Université de Montpellier
HAX912X - Modèles linéaires généralisés

Année 2023-2024

Examen de contrôle continu

Durée 1h - Sans document
Vendredi 20 octobre 2023

Exercice 1 (10 pts)

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

```
import os
import pandas as pd
import numpy as np

from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import RepeatedKFold
from sklearn.model_selection import GridSearchCV

path = os.getcwd()
os.chdir('/Users/marin/TEACHING/2324/M2-GLM-HAX912X/TP')

data = pd.read_csv("creditcard.csv")
data.shape

(284807, 31)

count_classes = pd.value_counts(data['Class'], sort = True).sort_index()
count_classes

0   284315
1   492

X = data.loc[:, data.columns != 'Class']
y = data['Class']
scaler = preprocessing.StandardScaler().fit(X)
X = scaler.transform(X)
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.3,
random_state = 1974)

print("Number transactions train dataset: ", len(X_train))
print("Number transactions test dataset: ", len(X_test))
print("Total number of transactions: ", len(X_train)+len(X_test))
```

```
Number transactions train dataset:  199364
Number transactions test dataset:  85443
Total number of transactions:  284807

logit = LogisticRegression()
logit.fit(X_train,y_train)
y_chap = logit.predict(X_test)
table = pd.crosstab(y_test,y_chap)
table

    0      1
0   85272  19
1   48     104

Crossvalid = RepeatedKFold(n_splits=2, n_repeats=2)
parameters = {'C':np.linspace(0.1,2.1,num=5)}
logit = GridSearchCV(LogisticRegression(), parameters,
                     cv=Crossvalid,scoring='neg_log_loss')
logit.fit(X_train,y_train)
print(logit.best_params_)

{'C': 0.1}

y_chap = logit.predict(X_test)
table=pd.crosstab(y_test,y_chap)
table

    0       1
0   85273   18
1   49      103
```

Exercice 2 (10 pts)

Nous considérons n variables aléatoires indépendantes $y_1, \ldots, y_n$ telles que y_i est distribué suivant une loi de Pareto de paramètre $(1, k_i)$ où $k_i > 0$. La densité de la loi de Pareto de paramètre $(1, k)$ est telle que

$$f(y; k) = \frac{k}{y^{k+1}} \mathbf{1}_{[1,+\infty[}(y) .$$

Pour tout $i = 1, \ldots, n$, nous supposons que $\log(k_i) = \beta_0 + \beta_1 x_i$ où $x_i \in \mathbb{R}$ est supposé connu, β_0 et β_1 sont des paramètres inconnus.

1 (4 pts) Montrer qu'il s'agit d'un modèle linéaire généralisée. La fonction de lien canonique a-t-elle été utilisée ?

2 (6 pts) Donner la log-vraisemblance et les équations de vraisemblance. Pouvons-nous calculer les expressions analytique des estimateurs du maximum de vraisemblance de β_0 et β_1 ?

Correction Examen Partiel
20 octobre 2023
HAX912X

Exercice 1

6 points doivent être mentionnés

1) Importation données (1 pt)
284 807 individus
31 variables

2) Variable binaire Class (1 pt)
grand déséquilibre entre
les modalités 0 et 1

3) y variable binaire à
expliquer Class (1 pt)
X matrice des variables
explicatives
les variables contenues dans X
sont centrées réduites.

2

4] Les données sont séparées dans un échantillon d'apprentissage contenant 199 364 individus et un échantillon de test contenant 85 443 individus

1pt

5] Modèle de régression logistique est mis en œuvre par défaut la matrice de confusion sur l'échantillon de test montre seulement 67 mal classés parmi 85 443 transactions frauduleuses (#8 pour 10 000)

3pts

Taux d'erreurs beaucoup plus fort parmi la classe 1 = 48 sur 152 (# 32 pour 100)

3

6] Modèle de régression logistique régularisé avec le paramètre de régularisation choisi par validation croisée à 2 ensembles répétée 2 fois sur 5 valeurs de C régulièrement espacées entre 0,1 et 2,1 et pour le critère de la log-vraisemblance négative du modèle de régression logistique.

3pts

Le paramètre de régularisation choisi est C = 0,1 c'est un peu discutable car au bord du domaine de recherche. Les résultats sur l'échantillon de test sont très similaires à ceux obtenus sans calibration de C

Exercice 2 ①

1) $f(y;k) = \frac{k}{y^{k+1}} \mathbb{1}_{[1,+\infty[}(y)\, \lambda(dy)$

$f(y;k) = \exp\{-k \log(y) + \log(k)\} \underbrace{\frac{1}{y}\mathbb{1}_{[1,+\infty[}(y)}_{\nu(dy)}$

Changement de variable

$u = \log(y) \Rightarrow y = e^{u}$

$du = \left(\frac{1}{y}\right) dy$

nous en déduisons

$f(u;k) = k e^{-ku} \mathbb{1}_{]0;+\infty[}(u)$

$u \sim \mathrm{Exp}(k)$

$\mathbb{P}_k(du) = \exp(-ku + \log(k)) \underbrace{\mathbb{1}_{]0,+\infty[}(u)\, d\, du}_{\nu(dy)}$

$\theta = -k$

$k(\theta) = -\log(-\theta)$

5

Par ailleurs,

$$\log(k) = \beta_0 + \beta_1 x \implies k = e^{\beta_0+\beta_1 x}$$

$$\mathbb{E}(Y) = \frac{1}{k} = e^{-\beta_0-\beta_1 x} = g^{-1}(\beta_0+\beta_1 x)$$

Moyennant le changement de variable $u = \ln(y)$, il s'agit d'un modèle linéaire généralisé associé à la fonction de lien $g(z) = -\log(z)$ $\left(g^{-1}(z) = e^{-z}\right)$

$$g(z) \neq b'(z) = -\frac{1}{z}$$

ce n'est pas le lien canonique qui a été utilisé.

2] $$L(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[\beta_0 + \beta_1 x_i\right] - \sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i} \log(y_i)$$

6

$$\frac{dL}{d\beta_0}(\beta_0, \beta_1) = n - \sum_{i=1}^{n} \log(y_i) e^{(\beta_0 + \beta_1 x_i)}$$

$$\frac{dL}{d\beta_1}(\beta_0, \beta_1) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i \log(y_i) e^{(\beta_0 + \beta_1 x_i)}$$

Nous ne pouvons pas calculer les expressions exactes des estimateurs du maximum de vraisemblance.