Régression (I) - Régression linéaire

1 En théorie

Exercice 1 (*). On se place dans le contexte de la régression linéaire simple au sens des moindres carrés. On considère donc :

- $N \geqslant 1$ données divisées en données d'entrée réelles $x = \{x_1, \dots, x_N\}$ et données de sortie réelles $y = \{y_1, \dots, y_N\}$,
- le modèle linéaire g_θ où $\theta = (\theta_0, \theta_1) \in \mathbb{R}^2$:

$$g_\theta(x_i) = \theta_0 + \theta_1 x_i,$$

- la fonction d'erreur

$$\mathscr{E}(\theta) = \sum_{i=1}^{N} (y_i - g_\theta(x_i))^2,$$

- et on suppose les x_i non égaux.

1. Calculer les dérivées partielles d'ordre 1 de $\mathscr{E}$.
2. Déterminer la hessienne de $\mathscr{E}$. On admettra qu'elle est définie positive : que peut-on en déduire ?
3. Montrer que les points critiques de $\mathscr{E}$ sont solution du système d'inconnues $\theta = (\theta_0, \theta_1)$ suivant :

$$\begin{pmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{pmatrix}$$

4. On note dans cette question $\overline{x}$ et $\overline{y}$ les moyennes respectives de x et y.
   (a) On définit la variance de x par $\mathrm{Var}(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$. Montrer qu'on a aussi $\mathrm{Var}(x) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \overline{x}^2$.
   (b) On définit la covariance de x et de y par $\mathrm{Cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$. Montrer qu'on a aussi $\mathrm{Cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{x}\,\overline{y}$.
   (c) En déduire que l'unique couple $\theta = (\theta_0, \theta_1)$ qui minimise $\mathscr{E}$ est donné par :

$$\begin{cases} \theta_1 &= \dfrac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} \\ \theta_0 &= \overline{y} - \theta_1 \overline{x} \end{cases}$$

5. Dans cette question, on modifie certaines notations et on en introduit de nouvelles :
   - on modifie $x^i \in \mathbb{R}$ en $x^i \in \mathbb{R}^2$ tel que $x^i = (1, x_1^i)$ noté plus simplement $x^i = (1, x_i)$,
   - de sorte que le modèle linéaire g_θ où $\theta = (\theta_0, \theta_1) \in \mathbb{R}^2$ s'exprime :

$$g_\theta(x^i) = {}^t\theta . x^i = \begin{pmatrix} \theta_0 & \theta_1 \end{pmatrix} \begin{pmatrix} 1 \\ x_i \end{pmatrix} = \theta_0 + \theta_1 x_i,$$

   - et on note $y = (y_1, \dots, y_N) \in \mathbb{R}^N$, $X = ({}^t x^1, \dots, {}^t x^N) = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \in \mathscr{M}_{N,2}(\mathbb{R})$, et $\|.\|^2$ la norme euclidienne sur $\mathbb{R}^N$.

(a) Montrer que le problème de régression linéaire simple au sens des moindres carrés revient à déterminer :

$$\widehat{\theta} = \arg\min_{\theta} \|y - X\theta\|^2.$$

(b) Montrer que cela revient à résoudre l'équation normale d'inconnue θ :

$${}^tXX\theta = {}^tXy.$$

(c) Montrer que l'équation normale a pour solution :

$$\theta = ({}^tXX)^{-1}\,{}^tXy,$$

et que cette solution est équivalente à celle trouvée à la question 4 (c).

2 En pratique

Exercice 2 (**). On considère les données du fichier `data.csv` pour lesquelles on souhaite effectuer une régression polynomiale de y en x au sens des moindres carrés.

1. Représenter les données et en déduire le modèle le plus adapté pour effectuer cette régression.
2. Donner la solution à ce problème de régression et représenter la courbe d'ajustement obtenue. Pour plus d'efficacité, vous pouvez utiliser directement la librairie `scikit-learn`.
3. Donner la solution à ce problème de régression en utilisant la famille des polynômes de Legendre à la place de celle des monômes. Comparer les résultats obtenus. Ces polynômes peuvent être obtenus par la méthode `eval_legendre` de `scipy.special`.

Exercice 3 (**). On considère les données du prix de logements en Californie (regroupés par groupes) [1].

group	MedInc	HouseAge	...	Longitude	Price
1	8.3252	41.	...	-122.23	4.526
2	8.3014	21.	...	-122.22	3.585
⋮	⋮	⋮	⋮	⋮	
20640	2.3886	16.	...	-121.24	0.894

Détail des variables :

- `MedInc` : revenu médian
- `HouseAge` : age médian
- ...
- `Longitude` : longitude
- `Price` : prix (en centaines de milliers de dollars)

Le but de cet exercice est de construire un modèle de régression linéaire pour expliquer la variable `Price` à partir des autres variables explicatives `MedInc`, `HouseAge`, ..., et `Longitude`.

1. Donner l'expression du modèle de régression utilisé et préciser la fonction d'erreur utilisée.
2. (a) Donner la solution au problème de régression linéaire.
   (b) Quelle valeur de la variable `Price` renvoie ce modèle si l'on donne en entrée les valeurs des variables explicatives du groupe 1 ? Commenter.
   (c) Pour quelle groupe la valeur de la variable `Price` que renvoie ce modèle est-elle la plus proche de la valeur réelle ?

[1] `https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset`