Université de Montpellier Année 2017-2018
HMMA304 - Modèles linéaires généralisés

Examen final - 10 novembre 2017
Durée 2h00 - Documents interdits

Exercice 1 (4 pts)

1 (2 pts) Expliquer en quoi consiste une regression lasso ?

2 (2 pts) Expliquer à quelle sous-classe de modèles log-linéaires appartiennent les modèles log-linéaires graphiques. Donner un exemple et quelques relations d'indépendance conditionnelle associées.

Exercice 2 (10 pts)

Nous considèrons n variables aléatoires indépendantes $y_1, \dots, y_n$ telles que y_i est distribuée suivant une loi de Bernoulli de paramètre $p_i \in]0,1[$

$$f(y_i; p_i) = p_i^{y_i}(1-p_i)^{1-y_i}\mathbf{1}_{y_i \in \{0,1\}}\,.$$

1 (2 pts) Montrer que la loi de Bernoulli appartient à la famille exponentielle et déduire $\mathbb{E}(y_i)$ et $\mathbb{V}(y_i)$.

Pour tout $i = 1, \dots, n$, nous supposons que $\log\left(\dfrac{p_i}{1-p_i}\right) = a + bx_i$ où $x_i \in \mathbb{R}$ est supposé connu, a et b sont des paramètres inconnus.

2 (1 pt) Montrer qu'il s'agit d'un modèle linéaire généralisé. La fonction de lien canonique a-t-elle été utilisée ?

3 (2,5 pts) Donner la log-vraisemblance et les équations de vraisemblance. Est-il possible de calculer (de manière générale) les expressions analytiques des estimateurs du maximum de vraisemblance de a et b ?

4 (2,5 pts) Calculer la matrice d'information de Fisher apportée par $(y_1, \dots, y_n)$ sur les paramètres a et b notée $I_n(a,b)$. Expliquer comment l'on peut construire un intervalle de confiance sur a.

5 (2 pts) Nous disposons de $n = 100$ observations (y_i, x_i) décrites dans la table ci-dessous

	$y=0$	$y=1$
$x=0$	20	40
$x=1$	10	30

.

Pour ce cas particulier, donner les expressions analytiques des estimateurs du maximum de vraisemblance de a et b.

Exercice 3 (6 pts)

1 (4 pts) Expliquer en détails les procédures mises en oeuvre par l'intermédiaire du code R ci-dessous et expliciter le résultat obtenu.

```
> library(gRim)
Le chargement a nécessité le package : gRbase
> data(reinis)
> model1 <- dmod(~ .^., reinis)
> model2 <- backward(model1, criterion = "aic")
  change.AIC  -19.7744 Edge deleted: mental systol
  change.AIC   -8.8511 Edge deleted: phys systol
  change.AIC   -4.6363 Edge deleted: mental protein
  change.AIC   -1.6324 Edge deleted: systol family
  change.AIC   -3.4233 Edge deleted: family protein
  change.AIC   -0.9819 Edge deleted: phys family
  change.AIC   -1.3419 Edge deleted: smoke family
```

Description du jeu de données
reinis gRbase R Documentation Risk factors for coronary heart disease
Data collected at the beginning of a 15 year follow-up study of probable risk factors for coronary thrombosis. Data are from all men employed in a car factory.
A table with 6 discrete variables

- A smoking,
- B strenous mental work,
- D strenuous physical work,
- E systolic blood pressure,
- F ratio of lipoproteins,
- G Family anamnesis of coronary heart disease.

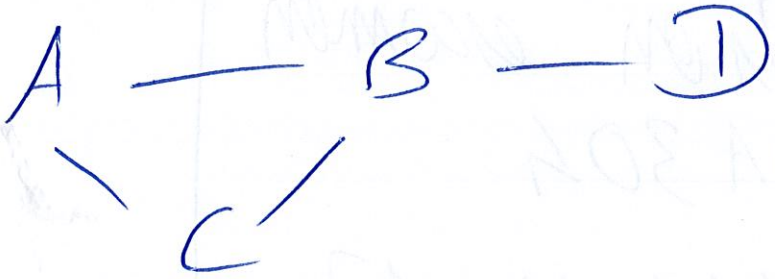
2 (2 pts) Soit le vecteur `y` contenant $n = 80$ réalisations d'une variable à expliquer réelle et la matrice `X` contenant les valeurs correspondantes de $p = 70$ variables explicatives réelles. Donner le code R permettant de mettre en oeuvre un modèle de regression lasso.

Correction examen
HMMA 304
10/11/2017

Exercice 1

1] Moindres carrés pénalisés par contrainte L_1, régularisation à appliquer lorsque n est proche de p.

2] Les modèles log-linéaires graphiques appartiennent à un sous-ensemble des modèles log-linéaires hiérarchiques définis par un graphe.



A*B*C + B*D est un modèle hiérarchique graphique, ce n'est pas le cas de A*B + B*C + A*C + B*D

Exercice 2

1] $f(y;p) = p^{y}(1-p)^{1-y}\,\Pi_{\{0,1\}}(y)$

$\Leftrightarrow f(y;p) = \exp\left\{ y \log\left(\frac{p}{1-p}\right) + \log(1-p)\right\}\Pi_{\{0,1\}}(y)$

$\theta = \log\left(\frac{p}{1-p}\right)$, $b(\theta) = -\log(1-p)$

$p = \frac{e^{\theta}}{1+e^{\theta}}$, $b(\theta) = \log(1+e^{\theta})$

$\nu(dy) = \Pi_{\{0,1\}}(y)\, S(dy)$ où S mesure de comptage

$$\mathbb{E}(y) = b'(\theta) = \frac{e^\theta}{1+e^\theta} = \rho$$

$$\mathbb{V}(y) = b''(\theta) = \frac{e^\theta(1+e^\theta) - e^\theta e^\theta}{(1+e^\theta)^2} = \rho(1-\rho)$$

2] $\mathbb{E}(y) = \rho = \dfrac{e^{\alpha+b x_i}}{1+e^{\alpha+b x_i}} = g(\alpha+b x_i)$

C'est bien un modèle linéaire généralisé.

Par ailleurs, $g(u) = \dfrac{e^u}{1+e^u} = b'(u)$

C'est bien le lien canonique qui a été utilisé.

3] $\mathcal{L}Vm(\alpha, b) = \sum y_i (\alpha + b x_i - \log(1+e^{\alpha+b x_i}))$

$$- \sum (1-y_i) \log(1+e^{\alpha+b x_i})$$

$$\mathcal{L}Vm(\alpha, b) = \alpha \sum y_i + b \sum y_i x_i$$

$$- \sum \log(1+e^{\alpha+b x_i})$$

$$\frac{\partial LVm}{\partial a}(a,b) = \sum y_i - \sum \left[\frac{e^{a+b x_i}}{1+e^{a+b x_i}} \right]$$

$$\frac{\partial LVm}{\partial b}(a,b) = \sum y_i x_i - \sum x_i \left[\frac{e^{a+b x_i}}{1+e^{a+b x_i}} \right]$$

Voici l'expression analytique pour les ETIV.

4] $$\frac{\partial^2 LVm}{(\partial a)^2}(a,b) = - \sum \frac{e^{a+b x_i}}{(1+e^{a+b x_i})^2}$$

$$\frac{\partial^2 LVm}{\partial a \partial b}(a,b) = - \sum x_i \frac{e^{a+b x_i}}{(1+e^{a+b x_i})^2}$$

$$\frac{\partial^2 LVm}{(\partial b)^2}(a,b) = - \sum x_i^2 \frac{e^{a+b x_i}}{(1+e^{a+b x_i})^2}$$

On en déduit $I_m(a,b)$

Intervalle de confiance asymptotique en utilisant le fait que

$$(\hat{\alpha}^{EMV} - \alpha) \overset{\#}{\sim} \mathcal{N}\left(0, \underbrace{\left[I_n(\alpha,\beta)\right]^{-1}_{(1,1)}}_{\text{élément } (1,1) \text{ de la matrice}}\right)$$

5] $\sum y_i = 70, \quad \sum y_i x_i = 30$

$$\begin{cases} \dfrac{\partial LV_n(\hat{\alpha},\hat{\beta})}{\partial \alpha} = 70 - 60\dfrac{e^{\hat{\alpha}}}{1+e^{\hat{\alpha}}} - 40\dfrac{e^{\hat{\alpha}+\hat{\beta}}}{1+e^{\hat{\alpha}+\hat{\beta}}} = 0 \\ \dfrac{\partial LV_n(\hat{\alpha},\hat{\beta})}{\partial \beta} = 30 - 40\dfrac{e^{\hat{\alpha}+\hat{\beta}}}{1+e^{\hat{\alpha}+\hat{\beta}}} = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \dfrac{40e^{\hat{\alpha}+\hat{\beta}}}{1+e^{\hat{\alpha}+\hat{\beta}}} = 30 \\ 40 - 60\dfrac{e^{\hat{\alpha}}}{1+e^{\hat{\alpha}}} \end{cases} \Leftrightarrow \begin{cases} 20e^{\hat{\alpha}} = 40 \\ 10e^{\hat{\alpha}+\hat{\beta}} = 30 \end{cases}$$

$$\Leftrightarrow \hat{\alpha} = \log 1 \,/\, \hat{\beta} = \log\left(\frac{3}{2}\right)$$

Exercice 3

1] Sélection de modèle log-linéaires graphiques par méthode descendante et critère AIC.

2] Voir TD