FONDEMENTS DU ML

15/10/2024

Programme

CM Modèle linéaire (2/2)

TD Fin ACP + Début modèle linéaire

PREMIERS PAS AVEC LE MODÈLE LINÉAIRE (suite)

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

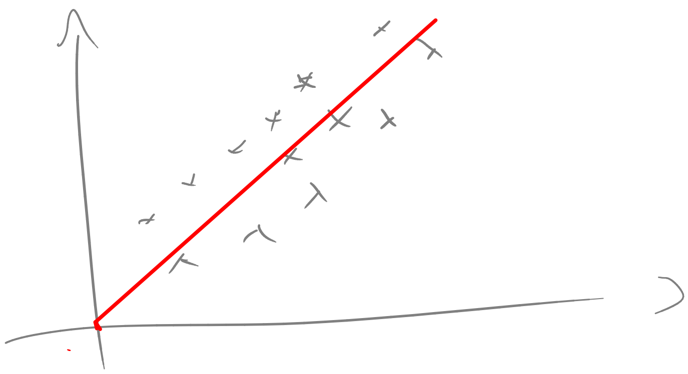
On cherche $\beta \in \mathbb{R}^n$ tel que

$$x_i^T \beta = y_i \quad \forall i = 1..m$$

On a vu que la pseudo-inverse de X donnait une solution de ce système quand il en existe une.

$\Rightarrow$ Que dire lorsque le système n'admet pas de solution ?

$\Rightarrow$ Est-ce que la solution donnée par la pseudo-inverse est la meilleure solution lorsqu'il y en a une infinité ?



Approche (par moindres carrés)

Trouver le meilleur modèle linéaire possible relativement aux données (X, y)

Mathématiquement, cette approche se modélise en utilisant un problème d'optimisation

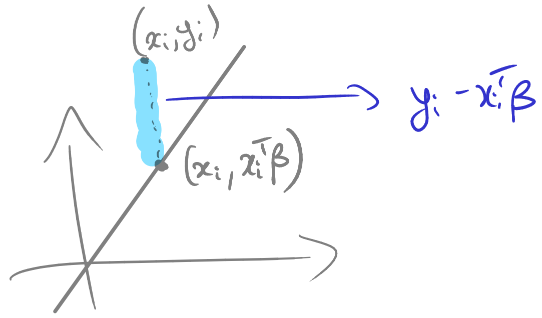
$\Rightarrow$ On va chercher à minimiser l'écart entre $x_i^T\beta$ (valeur donnée par le modèle linéaire) et y_i (vrai label que l'on cherche à prédire)

Si on peut interpoler les données, c'est-à-dire s'il existe β^* tel que $x_i^T\beta^* = y_i \ \forall i = 1..m$, alors ce β^* minimise les écarts.

$\Rightarrow$ On mesure ici l'écart via une perte aux moindres carrés

$$\frac{1}{2}\left(x_i^T\beta - y_i\right)^2 \qquad \left(\tfrac{1}{2} : \text{convention de normalisation}\right)$$

Graphiquement



$\hookrightarrow$ On mesure l'écart du modèle $x \mapsto x^T\beta$ sur le jeu de données $\{(x_i, y_i)\}_{i=1..m}$ par

$$\frac{1}{2}\sum_{i=1}^{m}\left(x_i^T\beta - y_i\right)^2 = \frac{1}{2}\|\overbrace{X\beta}^{\begin{bmatrix} x_1^T\beta \\ \vdots \\ x_m^T\beta \end{bmatrix}} - y\|^2$$

$\longrightarrow$ On va chercher le vecteur β qui minimise la valeur $\frac{1}{2}\|X\beta - y\|^2$, ce que l'on écrit comme un problème d'optimisation

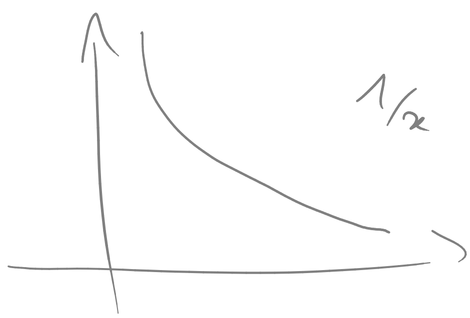
"minimiser" $\rightarrow$

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

"par rapport à β"

$\rightarrow$ "fonction objectif du problème"

$\beta \mapsto \frac{1}{2}\|X\beta - y\|^2$



Remarques

- $\forall a > 0, \forall b \in \mathbb{R}$,

$\min_{\beta \in \mathbb{R}^n} \frac{1}{2}\|X\beta - y\|^2$ et $\min_{\beta \in \mathbb{R}^n}\left(\frac{a}{2}\|X\beta - y\|^2 + b\right)$ ont le même ensemble de solutions (mais la valeur de la fonction objectif en la solution sera différente)

• On notera $\underset{\beta \in \mathbb{R}^m}{\operatorname{argmin}} \frac{1}{2}\|X\beta - y\|^2$ $\in \mathbb{R}^n$ l'ensemble des solutions du problème

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

(argmin ≡ argument minimal)

Définition : Etant donnés (X, y) et $\hat{\beta} \in \mathbb{R}^m$, on dit que :

• $\hat{\beta}$ est une solution au sens des moindres carrés si

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^m}{\operatorname{argmin}} \frac{1}{2}\|X\beta - y\|^2$$

(solution du problème d'optimisation)

• $\hat{\beta}$ est une solution au sens classique si :

$$X\hat{\beta} = y$$

(solution du système linéaire)

Remarque : → Toute solution au sens classique est une solution au sens des moindres carrés

Si $\hat{\beta}$ est une solution au sens classique, alors $X\hat{\beta} - y = 0_{\mathbb{R}^n}$

$$\frac{1}{2}\|X\hat{\beta} - y\|^2 = \frac{1}{2}\|0_{\mathbb{R}^n}\|^2 = 0$$

or $\forall \beta \in \mathbb{R}^m, \; \frac{1}{2}\|X\beta - y\|^2 \geq 0$

$$\forall \beta \in \mathbb{R}^m, \; \frac{1}{2}\|X\beta - y\|^2 \geq \frac{1}{2}\|X\hat{\beta} - y\|^2$$

$$\Leftrightarrow \hat{\beta} \in \underset{\beta \in \mathbb{R}^m}{\operatorname{argmin}} \frac{1}{2}\|X\beta - y\|^2$$

→ S'il existe une solution au sens classique, alors l'ensemble des solutions au sens classique coïncide avec l'ensemble des solutions au sens des moindres carrés

→ En revanche, il peut y avoir des solutions au sens des moindres carrés et pas de solution au sens classique

$m=2$

$x_1^T=[1]$

$x_2^T=[1]$

$y_1=0$

$y_2=1$

Ex) $X=\begin{bmatrix}1\\1\end{bmatrix}\in\mathbb{R}^{2\times 1}$ $\quad y=\begin{bmatrix}0\\1\end{bmatrix}\in\mathbb{R}^2$

$$X\beta=y \iff \begin{cases}\beta=0 \\ \beta=1\end{cases} \quad x_1^T\beta=y_1$$

pas de solution au sens classique

$$\min_{\beta\in\mathbb{R}} \frac{1}{2}\|X\beta-y\|^2=\frac{1}{2}\Big[(\underbrace{\beta-0}_{x_1^T\beta-y_1})^2+(\underbrace{\beta-1}_{x_2^T\beta-y_2})^2\Big]=\frac{1}{2}\left[\beta^2+(\beta-1)^2\right]$$

a pour solution $\hat\beta=\frac{1}{2}$

$$\frac{1}{2}\left[\hat\beta^2+(\hat\beta-1)^2\right]=\frac{1}{2}\left(\frac{1}{4}+\frac{1}{4}\right)=\frac{1}{4}$$

Théorème (Propriétés du problème d'optimisation "aux moindres carrés linéaires")
linear least squares

Soient $X\in\mathbb{R}^{m\times n}$ et $y\in\mathbb{R}^m$.

1) Le problème aux moindres carrés $\min_{\beta\in\mathbb{R}^n}\frac{1}{2}\|X\beta-y\|^2$ possède toujours au moins une solution.

2) Le vecteur $\hat\beta=X^\dagger y$ est toujours une solution du problème.

3) Parmi toutes les solutions possibles, $\hat\beta=X^\dagger y$ est la solution de norme minimale, c'est-à-dire

$$\forall\, \gamma\in\operatorname*{argmin}_{\beta\in\mathbb{R}^n}\frac{1}{2}\|X\beta-y\|^2,\quad \|\gamma\|\geq\|X^\dagger y\|$$

Démo : Cf semestre 2

Interprétation :

1) ⟹ Problème "bien posé" (contrairement à l'écriture en système linéaire)

$X\beta = y$: problème d'interpolation

$\min_{\beta \in \mathbb{R}^m} \frac{1}{2}\|X\beta - y\|^2$: problème de régression

2) On sait calculer une solution

3) La solution X^+y a de bonnes propriétés

Norme minimale ⟹ Simplicité du modèle
(+ variance réduite par rapport aux données)

Résolution du problème aux moindres carrés linéaires

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

On note $\hat{\beta} = X^+y$

9 cas à considérer

1) $n = m$

1-a) $\mathrm{rang}(X) = n = m$

→ $\hat{\beta}$ est l'unique solution au sens des moindres carrés et l'unique solution au sens classique

→ $\hat{\beta} = X^{-1}y$

1-b) $\mathrm{rang}(X) < m$ et $y \in \mathrm{Im}(X)$

→ $\hat{\beta}$ est solution au sens classique

→ $\hat{\beta}$ est solution au sens des moindres carrés de norme minimale

→ Il existe une infinité de solutions | au sens classique | au sens des moindres carrés

En effet, $\forall v \in \ker(X)$, $X(\hat{\beta}+v) = \underbrace{X\hat{\beta}}_{=y} + \underbrace{Xv}_{=0} = y$

donc $\hat{\beta}+v \in \operatorname{argmin}_{\beta \in \mathbb{R}^n} \frac{1}{2}\|X\beta - y\|^2$

1-c) $\operatorname{rang}(X) < m$, $y \notin \operatorname{Im}(X)$

→ Pas de solution au sens classique

→ Une infinité de solutions au sens des moindres carrés

et $\hat{\beta}$ est une solution de norme minimale au sens des moindres carrés

2) $m < n$ (cas sous-déterminé)

2-a) $\operatorname{rang}(X) = m$

→ $\hat{\beta}$ s'écrit $X^T(XX^T)^{-1}y$: C'est une solution au sens classique, et c'est la solution de norme minimale au sens des moindres carrés.

→ Il y a une infinité de solutions ($\ker(X) \neq \{0\}$)

2-b) $\operatorname{rang}(X) < m$, $y \in \operatorname{Im}(X)$

→ $\hat{\beta}$ solution au sens classique et solution au sens des moindres carrés de norme minimale

→ Infinité de solutions (dans les deux sens)

2-c) $\operatorname{rang}(X) < m$, $y \notin \operatorname{Im}(X)$

→ Pas de solution au sens classique

→ Infinité de solutions au sens des moindres carrés. Parmi celles-ci, $\hat{\beta}$ est la solution de norme minimale

3) $m > n$ (cas sur-déterminé)

3-a) $\mathrm{rang}(X) = n$

→ $\hat{\beta} = (X^T X)^{-1} X^T y$ est l'unique solution au sens des moindres carrés

→ $\hat{\beta}$ est une solution au sens classique si $y \in \mathrm{Im}(X)$, auquel cas c'est l'unique solution, sinon il n'y a pas de solution au sens classique

3.b) $\mathrm{rang}(X) < n$, $y \in \mathrm{Im}(X)$

→ $\hat{\beta}$ solution au sens classique et solution au sens des moindres carrés de norme minimale

→ Infinité de solutions

3-c) $\mathrm{rang}(X) < n$, $y \notin \mathrm{Im}(X)$

→ Pas de solution au sens classique

→ Infinité de solutions au sens des moindres carrés, dont $\hat{\beta}$ qui est la solution de norme minimale

---

BILAN Face à un problème de modèle linéaire :

- Pas toujours possible d'interpoler (dépend de si $y \in \mathrm{Im}(X)$)
- La formulation par moindres carrés permet toujours de calculer le meilleur modèle linéaire possible (il y en a une infinité lorsque $\mathrm{rang}(X) < \min(m, n)$)
- Ce(s) meilleur(s) modèle(s) permettent d'interpoler lorsque cela est possible.

- Le modèle donné par la pseudo-inverse ($X^+ y$, dont l'expression se simplifie lorsque $\text{rang}(X) = \min(m, n)$) est toujours solution aux moindres carrés et est une solution remarquable (simplicité induite par la norme minimale)