Introduction à l'apprentissage supervisé

Y. Pradat (*auteur*), L. Verlingue & D. Gautheret (*responsables cours*)

CentraleSupelec (labo MICS) & Institut Gustave Roussy (IGR)

23 janvier 2020

Notations

Notations

- $n, p \in \mathbb{N}^*$ nombre d'observations (unités, individus) et nombre de variables (facteurs, cofacteurs) ;
- X (resp $\mathbf{X}$) variable aléatoire scalaire (resp vectorielle) ;
- $\mathbf{X_{1:n}}$ vecteur (resp matrice) des variables aléatoires $\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n}$ (resp $\mathbf{X_1}, \ldots, \mathbf{X_n}$) ;
- x (resp $\boldsymbol{x}$) observation de X (resp $\mathbf{X}$) ;
- $\boldsymbol{x_{1:n}}$ observation de $\mathbf{X_{1:n}}$.

Objectif :

1. Proposer un modèle mathématique ;
2. Estimer les paramètres du modèle.

Definition vraisemblance

Définition 1. Modèle statistique

Un modèle statistique est une collection de lois (ou densités) candidates paramétrée par $\theta \in \Theta$

$$\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\} \tag{1}$$

Définition 2. Vraisemblance

Soient $\boldsymbol{x}_{\mathbf{1}:\boldsymbol{n}}$ *un échantillon d'observations de* $\mathbf{X}_{\mathbf{1:n}} \sim p_{\theta^*}$. *La vraisemblance du paramètre* θ *pour cet échantillon est*

$$\mathcal{L}(\theta; \mathbf{x}_{1:n}) = \prod_{i=1}^{n} p_\theta(x_i) \tag{2}$$

Definition vraisemblance

Définition 1. Modèle statistique

Un modèle statistique est une collection de lois (ou densités) candidates paramétrée par $\theta \in \Theta$

$$\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\} \tag{1}$$

Définition 2. Vraisemblance

Soient $\mathbf{x_{1:n}}$ *un échantillon d'observations de* $\mathbf{X}_{1:n} \sim p_{\theta^*}$. *La vraisemblance du paramètre* θ *pour cet échantillon est*

$$\mathcal{L}(\theta; \mathbf{x}_{1:n}) = \prod_{i=1}^{n} p_\theta(x_i) \tag{2}$$

Exemple 1

Soit $\mathbf{x_{1:n}}$ *un n-échantillon de* $\mathbf{X}_{1:n} \sim \mathcal{N}(\mu^*, \sigma^{2^*})$. *Alors, la vraisemblance de* (μ, σ^2) *est*

$$\mathcal{L}(\mu, \sigma^2; \mathbf{x_{1:n}}) = \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2} \tag{3}$$

Estimation par MV

Définition 3. Estimateur du MV

Pour un échantillon $\boldsymbol{x}_{1:n}$ *de* $\mathbf{X}_{1:n} \sim p_{\theta^*}$*, l'équation*

$$\widehat{\theta^*}(\boldsymbol{x}_{1:n}) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \, \mathcal{L}(\theta; \boldsymbol{x}_{1:n}) \tag{4}$$

définit un estimateur du maximum de vraisemblance de θ^**.*

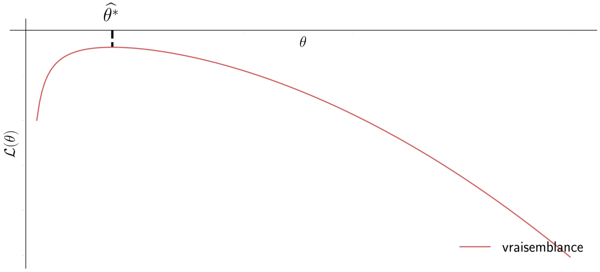
Estimation par MV

Définition 3. Estimateur du MV

Pour un échantillon $\boldsymbol{x}_{1:n}$ *de* $\mathbf{X}_{1:n} \sim p_{\theta^*}$*, l'équation*

$$\widehat{\theta^*}(\boldsymbol{x}_{1:n}) \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(\theta; \boldsymbol{x}_{1:n}) \tag{4}$$

définit un estimateur du maximum de vraisemblance de θ^**.*



Exemple estimation par MV

Définition 3. Estimateur du MV

Pour un échantillon $\boldsymbol{x}_{\mathbf{1:n}}$ *de* $\mathbf{X}_{1:n} \sim p_{\theta^*}$*, l'équation*

$$\widehat{\theta^*}(\boldsymbol{x}_{\mathbf{1:n}}) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(\theta; \boldsymbol{x}_{\mathbf{1:n}}) \tag{5}$$

définit un estimateur du maximum de vraisemblance de θ^**.*

Exemple 2

Soit $\boldsymbol{x}_{\mathbf{1:100}}$ *un* 100*-échantillon de* $\mathbf{X}_{1:100} \sim \mathcal{B}(\theta^*)$ *(100 lancés de pièces identiques). L'estimateur par MV de* θ *maximise*

$$\prod_{i=1}^{100} p_{\mathrm{X_i}}(x_i; \theta) \tag{6}$$

Exemple estimation par MV

Définition 3. Estimateur du MV

Pour un échantillon $\boldsymbol{x_{1:n}}$ *de* $\mathbf{X}_{1:n} \sim p_{\theta^*}$*, l'équation*

$$\widehat{\theta^*}(\boldsymbol{x_{1:n}}) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(\theta; \boldsymbol{x_{1:n}}) \tag{5}$$

définit un estimateur du maximum de vraisemblance de θ^**.*

Exemple 2

Soit $\boldsymbol{x_{1:100}}$ *un* 100*-échantillon de* $\mathbf{X}_{1:100} \sim \mathcal{B}(\theta^*)$ *(100 lancés de pièces identiques). L'estimateur par MV de* θ *maximise*

$$\prod_{i=1}^{100} p_{\mathrm{X_i}}(x_i; \theta) \tag{6}$$

Pour une loi de Bernoulli, $p(x; \theta) = \theta^x (1-\theta)^{1-x}$ *de sorte que*

$$\hat{\theta^*}(\boldsymbol{x_{1:100}}) = \frac{1}{100} \sum_{i=1}^{100} x_i \tag{7}$$

Construction modèle

<u>Le modèle</u> Soit $\boldsymbol{x_{1:n}} \in (\mathbb{R}^p)^n$ un n-échantillon de $\mathbf{X_{1:n}}$ et soient, pour $\boldsymbol{x_{1:n}}$ fixé, les variables aléatoires $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n}$ données par

$$\forall i = 1, \ldots, n, \qquad \mathrm{Y_i} = \boldsymbol{x_i}^\top \beta + \mathcal{E}_i \tag{8}$$

avec $\mathcal{E}_i \sim \mathcal{N}(0, \sigma^2)$ (résidus).

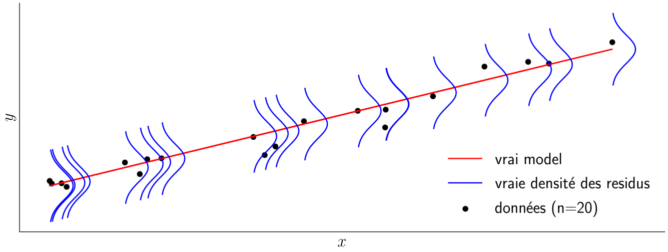
Construction modèle

Le modèle Soit $\boldsymbol{x_{1:n}} \in (\mathbb{R}^p)^n$ un n-échantillon de $\mathbf{X_{1:n}}$ et soient, pour $\boldsymbol{x_{1:n}}$ fixé, les variables aléatoires $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n}$ données par

$$\forall i = 1, \ldots, n, \qquad \mathrm{Y}_\mathrm{i} = \boldsymbol{x_i}^\top \beta + \mathcal{E}_i \tag{8}$$

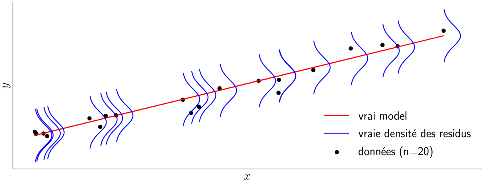
avec $\mathcal{E}_i \sim \mathcal{N}(0, \sigma^2)$ (résidus).

Explication



Construction modèle

Explication



- Les paramètres du modèle sont (β, σ^2).
- C'est un modèle statistique $\mathcal{M}_{\beta,\sigma^2}$ sur les densités conditionnelles $p_{\mathrm{Y_i}|\mathbf{x_i}=x_i}$.
- **Hypothèses du modèle**
  1. Relation linéaire entre $\mathbf{x_i}$ et $\mathrm{Y_i}$;
  2. Résidus indépendants et gaussiens ;
  3. Homoscedasticité.

Estimation paramètre β

Par maximum de vraisemblance,

$$\mathcal{L}(\beta, \sigma^2; \boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}}) = \prod_{i=1}^{n} p_{\mathrm{Y_i|X_i}=\boldsymbol{x_i}}(y_i) \tag{9}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x_i}^\top \beta)^2} \tag{10}$$

<u>Exercice</u> : Calculer l'estimateur $\hat{\beta}(\boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}})$ de β par MV en minimisant $\ell(\beta, \sigma^2) = -\log \mathcal{L}(\beta, \sigma^2)$.

Estimation paramètre β

Par maximum de vraisemblance,

$$\mathcal{L}(\beta, \sigma^2; \boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}}) = \prod_{i=1}^{n} p_{\mathrm{Y_i}|\mathrm{X_i}=\boldsymbol{x_i}}(y_i) \tag{9}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x_i}^\top \beta)^2} \tag{10}$$

<u>Exercice</u> : Calculer l'estimateur $\hat{\beta}(\boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}})$ de β par MV en minimisant $\ell(\beta, \sigma^2) = -\log \mathcal{L}(\beta, \sigma^2)$.

Tous calculs faits, si $\boldsymbol{x_{1:n}^\top x_{1:n}}$ inversible,

$$\hat{\beta}(\boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}}) = \left(\boldsymbol{x_{1:n}^\top x_{1:n}}\right)^{-1} \boldsymbol{x_{1:n}^\top y_{1:n}} \tag{11}$$

Construction modèle

Le modèle Soit $\boldsymbol{x_{1:n}} \in (\mathbb{R}^p)^n$ un n-échantillon de $\mathbf{X_{1:n}}$ et soient, pour $\boldsymbol{x_{1:n}}$ fixé, les variables aléatoires $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n}$ données par

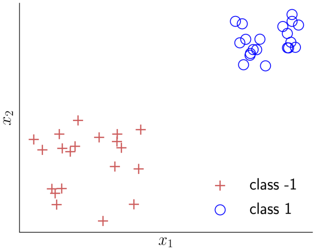
$$\forall i = 1, \ldots, n, \qquad \mathrm{Y_i} \sim \mathcal{B}(\sigma(x_i^\top \theta)) \tag{12}$$

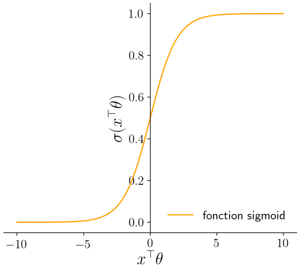
Construction modèle

Le modèle Soit $\boldsymbol{x_{1:n}} \in (\mathbb{R}^p)^n$ un n-échantillon de $\mathbf{X_{1:n}}$ et soient, pour $\boldsymbol{x_{1:n}}$ fixé, les variables aléatoires $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n}$ données par

$$\forall i = 1, \ldots, n, \qquad \mathrm{Y}_\mathrm{i} \sim \mathcal{B}(\sigma(x_i^\top \theta)) \tag{12}$$

Explication





Estimation paramètre θ

Explications

- Les paramètres du modèle sont θ.
- C'est un modèle statistique $\mathcal{M}_\theta$ sur les densités conditionnelles $p_{\mathrm{Y_i}|\mathbf{X_i}=x_i}$.
- **Hypothèses du modèle**
  1. Relation linéaire entre $\mathrm{logit}\, p_{\mathrm{Y_i}|\mathbf{X_i}=\boldsymbol{x_i}}(1)$ et $\boldsymbol{x_i}$;

Estimation Par maximum de vraisemblance,

$$\mathcal{L}(\theta; \boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}}) = \prod_{i=1}^{n} p_{\mathrm{Y_i}|\mathrm{X_i}=x_i}(y_i) \tag{13}$$

$$= \prod_{i=1}^{n} \sigma(\theta^\top x_i)^{y_i} (1 - \sigma(\theta^\top x_i))^{1-y_i} \tag{14}$$

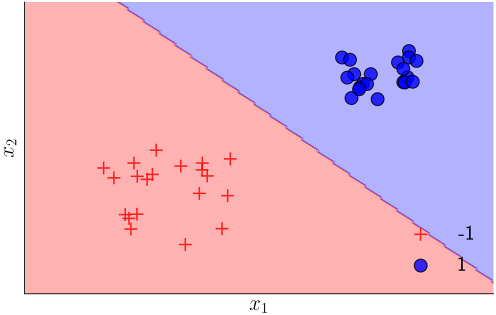
$$\ell(\theta; \boldsymbol{x_{1:n}}, \boldsymbol{y_{1:n}}) = -\sum_{i=1}^{n} y_i \log \sigma(\theta^\top \boldsymbol{x_i}) + (1 - y_i) \log(1 - \sigma(\theta^\top \boldsymbol{x_i})) \tag{15}$$

Estimation paramètre θ

Comme $\sigma'(z) = \sigma(z)(1 - \sigma(z)$ on obtient que :

$$\nabla \ell(\theta) = -\sum_{i=1}^{n} \boldsymbol{x_i}(y_i - \sigma(\boldsymbol{x_i}^\top \theta)) \tag{16}$$

qui nous sert à minimiser $\ell(\theta)$ *numériquement*.



Modélisation de survie