FONDEMENTS DU ML

05/11/2024

Aujourd'hui

→ Maximum a posteriori (+ suite la semaine prochaine ?)

→ TD 4 (Partie 1)

Dernier cours (12/11) : Présentation projet

→ TP régression linéaire

MAXIMUM A POSTERIORI

Contexte . Données $\left(X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times n},\ y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m\right)$

• Hypothèse : $y = X\beta^* + \varepsilon$

$\Leftrightarrow y_i = x_i^T \beta^* + \varepsilon_i \quad \forall i = 1..m$

$\beta^* \in \mathbb{R}^n$

$\varepsilon \in \mathbb{R}^m$ vecteur de bruit (aléatoire)

• Bruit gaussien : $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_m) \quad \sigma > 0$

$\Leftrightarrow \varepsilon_1, \dots, \varepsilon_m$ sont i.i.d $\sim \mathcal{N}(0, \sigma^2)$

$\Leftrightarrow \mathbb{E}[\varepsilon] = 0_{\mathbb{R}^m}$ et $\mathbb{E}[\varepsilon\varepsilon^T] = \sigma^2 I_m$

(1) Maximum de vraisemblance → Maximum a posteriori

→ L'estimateur du maximum de vraisemblance $\hat{\beta}^{MV} \in \mathbb{R}^n$ est défini comme une solution de

$$(P_{MV}) \quad \max_{\beta \in \mathbb{R}^n} \mathcal{L}(y_1, \dots, y_m ; \beta) \overset{\text{sous nos hypothèses}}{=} \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\sum_{i=1}^m \frac{(y_i - x_i^T\beta)^2}{2\sigma^2}\right)$$

Vraisemblance : Probabilité que $Y_1 = y_1, \dots, Y_m = y_m$ si $Y_i \sim \mathcal{N}(x_i^T\beta, \sigma^2)$

→ L'ensemble des solutions de (P_{MV}) est identique à celui de

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2\sigma^2} \|X\beta - y\|^2$$

ou encore

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2} \|X\beta - y\|^2$$

⇒ Problème aux moindres carrés linéaires, dont on sait caractériser les solutions.

↳ Plutôt que la vraisemblance, on aurait pu vouloir maximiser $\mathcal{L}(\beta; y_1, \dots, y_m)$

mais cette loi n'a pas d'expression directe simple (on ne peut pas calculer cette probabilité en général)

→ En revanche, comme

$$\mathcal{L}(\beta; y_1, \dots, y_m) \propto \mathcal{L}(y_1, \dots, y_m; \beta)\, \mathcal{L}(\beta)$$

("proportionnel à")

$\mathcal{L}(y_1, \dots, y_m; \beta)$: vraisemblance

$\mathcal{L}(\beta)$: loi inconnue en général, mais que l'on peut définir via un a priori sur notre modèle linéaire

(formule de Bayes)

Définition Etant donnée une loi de probabilité $\mathcal{L}(\beta)$ sur $\mathbb{R}^n$, l'estimateur du maximum a posteriori pour (X, y) est défini comme solution du problème

$$\max_{\beta \in \mathbb{R}^n} \ln\left[\mathcal{L}(y_1, \dots, y_m; \beta)\, \mathcal{L}(\beta)\right]$$

On le note $\hat{\beta}^{MAP}$

NB : Peut être défini sans le logarithme

$$\max_{\beta \in \mathbb{R}^n} \mathcal{L}(y_1, \dots, y_m; \beta)\, \mathcal{L}(\beta)$$

Implication : . Pour toute loi a priori sur β, on peut définir un maximum a posteriori

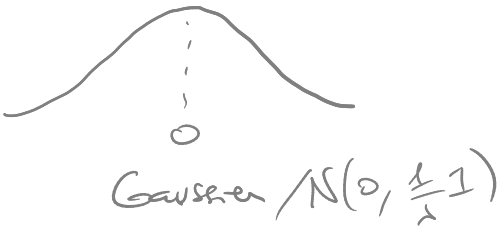
. Selon la loi, on peut disposer ou non d'une formule explicite pour $\hat{\beta}^{MAP}$, ainsi que de garanties statistiques.

. La loi a priori dépend souvent d'un paramètre qui permet de quantifier le poids de l'a priori par rapport à la vraisemblance

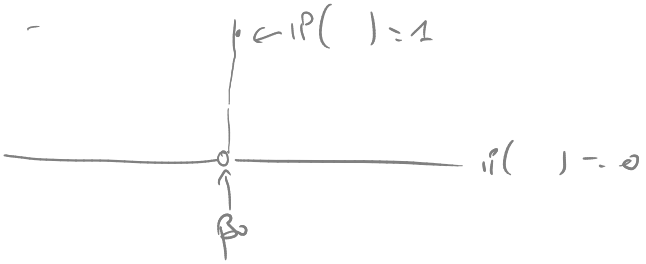
Exemples d'a priori

- $\beta \sim \mathcal{N}(0, (1/\lambda) I_m)$ $\lambda > 0$
- $\beta \sim \text{Laplace}(0, \frac{1}{\lambda})$ $\lambda > 0$

$z \in \mathbb{R}$, $z \sim \text{Laplace}(\bar{z}, 1/\lambda)$, $\mathbb{P}(z = t) \propto \exp\left(-\frac{|t - \bar{z}|}{\lambda}\right)$



- $\beta \sim \text{Dirac}(\beta_0)$ $\mathbb{P}(\beta = \beta_0) = 1$ $\mathbb{P}(\beta \neq \beta_0) = 0$



...

=> Le choix de l'a priori dépend du problème / de l'expertise métier

=> On peut toujours faire sans a priori (maximum de vraisemblance)

② Cas d'un a priori gaussien

On suppose un a priori gaussien centré en 0 sur β, càd que on considère $\beta \sim N\left(0, \left(\frac{1}{\lambda}\right) I_m\right)$ $\quad \lambda > 0$

$\Rightarrow$ Revient à dire que l'on cherche une solution proche du vecteur nul

$\Rightarrow$ La valeur de λ détermine/quantifie à quel point on souhaite considérer des valeurs éloignées de 0

$\lambda \to 0$: A priori $\to$ Pas d'a priori (variance infinie)

$\lambda \to \infty$: A priori $\to$ Dirac(0)

Sous les hypothèses de départ ($y = X\beta^* + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I_m)$) et avec cet a priori, on a :

$$\mathcal{L}(y_1, \dots, y_m ; \beta) = \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(- \frac{1}{2\sigma^2} \sum_{i=1}^{m} \left(x_i^T \beta - y_i \right)^2 \right)$$

et

$$\mathcal{L}(\beta) = \frac{1}{(2\pi/\lambda)^{m/2}} \exp\left(- \frac{1}{2/\lambda} \sum_{j=1}^{m} \beta_j^2 \right)$$

$\forall j \quad \beta_j \sim N(0, 1/\lambda)$

Par conséquent,

$$\ln\left[\mathcal{L}(y_1, \dots, y_m ; \beta) \mathcal{L}(\beta) \right] = \ln\left(\frac{1}{(2\pi\sigma^2)^{\frac{m}{2}}} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} \left(x_i^T \beta - y_i \right)^2$$
$$+ \ln\left(\frac{1}{(2\pi/\lambda)^{m/2}} \right) - \frac{1}{2/\lambda} \sum_{j=1}^{m} \beta_j^2$$

Le problème d'estimation du maximum a posteriori est

$$\max_{\beta \in \mathbb{R}^n} \left\{ \ln\left(\frac{1}{(2\pi\sigma^2)^{\frac{m}{2}}}\right) + \ln\left(\frac{1}{(2\pi/\lambda)^{n/2}}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i^T\beta - y_i)^2 - \frac{1}{2/\lambda} \sum_{j=1}^{n} \beta_j^2 \right\}$$

constantes (ne dépendent pas de β)

qui a le même ensemble de solutions que

$$\max_{\beta \in \mathbb{R}^n} -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i^T\beta - y_i)^2 - \frac{1}{2/\lambda} \sum_{j=1}^{n} \beta_j^2 = -\frac{1}{2\sigma^2}\|X\beta - y\|^2 - \frac{\lambda}{2}\|\beta\|^2$$

Théorème Sous les hypothèses ci-dessus

- Le problème d'estimation du maximum a posteriori a le même ensemble de solutions que

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2\sigma^2}\|X\beta - y\|^2 + \frac{\lambda}{2}\|\beta\|^2$$

qui est un problème aux moindres carrés linéaires.

- Pour $\lambda > 0$, les deux problèmes possèdent une unique solution donnée par

$$\hat{\beta}^{MAP} = \left(X^TX + \lambda\sigma^2 I_n\right)^{-1} X^T y$$

Principe de preuve

$\rightarrow \max_{\beta \in \mathbb{R}^n} -f(\beta) \iff \min_{\beta \in \mathbb{R}^n} f(\beta)$

$$\rightarrow \frac{1}{2\sigma^2}\|X\beta - y\|^2 + \frac{\lambda}{2}\|\beta\|^2 = \frac{1}{2\sigma^2}\left(\|X\beta - y\|^2 + \lambda\sigma^2\|\beta\|^2 \right)$$

$$= \frac{1}{2\sigma^2}\left(\beta^T X^T X\beta - 2\beta^T X^T y + y^T y + \lambda\sigma^2 \beta^T\beta \right)$$

$$= \frac{1}{2\sigma^2}\left(\beta^T\left(X^TX + \lambda\sigma^2 I\right)\beta - 2\beta^T X^T y + y^T y \right)$$

$\forall X \in \mathbb{R}^{n\times m}$, $X^TX + \lambda\sigma^2 I \succ 0$

car $\left(X^TX + \lambda\sigma^2 I\right)^T = \left(X^TX\right)^T + \lambda\sigma^2 I^T = X^TX + \lambda\sigma^2 I$

et $\forall v \in \mathbb{R}^m$, $v^T\left(X^TX + \lambda\sigma^2 I\right)v = \underbrace{\|Xv\|^2}_{\geq 0} + \underbrace{\lambda\sigma^2\|v\|}_{>0 \text{ si } v\neq 0}$

> 0 si $v \neq 0$

Par conséquent, il existe $M \succ 0$ telle que $M^2 = M^TM = X^TX + \lambda\sigma^2 I$

En utilisant cette matrice, on écrit

$$\beta^T\left(X^TX + \lambda\sigma^2 I\right)\beta - 2\beta^T X^T y + y^T y$$

$$= \beta^T M^T M \beta - 2\beta^T \underbrace{M^T M^{-T}}_{I} X^T y + y^T y$$

$$= \|M\beta - M^{-T}X^T y\|^2 - y^T X M^{-1} M^{-T} X^T y + y^T y$$

$$a^2 - 2ab = (a-b)^2 - b^2$$

$$a^Ta - 2a^Tb = \|a-b\|^2 - b^Tb$$

Le problème aux moindres carrés s'écrit

$$\min_{\beta\in\mathbb{R}^m} \frac{1}{2\sigma^2}\left(\|M\beta - M^{-T}X^T y\|^2 - y^T X M^{-1} M^{-T} X^T y + y^T y \right)$$

$\frac{1}{2\sigma^2}$: constante > 0

$y^T X M^{-1} M^{-T} X^T y + y^T y$: constante indépendante de β

qui a le même ensemble de solutions que

$$\min_{\beta\in\mathbb{R}^m} \frac{1}{2}\|M\beta - M^{-T}X^T y\|^2$$

$\rightarrow M > 0 \Rightarrow M$ est de rang plein et carrée

$\Rightarrow$ le problème a une unique solution donnée par

$$\hat{\beta} = M^{-1} M^{-T} X^T y$$

$$= (M^{-1})^2 X^T y$$

$$= (M^2)^{-1} X^T y = (X^T X + \lambda \sigma^2 I)^{-1} X^T y$$

Remarques:

- Lorsque $\varepsilon \sim N(0, \sigma^2 I_m)$, on définit l'a priori gaussien en posant $\beta \sim N(0, \frac{\sigma^2}{\lambda} I_n)$, ce qui permet d'éliminer σ^2 dans le résultat final.

- Le problème aux moindres carrés $\min_{\beta \in \mathbb{R}^n} \frac{1}{2\sigma^2} \|X\beta - y\|^2 + \frac{\lambda}{2} \|\beta\|^2$ s'appelle un problème avec notions équivalentes
  - régularisation ℓ_2
  - régularisation de Tychonoff
  - régularisation écrêtée ("ridge")

- On dit que la régularisation ℓ_2 réduit la variance de la solution par rapport aux données

Ex) $X = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ $y = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$\hat{\beta}^{MV} = \hat{\beta}^{OLS} = X^\dagger y = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$= \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$\|\hat{\beta}^{MV}\| = 1$

On perturbe $X \rightarrow \tilde{X} = \begin{bmatrix} 1 & 0 \\ 0 & \eta \end{bmatrix}$ $0 < \eta \ll 1$

$$(\tilde{X}, y) \longrightarrow \tilde{X}^T y = \begin{bmatrix} 1 & 0 \\ 0 & 1/\eta \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1/\eta \end{bmatrix}$$

$$\|\tilde{X}^T y\| \xrightarrow[\eta \to 0]{} \infty$$

Avec régularisation ℓ_2 / a priori gaussien

$$\hat{\beta}^{MAP} = (\tilde{X}^T \tilde{X} + \lambda I)^{-1} \tilde{X}^T y$$

$$= \begin{bmatrix} 1+\lambda & 0 \\ 0 & \eta^2+\lambda \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \eta \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{1+\lambda} \\ \frac{\eta}{\eta^2+\lambda} \end{bmatrix} \qquad \|\hat{\beta}^{MAP}\| \xrightarrow[\eta \to \infty]{} \frac{1}{1+\lambda}$$

$$(X^T X + \lambda I)^{-1} X^T y = \begin{bmatrix} \frac{1}{1+\lambda} \\ 0 \end{bmatrix}$$

Quand $\lambda \to \infty$, $\hat{\beta}^{MAP} \to \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

Propriétés statistiques pour $\hat{\beta}^{MAP}$ dans le cas d'un a priori gaussien

Sous les hypothèses $y = X\beta^* + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I_m)$ et l'a priori $\beta \sim N(0, (1/\lambda) I_m)$, on pose $\hat{\beta}^{MAP} = (X^T X + \lambda \sigma^2 I)^{-1} X^T y$

Alors 1) $\mathbb{E}[\hat{\beta}^{MAP}] = (X^T X + \lambda \sigma^2 I)^{-1} X^T X \beta^* \neq \beta^*$ en général

2) La matrice de covariance de $\hat{\beta}^{MAP}$ vaut

$\Sigma_{\hat{\beta}^{MAP}} = (X^T X + \lambda \sigma^2 I)^{-1}$, et $\| \Sigma_{\hat{\beta}^{MAP}} \| \leq \| \Sigma_{\hat{\beta}^{MV}} \|$

avec $\hat{\beta}^{MV} = X^+ y$ l'estimateur du maximum de vraisemblance.

→ On dit que l'estimateur $\hat{\beta}^{MAP}$ réduit la variance au prix d'un biais plus élevé

Bilan

→ Principe du maximum a posteriori et problème d'estimation / d'optimisation associé

→ le cas de l'a priori gaussien : équivalence avec un problème aux moindres carrés, solution (unique !) et propriétés statistiques