Université de Montpellier
HAX912X - Modèles linéaires généralisés

Année 2023-2024

Examen de contrôle continu

Durée 1h - Sans document
Vendredi 6 octobre 2023

Exercice 1 (10 pts)

Décrire et commenter les résultats obtenus à l'aide du code R ci-dessous.

```
> library(glmnet)
Loading required package: Matrix
Loaded glmnet 4.1-6
> library(plotmo)
Loading required package: Formula
Loading required package: plotrix
Loading required package: TeachingDemos
> library(caret)
Loading required package: ggplot2
Loading required package: lattice

> set.seed(1974)

> n <- 100
> d <- 50
> X <- matrix(rnorm(n*d),n,d)
> colnames(X) <- paste("x",1:d,sep="")
> y <- as.vector(2+X%*%rep(1,d)+rnorm(n,sd=2))
> training.set <- data.frame(y=y,X)

> Xtest <- matrix(rnorm(n*d),n,d)
> colnames(Xtest) <- paste("x",1:d,sep="")
> ytest <- 2+Xtest%*%rep(1,d)+rnorm(n,sd=2)
> test.set <- data.frame(Xtest)

> yhat <- 2+Xtest%*%rep(1,d)
> sqrt(mean((yhat-ytest)^2))
[1] 2.362255

> model <- lm(y~.,data=training.set)
> yhat <- predict(model,test.set)
> sqrt(mean((yhat-ytest)^2))
[1] 3.513195
```

```
> model <- train(X,y,method="glmnet",metric="RMSE",
+         trControl=trainControl(method="repeatedcv",
+                                number=n,repeats=10),
+         tuneGrid=data.frame(alpha=0,lambda = seq(0,1,length=50)))
> model
glmnet

100 samples
50 predictor

No pre-processing
Resampling: Cross-Validated (100 fold, repeated 10 times)
Summary of sample sizes: 99, 99, 99, 99, 99, 99, ...

Tuning parameter 'alpha' was held constant at a value of 0
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0 and lambda = 0.

> yhat <- predict(model,test.set)
> sqrt(mean((yhat-ytest)^2))
[1] 3.513195

> y <- as.vector(2+X%*%rep(1,d)+rnorm(n,sd=5))
> training.set$y <- y
> ytest <- 2+Xtest%*%rep(1,d)+rnorm(n,sd=5)

> yhat <- 2+Xtest%*%rep(1,d)
> sqrt(mean((yhat-ytest)^2))
[1] 4.760175

> model <- lm(y~.,data=training.set)
> yhat <- predict(model,test.set)
> sqrt(mean((yhat-ytest)^2))
[1] 6.816788

> model <- train(X,y,method="glmnet",metric="RMSE",
+               trControl=trainControl(method="repeatedcv",
+                                      number=n,repeats=10),
+               tuneGrid=data.frame(alpha=0,lambda = seq(0,20,length=50)))
> model
glmnet

100 samples
50 predictor

No pre-processing
Resampling: Cross-Validated (100 fold, repeated 10 times)
Summary of sample sizes: 99, 99, 99, 99, 99, 99, ...
```

```
Tuning parameter 'alpha' was held constant at a value of 0
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0 and lambda = 1.632653.

> yhat <- predict(model,test.set)
> sqrt(mean((yhat-ytest)^2))
[1] 6.555984
```

Exercice 2 (2 pts)

Expliquer en quoi consiste une regression LASSO.

Exercice 3 (2 pts)

Expliquer la construction de l'estimateur des moindres carrés généralisés.

Exercice 4 (6 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.

Correction examen partiel
6 octobre 2023
HAX912X

Exercice 1

5 points doivent être mentionnés

1) Nous simulons deux jeux de données un échantillon d'apprentissage contenant 100 individus et un échantillon de test contenant également 100 individus suivant le modèle de régression ci-dessous

$$y = 2 + \sum_{j=1}^{d} x_j + u$$ (2 pts)

où $u \sim N(0,4)$

$x_j \sim N(0,1)$ $\forall j = 1, \dots, 50$

Tous les coefficients du modèle de régression hormis la constante sont égaux à 1 -

2| Nous comparons la racine de l'erreur quadratique moyenne de deux modèles : celle du modèle oracle où les valeurs des coefficients et de la constante sont supposées connues, erreur que nous ne pourrons battre, celle d'un modèle de régression linéaire multiple non régularisé. Pour l'oracle, nous obtenons # 2,5 et pour le modèle de régression # 3,5 -

3| Un modèle de régression régularisé par contrainte L2 (ridge) est ensuite calibré en utilisant une stratégie de validation croisée à 100 ensembles répétée 10 fois -

Il est finalement estimé que la régularisation n'est pas necessaire $\lambda = 0$. La racine de l'erreur quadratique moyenne est identique à celle du modèle de régression multiple. (3)

4] Deux nouveaux jeux de données sont simulés suivant le modèle ci-dessous

$$y = 2 + \sum_{j=1}^{d} x_j + u$$ (2pts)

où $u \sim N(0,25)$

$x_j \sim N(0,1) \quad \forall j = 1, \ldots, 50$

La racine de l'erreur quadratique moyenne augmente simultanément avec la régression multiple.

5] Le modèle de regularisation ridge calibré comme précédemment améliore légèrement les résultats. (2)

Exercice 2

Régularisation avec contrainte l2

On minimise les moindres carrés sous contrainte $\sum|\beta_i|$ soit petit.
On introduit ainsi un biais forçant les coefficients à se rapprocher de 0 et ce faisant nous réduisons la variance des estimateurs des coefficients de régression. Cela peut engendrer des gains importants sur l'erreur quadratique moyenne et donc permettre d'obtenir de meilleures prédictions.

Exercice 3

Estimateur MCG tenant compte de l'hétéroscédasticité des erreurs d'observation

On minimise en β $(y-X\beta)^T\Sigma^{-1}(y-X\beta)$.

Exercice 4

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

$$E(u_i) = 0$$

$$V(u_i) = 1 \text{ si } i \text{ est pair}$$

$$V(u_i) = 2 \text{ si } i \text{ est impair}$$

$$C(u_i, u_j) = 0 \quad \forall i \neq j$$

$$\Sigma = \begin{bmatrix} 2 & & & & & \\ & 1 & & & 0 & \\ & & 2 & & & \\ & & & 1 & & \\ & 0 & & & \ddots & \\ & & & & & 2 \\ & & & & & & 1 \end{bmatrix}$$

nous supposons n pair

$$\underbrace{\sum_{k=1}^{n/2} \left(y_{2k} - \beta_1 - \beta_2 x_{2k}\right)^2 + \frac{1}{2} \sum_{k=1}^{n/2} \left(y_{2k-1} - \beta_1 - \beta_2 x_{2k-1}\right)^2}_{= h(\beta_1, \beta_2)}$$

On minimise $h(\cdot)$ par rapport à (β_1, β_2).

$$\frac{\partial h}{\partial \beta_1}(\beta_1, \beta_2) = -2 \sum_{k=1}^{n/2} \left(y_{2k} - \beta_1 - \beta_2 x_{2k}\right) - \sum_{k=1}^{n/2} \left(y_{2k-1} - \beta_1 - \beta_2 x_{2k-1}\right)$$

$$\frac{\partial h}{\partial \beta_2}(\beta_1,\beta_2) = -2\sum_{k=1}^{n/2} x_{2k}\left(y_{2k}-\beta_1-\beta_2 x_{2k}\right) - \sum_{k=1}^{n/2} x_{2k-1}\left(y_{2k-1}-\beta_1-\beta_2 x_{2k-1}\right)$$

6

$$\Longleftrightarrow \begin{cases} \dfrac{\partial h}{\partial \beta_1}(\beta_1,\beta_2)=0 \\ \dfrac{\partial h}{\partial \beta_2}(\beta_1,\beta_2)=0 \end{cases}$$

$$\Longleftrightarrow \begin{cases} -\sum_{i=1}^{n} y_i - \sum_{k=1}^{n/2} y_{2k} + \dfrac{3n}{2}\beta_1 + \beta_2\left(\sum_{i=1}^{n} x_i + \sum_{k=1}^{n/2} x_{2k}\right) = 0 \\ -\sum_{i=1}^{n} x_i y_i - \sum_{k=1}^{n/2} x_{2k} y_{2k} + \beta_2\left(\sum_{i=1}^{n} x_i^2 + \sum_{k=1}^{n/2} x_{2k}^2\right) + \beta_1\left(\sum_{i=1}^{n} x_i + \sum_{k=1}^{n/2} x_{2k}\right) = 0 \end{cases}$$

7

$$C_y = \sum_{i=1}^{m} y_i + \sum_{k=1}^{m/2} y_{2k}$$

$$C_x = \sum_{i=1}^{m} x_i + \sum_{k=1}^{m/2} x_{2k}$$

$$C_{xy} = \sum_{i=1}^{m} x_i y_i + \sum_{k=1}^{m/2} x_{2k} y_{2k}$$

$$C_{x^2} = \sum_{i=1}^{m} x_i^2 + \sum_{k=1}^{m/2} x_{2k}^2$$

$$\begin{cases} -C_y + \frac{3m}{2}\beta_1 + C_x \beta_2 = 0 \\ -C_{xy} + C_x \beta_1 + C_{x^2} \beta_2 = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \beta_1 = \dfrac{2(C_y - C_x\beta_2)}{3m} \\ -C_{xy} + C_x \left[\dfrac{2(C_y - C_x\beta_2)}{3m}\right] + C_{x^2}\beta_2 = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \beta_1 = \dfrac{2(C_y - C_x\beta_2)}{3m} \\ \beta_2 = \dfrac{\dfrac{3mC_{xy} - 2C_xC_y}{3m}}{\dfrac{3mC_{x^2} - 2(C_x)^2}{3m}} \end{cases}$$

$$\Leftrightarrow \begin{cases} \beta_2^* = \dfrac{3m\, C_{xy} - 2 C_x C_y}{3m\, C_{x2} - 2(C_x)^2} \\[2ex] \beta_1^* = \left[\dfrac{2(C_y - C_x \beta_2^*)}{3m} \right] \end{cases}$$

$$\hat{\beta}_1^{MCO} = \frac{2(C_y - C_x \hat{\beta}_2^{MCO})}{3m}$$

$$\hat{\beta}_1^{MCO} = \frac{2\,C_y}{3m} - \frac{2C_x}{3m} \hat{\beta}_2^{MCO}$$

$$\hat{\beta}_1^{MCO} = \bar{y} - \bar{x}\, \hat{\beta}_2^{MCO}$$