Université de Montpellier Année 2024-2025
HAX912X - Modèles linéaires généralisés

Examen de contrôle continu

Durée 1h - Aucun document autorisé
Vendredi 18 octobre 2024

Exercice 1 (6 pts)

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

Aide paramètre `C` de la fonction `LogisticRegression` de `sklearn`

```
C : float, default=1.0
Inverse of regularization strength; must be a positive float.
Like in support vector machines, smaller values specify stronger regularization.
```

```
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import make_classification

X, y = make_classification(n_samples=300, n_features=40,
    n_classes=2,flip_y=0.1,class_sep=0.5, random_state=42)

X = StandardScaler().fit_transform(X)

model1 = LogisticRegression(penalty='l1', C=1, class_weight=None,
dual=False, solver = 'saga', fit_intercept=True, l1_ratio=None, max_iter=10000, tol=0.0001)
model1.fit(X,y)
print(model1.score(X,y))
0.8233333333333334

model2 = LogisticRegression(penalty='l1', C=0.1, class_weight=None,
dual=False, solver = 'saga', fit_intercept=True, l1_ratio=None, max_iter=10000, tol=0.0001)
model2.fit(X,y)
print(model2.score(X,y))
0.7866666666666666

from sklearn.model_selection import train_test_split
X,X_test,y,y_test = train_test_split(X,y,test_size=0.25)

model1 = LogisticRegression(penalty=None, class_weight=None,
dual=False, solver = 'saga', fit_intercept=True, l1_ratio=None, max_iter=10000, tol=0.0001)
model1.fit(X,y)
print(model1.score(X_test,y_test))
0.68

model2 = LogisticRegression(penalty='l1', C=0.1, class_weight=None,
dual=False, solver = 'saga', fit_intercept=True, l1_ratio=None, max_iter=10000, tol=0.0001)
model2.fit(X,y)
print(model2.score(X_test,y_test))
0.7466666666666667
```

Exercice 2 (5 pts)

Nous considérons n variables aléatoires indépendantes $y_1, \ldots, y_n$ telles que y_i est distribuée suivant une loi binomiale de paramètre $(100, p_i)$. Pour tout $i = 1, \ldots, n$, nous supposons que $\log(-\log(1-p_i)) = x_i^T \beta$ où $x_i \in \mathbb{R}^p$ est un vecteur supposé connu et β un vecteur de paramètres inconnus. Montrer qu'il s'agit d'un modèle linéaire généralisé. La fonction de lien canonique a-t-elle été utilisée ?

Exercice 3 (4 pts)

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

```
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import make_regression
import numpy as np

X, y = make_regression(n_samples=100, n_features=80, noise=0.1, random_state=42)

scaler = StandardScaler()
X_normalized = scaler.fit_transform(X)

model1 = LinearRegression()
scores1 = cross_val_score(model1, X_normalized, y, cv=5, scoring='neg_mean_absolute_error')
print(scores1)
print(np.mean(scores1))

[-10.96963938 -44.27345903 -40.90757601 -32.031635   -18.96821433]
-29.430104749593802

alpha = 0.1
model2 = Lasso(alpha=alpha)
scores2 = cross_val_score(model2, X_normalized, y, cv=5, scoring='neg_mean_absolute_error')
print(scores2)
print(np.mean(scores2))

[-0.34950166 -0.3056262  -0.30595072 -0.46432015 -0.29677541]
-0.3444348268695149
```

Exercice 4 (5 pts)

1 (2,5 pts) Quelle est la différence entre les modèles logit et probit ?

2 (2,5 pts) Quelle est l'objectif des méthodes de régularisation en regression ?

Correction Controlli Continui 3
HAX912X
25/10/2024

Esercizio 1

1) $f(y,\pi) = \binom{k+y-1}{y} \pi^k (1-\pi)^y \mathbb{1}_{\mathbb{N}}(y)$

$f(y;\pi) = \exp\{k \log(\pi) + y \log(1-\pi)\} \mathbb{1}(y)$

$\theta = \log(1-\pi) \Rightarrow \pi = 1 - e^\theta$

$b(\theta) = -k \log(1-e^\theta)$

$b'(\theta) = \frac{k e^\theta}{1-e^\theta}$ Famiglia esponenziale 1-dim

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

$x \quad E(y) = b'(\theta) = \frac{k e^\theta}{1-e^\theta} = k\left(\frac{1-\pi}{\pi}\right)$

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

$$\Rightarrow E(y) = n \left\{ \frac{e^{\beta_1 + \beta_2 x}}{1 - e^{\beta_1 + \beta_2 x}} \right\}$$

C'est bien un modèle linéaire généralisé

Car $\eta = \log(1-\pi) = \beta_1 + \beta_2 x$

donc c'est bien la bien fonction qui a été utilisé.

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

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

$$\frac{\partial LV_n(\beta_1, \beta_2)}{\partial \beta_1} = -n \sum_{i=1}^{m} \left[\frac{e^{\beta_1 + \beta_2 x_i}}{1 - e^{\beta_1 + \beta_2 x_i}} \right] + \sum_{i=1}^{m} y_i$$

$$\frac{\partial LV_m}{\partial \beta_2}(\beta_1,\beta_2) = -h \sum_{i=1}^{m} x_i \left[\frac{e^{\beta_1+\beta_2 x_i}}{1-e^{\beta_1+\beta_2 x_i}} \right] + \sum_{i=1}^{m} y_i x_i$$

Exercice 2

Modèle de régression logistique

→ $y \in \{0,1\}$ = variable indiquant la présence ou l'absence d'un défaut, variable à expliquer

→ $x \in \mathbb{R}$ température au moment du lancement, variable explicative

$$\mathbb{P}(y=1 \mid x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

On estime (β_0, β_1) avec les 24 données disponibles, on obtient $(\hat{\beta}_0, \hat{\beta}_1)$.

Enfin, on prédit

$P(y=1 \mid x=31)$ avec

$$\left[\frac{e^{\hat{\beta}_0 + 31\hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + 31\hat{\beta}_1}}\right]$$

Code associé :

```
modèle <- glm(vol_ ~ temp, family = "binomial"
              , data = challenger)
coeff <- modèle$coefficients
x <- exp(coeff[1] + coeff[2] * 31)
x / (1 + x)
```

Exercice 3

(5)

1 point → les densités montrent que la variable Precip est intéressante pour expliquer le pourcentage de ruissellement mais que ce n'est pas le cas de la variable StormDuration.

1 point → un premier modèle logistique est mis en œuvre Precip ayant gardé : test asymptotique de nullité du coefficient a une p-valeur très faible

1 point → un deuxième modèle logistique est mis en œuvre StormDuration on ne garde pas p-valeur très (trop) forte

1 point → pour le modèle 1 avec Recipe, on estime par validation croisée à 10 ensembles répétés 50 fois la matrice de confusion. difficulté à prédire classe 1 (ruissellement) mais on y arrive un peu.

1 point → pour le modèle 2 avec stormDuration, incapable à prédire la classe 1

Exercice 4

1 point → $n = 5000$ individus
$p = 10$ variables
régression → X_1
en présent 100 variables différentes → X_2

2 points → 2 modalités de forêts aléatoires, l'un sans les variables de bruit, l'autre avec

données séparées en deux
- un échantillon de training
- un échantillon de test

5000 5000

2 points → comparaison sur échantillons de test

même résultats

aucune influence du bruit