les résultats produits par des régressions ridge, lasso et classique sur l'ensemble des variables prédictives. Les paramètres de régularisation des régressions lasso et ridge est fixé à l'aide d'une stratégie de validation croisée à 5 ensembles répétée 10 fois.