Régression linéaire

Leçons 161,162,219

Théorème (Régression linéaire)

Soit $m \in \mathbb{N}^*$ (en pratique $m >> 1$). Si $(U, V) \in \mathbb{R}^m \times \mathbb{R}^m$ sont deux séries de données connues, alors on peut trouver explicitement $(x_0, x_1) \in \mathbb{R}^2$ minimisant la quantité:

$$\left\| x_0 \begin{bmatrix} 1 \\ | \\ 1 \end{bmatrix} + x_1 U - V \right\|$$

sous réserve que $\sigma(U) > 0$ (écart-type de U).

Remarque. $\|\cdot\|$ *désigne la norme euclidienne sur* $\mathbb{R}^m$. *De plus, on a:*

$$\sigma(U) = 0 \quad \Leftrightarrow \quad U \in Vect \left\{ \begin{bmatrix} 1 \\ | \\ 1 \end{bmatrix} \right\}$$

Voici le plan de la démonstration:

1. Etablir l'équation normale et donner une condition suffisant d'unicité de la solution dans une proposition
2. Résoudre le système linéaire obtenu

Avant de se lancer dans la démonstration, introduisons quelques notations:

Définition (Moyenne et produit terme à terme)

1. On définit la moyenne de $U \in \mathbb{R}^m$ par:
$$\overline{U} = \frac{1}{m} \sum_{i=1}^{m} u_i$$
2. On définit le produit composante par composante de $U, V \in \mathbb{R}^m$, noté $U \cdot V \in \mathbb{R}^m$ en posant: $\forall i \in [\![1, m]\!]$, $(U \cdot V)_i = u_i v_i$.

Remarque. *On a ainsi* $\sigma(U) = \sqrt{\overline{U \cdot U} - \overline{U}^2}$

Proposition (Distance euclidienne et équation normale)

1. Soit E un espace euclidien (réel) et soit F un sous-espace vectoriel de E. Soit p_F la projection orthogonale sur F. Soit $x \in E$. On a alors:
$$d(x, F) = \underset{f \in F}{Inf} \|x - f\| = \|x - p_F(x)\|$$
$p_f(x)$ est l'unique point de F réalisant ce minimum (il est bien atteint !)
2. Si $E = \mathbb{R}^m$, $V \in \mathbb{R}^m$, $A \in \mathcal{M}_{m,n}(\mathbb{R})$. Si $X \in \mathbb{R}^n$ vérifie:

$$\|AX - V\| = \underset{Z\in\mathbb{R}^n}{Min}\|AZ - V\|$$

alors X vérifie l'équation normale: $A^TAX = A^TV$. De plus, X est unique si et seulement si $rg(A) = n$

Démonstration. *1. Soient $x \in E$ et $f \in F$. On a:*

$$x - f = \underbrace{x - p_F(x)}_{\in\ker(p_F)=F^\perp} + \underbrace{p_F(x) - f}_{\in Im(p_F)=F}$$

donc, par le théorème de Pythagore, on a:

$$\|x - f\| = \sqrt{\|x - p_F(x)\|^2 + \|p_F(x) - f\|^2}$$

Ainsi, on a: $\|x - f\| \geqslant \|x - p_F(x)\|$, et si $f \in F\setminus\{p_F(x)\}$, alors on a $\|x - f\| > \|x - p_F(x)\|$. Donc $\underset{f\in F}{Inf}\|x - f\|$ est atteint uniquement en $f = p_F(x)$ et vaut donc $\|x - p_F(x)\|$

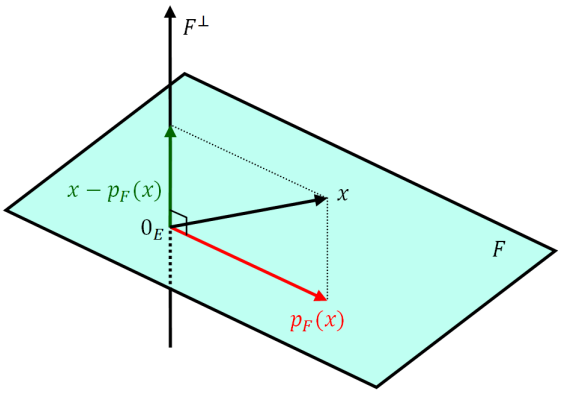

Figure 1: Illustration de la notion de distance euclidienne

2. Ici, on se place dans le cas $F = Im(A)$. Soit $X \in \mathbb{R}^n$ vérifiant:

$$\|AX - V\| = \underset{Z\in\mathbb{R}^n}{Min}\|AZ - V\|$$

Alors X vérifie $\|AX - V\| = \underset{Y\in F}{Min}\|Y - V\|$. Donc AX est unique. Si $rg(A) = n$, alors $X \mapsto AX$ est injective et X est unique. Comme on a $AX - V \in F^\perp$ on a:

$$\forall W \in \mathbb{R}^n, \langle AX - V|AW\rangle = 0 \Leftrightarrow \forall W \in \mathbb{R}^n, \langle A^TAX - A^TV|W\rangle = 0$$

Donc $A^TAX = A^TV$, ce qui est notre équation normale, où $A^TA \in \mathcal{M}_n(\mathbb{R})$

■

Passons à la démonstration du théorème à proprement parler:

Démonstration. *Dans notre cas, on a:*

$$A = \begin{bmatrix} 1 & u_1 \\ | & \vdots \\ 1 & u_m \end{bmatrix}, \quad X = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}, \quad V = \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}$$

On a donc $n = 2$*, de plus, on a:*

$$rg(A) = 2 \quad \Leftrightarrow \quad \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \notin Vect\left\{ \begin{bmatrix} 1 \\ | \\ 1 \end{bmatrix} \right\} \tag{1}$$

Donc, pour trouver le X *"optimal", on doit résoudre l'équation normale* $A^TAX = A^TV$*. Or, on a:*

$$\begin{aligned} A^TA &= \begin{bmatrix} 1 & - & 1 \\ u_1 & \dots & u_m \end{bmatrix} \begin{bmatrix} 1 & u_1 \\ | & \vdots \\ 1 & u_m \end{bmatrix} \\ &= \begin{bmatrix} m & \sum_{i=1}^m u_i \\ \sum_{i=1}^m u_i & \sum_{i=1}^m u_i^2 \end{bmatrix} \end{aligned}$$

et A^TA *est inversible si et seulement si* $\det(A^TA) \neq 0$*. Or, on a:*

$$\begin{aligned} \det(A^TA) &= m\sum_{i=1}^m u_i^2 - \left(\sum_{i=1}^m u_i\right)^2 \\ &= m^2\sigma(U)^2 \end{aligned}$$

On doit ainsi avoir la condition $\sigma(U) > 0$ *pour que* A^TA *soit inversible, ce qui correspond bien à la condition* (1)*. Ainsi, la formule de la comatrice assure que:*

$$A^TAX = A^TV \quad \Leftrightarrow \quad X = \frac{1}{\det(A^TA)} Com(A^TA)^T A^TV$$

Or, on a:

$$Com(A^TA) = \begin{bmatrix} \sum_{i=1}^m u_i^2 & -\sum_{i=1}^m u_i \\ -\sum_{i=1}^m u_i & m \end{bmatrix}$$

Donnant ainsi:

$$\begin{aligned} X &= \frac{1}{m^2\sigma(U)^2}\begin{bmatrix} \sum_{i=1}^m u_i^2 & -\sum_{i=1}^m u_i \\ -\sum_{i=1}^m u_i & m \end{bmatrix}\begin{bmatrix} 1 & - & 1 \\ u_1 & \dots & u_m \end{bmatrix}\begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} \\ &= \frac{1}{m^2\sigma(U)^2}\begin{bmatrix} \sum_{i=1}^m u_i^2 & -\sum_{i=1}^m u_i \\ -\sum_{i=1}^m u_i & m \end{bmatrix}\begin{bmatrix} \sum_{i=1}^m v_i \\ \sum_{i=1}^m u_i v_i \end{bmatrix} \\ &= \frac{1}{\sigma(U)^2}\begin{bmatrix} \overline{U\cdot U}\cdot\overline{V} - \overline{U\cdot V}\cdot\overline{U} \\ \overline{U\cdot V} - \overline{U}\cdot\overline{V} \end{bmatrix} \\ &= \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} \end{aligned}$$

■

On peut donc calculer x_0 (ordonnée à l'origine) et x_1 (coefficient directeur) uniquement en fonction de ces quatre paramètres:

1. Moyenne des données de U
2. Moyenne des données de V
3. Moyenne des données du produit $U \cdot V$ (composantes par composantes)
4. Ecart-type de U