

Travaux dirigés sur ordinateur n°6 : Méthode des moindres carrés, régression linéaire (↔ Q 1)

Question 1 : Régression linéaire

On commence par quelques rappels théoriques, étant donnés n points du plan de coordonnées (x_i, y_i) , on cherche un couple de réel (m, p) qui minimise la quantité

$$Q(m, p) = \sum_i \left(y_i - (mx_i + p) \right)^2$$

On peut interpréter cela de la façon suivante si $\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$, $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ et $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, alors

$Q(m, p) = \|Y - (mX + p\mathbf{1})\|^2$. Cette quantité est minimum lorsque $mX + p\mathbf{1}$ est le projeté orthogonal de Y sur le plan vectoriel $H = \text{vect}(\mathbf{1}, X)$ de $\mathbb{R}^n$. Les relations qui caractérisent ce projeté orthogonal sur H sont

$$\begin{cases} Y - (mX + p\mathbf{1}) \perp \mathbf{1} \\ Y - (mX + p\mathbf{1}) \perp X \end{cases}$$

C'est à dire que le vecteur $Y - P_H(Y)$ est orthogonal au plan H .

Matriciellement cela s'écrit ${}^t\mathbf{1} \cdot (Y - (mX + p\mathbf{1})) = 0$ et ${}^tX \cdot (Y - (mX + p\mathbf{1})) = 0$ ce que l'on peut encore

écrire $\begin{pmatrix} {}^t\mathbf{1} \\ {}^tX \end{pmatrix} (Y - (mX + p\mathbf{1})) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ ou encore $\begin{pmatrix} {}^t\mathbf{1} \\ {}^tX \end{pmatrix} Y = \begin{pmatrix} {}^t\mathbf{1} \\ {}^tX \end{pmatrix} \begin{pmatrix} \mathbf{1} & X \end{pmatrix} \begin{pmatrix} p \\ m \end{pmatrix}$. La matrice $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}$ a n lignes et deux colonnes, la première colonne ne contient que des 1 la deuxième est le vecteur X .

pour obtenir p et m il suffit donc d'inverser la matrice carrée $M = \begin{pmatrix} {}^t\mathbf{1} \\ {}^tX \end{pmatrix} \begin{pmatrix} \mathbf{1} & X \end{pmatrix} = {}^t(\mathbf{1} \ X) \begin{pmatrix} \mathbf{1} & X \end{pmatrix}$.

L'objectif de cet exercice est de retrouver une relation linéaire entre deux variables, qui est perturbé par du bruit (par exemple des erreurs de mesures qui sont importantes «mais homogènes»), vous aller prendre une famille de $2n$ points du plan qui sont de la forme $y_i = \hat{m}x_i + \hat{p} + \varepsilon_i$ où ε_i est le bruit, et vous devez retrouver $\hat{m}$ et $\hat{p}$.

1. Aller chercher dans la page scilab de la page web d'Alexandre Mizrahi les fichiers donX.txt et donY.txt et enregistrez les dans un dossier (par exemple .../Documents/scilab). Ensuite dans la console scilab sous l'onglet fichier en haut à gauche choisissez «changer le répertoire courant» et choisissez comme répertoire celui où vous venez de copier les fichiers donX.txt et donY.txt.
2. Taper le code scilab suivant :

```
X=read("donX.txt",-1,1);
Y=read("donY.txt",-1,1);
```

Vous avez maintenant dans les variables X et Y deux vecteurs colonnes, représentant les 300 abscisses et les 300 ordonnées du nuage de points que l'on étudie. Tracer le nuage de points à l'écran. Utiliser une option `style` avec une valeur négative pour que les points tracés ne soient pas reliés entre eux.
3. Construire la matrice C , à deux colonnes et 300 lignes, suivante $\begin{pmatrix} \mathbf{1} & X \end{pmatrix}$.
4. Construire la matrice M , calculer m et p .
5. Tracer la droite d'équation $y = mx + p$.
6. Êtes-vous arrivé à retrouver $\hat{m}$ et $\hat{p}$?
7. Déterminer la droite de régression de x en fonction de y .

Question 2 : Un petit jeu

On a construit une famille de points de la forme $y_i = \tilde{p} + \tilde{m}x_i + \tilde{q}x_i^\alpha + \varepsilon_i$ où ε_i représente du bruit et $\alpha \in [2; 3]$. On cherche à retrouver p, m, q et α .

1. Aller chercher la matrice des abscisses `jeuX.txt` et la matrice des ordonnées `jeuY.txt`, sur la page d'Alexandre Mizrahi.
2. Tracer le nuage de points correspondant.
3. Déterminer la droite de regression de y en fonction de x . La tracer.
4. Déterminer la parabole obtenue par la méthodes des moindres carrés. C'est à dire déterminer p, m, q tel que

$$\sum_i \left(y_i - (p + mx_i + qx_i^2) \right)^2$$

soit minimale. Ceci ressemble beaucoup à la régression linéaire puisque c'est une projection orthogonale sur un sous espace vectoriel de dimension 3. Tracer la parabole sur le nuage le points.

5. On cherche à rajouter un paramètre pour trouver une relation de la forme : $y = p + mx + qx^a$, mais il faut s'y prendre un peu différemment car ce n'est plus une projection orthogonale, c'est une projection orthogonale pour chaque a fixé : bon courage....