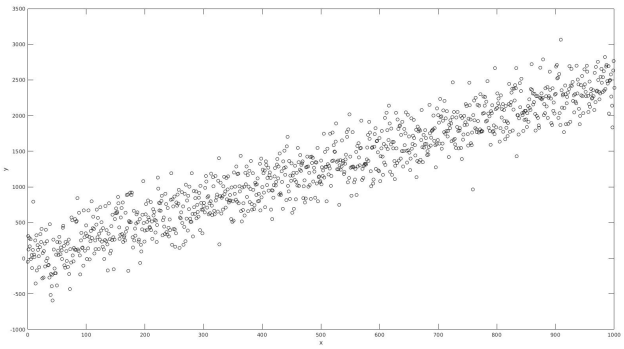
Fondements du ML

29/10/2024

Aujourd'hui : Régression linéaire (1/3)
TD (Fin modèle linéaire + Début régression)

Projet : A venir (en binôme, sur de la classification "linéaire")

RÉGRESSION LINÉAIRE ET LIEN AVEC LES MOINDRES CARRÉS



$\leftarrow y_i = x_i^T \beta^* + \varepsilon_i$ avec $\varepsilon_i \sim \mathcal{N}(0,1)$

Contexte : $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times m}$, $y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m$

avec $y_i = x_i^T \beta^* + \varepsilon_i$

β^* : Vérité terrain ("ground truth")

ε_i : bruit (variable aléatoire)

But : Trouver un modèle linéaire $x \longmapsto x^T \beta$ tel que

$y_i \simeq x_i^T \beta$

→ On a vu comment calculer une solution de ce problème au sens des moindres carrés

→ Mais en général, β^* n'est pas une solution de ce problème

Q) Quel est le lien entre la solution au sens des moindres carrés et β^* ?

⇒ Besoin d'outils d'estimation statistique

(1) Régression linéaire simple

$(n=1)$ $\quad X \in \mathbb{R}^{n\times 1}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$

$$y_i = \underbrace{x_i}_{1\times 1}\underbrace{\beta^*}_{1\times 1} + \varepsilon_i \qquad \varepsilon_i \text{ variable aléatoire}$$

Approche par moindres carrés linéaires

On cherche β solution de $\min_{\beta \in \mathbb{R}} \frac{1}{2}\|X\beta - y\|^2 = \frac{1}{2}\sum_{i=1}^{n}(x_i\beta - y_i)^2$

→ Le vecteur $X^\dagger y$ est toujours une solution, et c'est même la solution de norme minimale. On l'appelle l'estimateur des moindres carrés ordinaires, et on le note $\hat{\beta}^{OLS} = X^\dagger y$

(OLS = Ordinary Least Squares)

→ Dans le cas qui nous intéresse $(n=1)$, on a 2 possibilités:

- Soit $\text{rang}(X) = 0$: alors $X = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{n\times 1}$, $X^\dagger = [0 \ \cdots \ 0] \in \mathbb{R}^{1\times n}$

et donc $\hat{\beta}^{OLS} = 0$

- Soit $\text{rang}(X) = 1 \leq \min(\overbrace{1}^{n}, n)$, alors on sait que

$$\hat{\beta}^{OLS} = X^\dagger y = \underbrace{(X^T X)^{-1}}_{1\times n \times n\times 1 \ = \ 1\times 1} \underbrace{X^T y}_{1\times n \ n\times 1 \ = \ 1\times 1} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

⊕ Formule explicite

⊖ A priori indépendante de la distribution du bruit $\{\varepsilon_i\}_{i=1\dots n}$

Approche par maximum de vraisemblance

Hypothèse : $y_i = x_i \beta^* + \varepsilon_i$ où $\varepsilon_1, \dots, \varepsilon_m$ sont des variables aléatoires iid suivant une loi gaussienne $\mathcal{N}(0,1)$ (moyenne 0, variance 1)

(Valable aussi avec $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$)

x_i fixés/connus

$\varepsilon_i \sim \mathcal{N}(0,1) \Rightarrow y_i \sim \mathcal{N}(x_i \beta^*, 1)$

→ Calculer un modèle linéaire, c'est trouver un estimateur de β^*

→ Principe du maximum de vraisemblance : Trouver la valeur la plus probable étant données les observations $y_1, \dots, y_m$

Def : L'estimateur du maximum de vraisemblance (sous l'hypothèse ci-dessus) est défini comme la/une solution, notée $\hat{\beta}^{MLE}$, du problème d'optimisation

MLE : Maximum Likelihood Estimator

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

Vraisemblance : loi jointe de $y_1 - x_1\beta, \dots, y_m - x_m\beta$

Interprétation

Si β définit un bon modèle, alors $y_i - x_i^T \beta$ suit une loi gaussienne $\mathcal{N}(0,1)$ (ou $y_i \sim \mathcal{N}(x_i \beta, 1)$)

Vraisemblance : $\mathcal{L}(y_1, \dots, y_m | \beta) = \mathbb{P}\left(Y_1 = y_1, \dots, Y_m = y_m \;\middle|\; Y_i - x_i\beta \sim \mathcal{N}(0,1) \;\forall i \right)$

↳ On aurait pu vouloir maximiser $\mathbb{P}(\beta = \beta \mid y_i - x_i \beta \sim \mathcal{N}(0,1)\ \forall i)$

mais cette probabilité n'est pas calculable de manière explicite,

alors on utilise $\mathbb{P}(\beta \mid Y_1, \dots, Y_m) \propto \mathbb{P}(Y_1, \dots, Y_m \mid \beta)$

(formule de Bayes)

↳ Formule de la vraisemblance dans notre cas :

$$\mathcal{L}(y_1, \dots, y_m \mid \beta) = \mathbb{P}(Y_1 = y_1, \dots, Y_m = y_m \mid Y_i - x_i \beta \sim \mathcal{N}(0, \sigma_i^2))$$

$y_i - x_i^T \beta = \varepsilon_i$ iid $\rightarrow (=) \prod_{i=1}^{m} \mathbb{P}(Y_i = y_i \mid Y_i - x_i \beta \sim \mathcal{N}(0,1))$

$$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i - x_i \beta)^2\right)$$

$\varepsilon \sim \mathcal{N}(0,1)$ $\quad \mathbb{P}(\varepsilon = t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^m \exp\left(-\frac{1}{2} \sum_{i=1}^{m} (y_i - x_i \beta)^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^m \exp\left(-\frac{1}{2} \|X\beta - y\|^2\right)$$

Théorème : Les problèmes $\min_{\beta \in \mathbb{R}} \frac{1}{2} \|X\beta - y\|^2$

et $\max_{\beta \in \mathbb{R}} \mathcal{L}(y_1, \dots, y_m \mid \beta)$

ont le même ensemble de solutions.

⚠ Sur l'hypothèse $y_i - x_i\beta \sim \mathcal{N}(0,1)$

Remarque : $\max_{\beta \in \mathbb{R}} \mathcal{L}(y_1, \dots, y_m | \beta)$ et $\max_{\beta \in \mathbb{R}} \ln\left(\mathcal{L}(y_1, \dots, y_m | \beta)\right)$

ont le même ensemble de solutions

$$\ln\left(\mathcal{L}(y_1, \dots, y_m | \beta)\right) = -\frac{1}{2}\|X\beta - y\|^2 + \underbrace{\ln\left(\left(\frac{1}{\sqrt{2\pi}}\right)^m\right)}_{\text{Constante}}$$

$\max_{\beta \in \mathbb{R}} \ln\left(\mathcal{L}(y_1, \dots, y_m | \beta)\right)$ revient à maximiser $\beta \longmapsto -\frac{1}{2}\|X\beta - y\|^2$

$\max_{\beta \in \mathbb{R}} -\frac{1}{2}\|X\beta - y\|^2$ et $\min_{\beta \in \mathbb{R}} \frac{1}{2}\|X\beta - y\|^2$ ont le même ensemble de solutions

↳ Dans notre cas (hypothèse ε_i iid $\sim \mathcal{N}(0,1)$), on peut donc poser

$$\hat{\beta}^{MLE} = \hat{\beta}^{OLS} = X^\dagger y \in \mathbb{R} \quad (n = 1)$$

Propriétés statistiques de $\hat{\beta}^{MLE}$

- $\hat{\beta}^{MLE}$ est un estimateur sans biais : $\mathbb{E}_{\varepsilon_1, \dots, \varepsilon_m}\left[\hat{\beta}^{MLE}\right] = \beta^*$

(si $X = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$, on considère que $\beta^* = 0$)

Si $\text{rang}(X) = 1$, $\hat{\beta}^{MLE} = \dfrac{\sum_{i=1}^m x_i y_i}{\sum_{i=1}^m x_i^2} = \dfrac{\sum_{i=1}^m x_i (x_i \beta^* + \varepsilon_i)}{\sum_{i=1}^m x_i^2}$

et donc $\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_m}\left[\hat{\beta}^{MLE}\right] = \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_m}\left[\frac{\sum_{i=1}^{m}(x_i^2\beta^* + x_i\varepsilon_i)}{\sum_{i=1}^{m}x_i^2}\right]$

linéarité de l'espérance $\hookrightarrow = \frac{1}{\sum_{i=1}^{m}x_i^2}\sum_{i=1}^{m}\left(x_i^2\beta^* + x_i\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_m}[\varepsilon_i]\right)$

ε_i iid $\sim N(0,1)$ $\hookrightarrow = \frac{1}{\sum_{i=1}^{m}x_i^2}\sum_{i=1}^{m}(x_i^2\beta^* + 0)$

$= \frac{1}{\sum_{i=1}^{m}x_i^2}\sum_{i=1}^{m}x_i^2\beta^* = \beta^*$

- $\hat{\beta}^{MLE}$ est un estimateur convergent

$$\mathrm{Var}_{\varepsilon_1,\dots,\varepsilon_m}\left[\hat{\beta}^{MLE}\right] = \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_m}\left[\left(\hat{\beta}^{MLE} - \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_m}\left[\hat{\beta}^{MLE}\right]\right)^2\right]$$

$$= \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_m}\left[\left(\hat{\beta}^{MLE} - \beta^*\right)^2\right]$$

$\hat{\beta}^{MLE}$ convergent $\iff \mathrm{Var}_{\varepsilon_1,\dots,\varepsilon_m}\left[\hat{\beta}^{MLE}\right] \xrightarrow[m\to\infty]{} 0$

Lorsque $\mathrm{rang}(X) = 1$, on a

$$\mathrm{Var}_{\varepsilon_1,\dots,\varepsilon_m}\left[\hat{\beta}^{MLE}\right] = \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_m}\left[\left(\frac{\sum_{i=1}^{m}x_iy_i}{\sum_{i=1}^{m}x_i^2} - \beta^*\right)^2\right]$$

$$= \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_m}\left[\left(\frac{\sum_{i=1}^{m}x_i^2\beta^* + x_i\varepsilon_i}{\sum_{i=1}^{m}x_i^2} - \frac{\sum_{i=1}^{m}x_i^2\beta^*}{\sum_{i=1}^{m}x_i^2}\right)^2\right]$$

$$= \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_m}\left[\left(\frac{\sum_{i=1}^m x_i \varepsilon_i}{\sum_{i=1}^m x_i^2}\right)^2\right]$$

$$= \frac{1}{\left(\sum_{i=1}^m x_i^2\right)^2}\,\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_m}\left[\left(\sum_{i=1}^m x_i\varepsilon_i\right)^2\right]$$

$\varepsilon_i \sim \mathcal{N}(0,1)$ iid

$\mathbb{E}\left[\sum_{i=1}^m x_i \varepsilon_i\right] = 0$

$\hookrightarrow$

$$= \frac{1}{\left(\sum_{i=1}^m x_i^2\right)^2}\,\mathrm{Var}_{\varepsilon_1,\dots,\varepsilon_m}\left(\sum_{i=1}^m x_i\varepsilon_i\right)$$

$\mathrm{Var}[X] = \mathbb{E}\left[(X-\mathbb{E}[X])^2\right]$

$\hookrightarrow$

$$= \frac{1}{\left(\sum_{i=1}^m x_i^2\right)^2}\sum_{i=1}^m x_i^2\,\mathrm{Var}_{\varepsilon_i}[\varepsilon_i]$$

ε_i iid

$\mathrm{Var}_{\varepsilon_1,\dots,\varepsilon_m}\left[\sum_{i=1}^m x_i\varepsilon_i\right] = \sum_{i=1}^m x_i^2\,\mathrm{Var}_{\varepsilon_i}[\varepsilon_i]$

$$= \frac{\sum_{i=1}^m x_i^2}{\left(\sum_{i=1}^m x_i^2\right)^2} = \frac{1}{\sum_{i=1}^m x_i^2} \longrightarrow 0 \quad m\to\infty$$

Si on mesure en une infinité de $x_i \neq 0$

Bilan : Dans le cas $n=1$ et sous l'hypothèse de bruit gaussien, l'approche par maximum de vraisemblance est équivalente à l'approche par moindres carrés, dans le sens où elle conduit au même estimateur.

En connaissant la loi du bruit, on peut en plus donner des garanties statistiques pour cet estimateur (non biaisé $\mathbb{E}[\hat{\beta}^{MLE}] = \beta^*$, convergent $\mathrm{Var}[\hat{\beta}^{MLE}] \to 0$ $m\to\infty$)

(2) Régression linéaire multiple

$n \geq 1$ $\qquad y_i = \underbrace{x_i^T \beta^*}_{\in \mathbb{R}} + \varepsilon_i$, ε_i iid $\sim \mathcal{N}(0,1)$

Les résultats précédents se généralisent à ce cas :

- Equivalence entre la maximisation de la vraisemblance et les moindres carrés $\left(\hat{\beta}^{MLE} = \hat{\beta}^{OLS} = X^\dagger y \in \mathbb{R}^n \right)$

- Estimateur sans biais

$$\mathbb{E}_{\varepsilon_1, \dots, \varepsilon_m} \left[\hat{\beta}^{MLE} \right] = \beta^* \in \mathbb{R}^n$$

$v \in \mathbb{R}^n$ aléatoire

$$\mathbb{E}[v] = \begin{bmatrix} \mathbb{E}[v_1] \\ \vdots \\ \mathbb{E}[v_n] \end{bmatrix}$$

$$Cov[v] = \mathbb{E}\left[\overbrace{(v - \mathbb{E}[v])}^{n \times 1} \overbrace{(v - \mathbb{E}[v])^T}^{1 \times n} \right] = \left[\mathbb{E}\left[(v_i - \mathbb{E}[v_i])(v_j - \mathbb{E}[v_j]) \right] \right]_{ij}$$

- Estimateur convergent

$$\text{trace}\left(Cov\left[\hat{\beta}^{MLE} \right] \right) \xrightarrow[n \to \infty]{} 0$$

Pour la suite

→ ε_i suit une loi non gaussienne ?

→ A priori sur β^* ?