Université Joseph Fourier – Grenoble 1

Cours de Mathématiques Appliquées – MAP35G

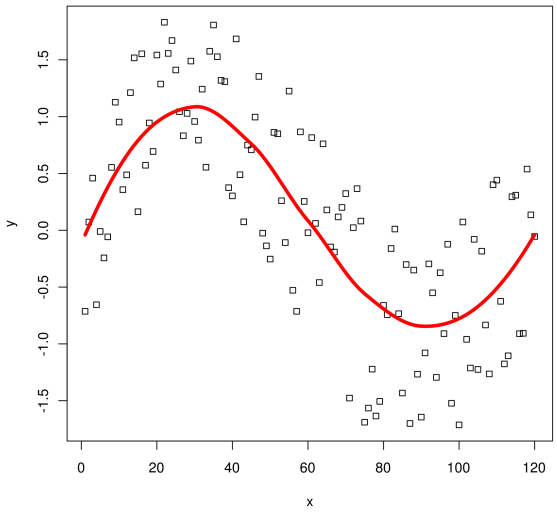
Année 2015-2016

Travaux Pratiques n° 1

Algèbre Linéaire

I Introduction à la régression

Imaginons que l'on dispose d'une collection de mesures effectuées pour des temps, lieux, ou situations différentes. Afin de construire des modèles physiques du phénomène observé, il est intéressant de trouver une *fonction* qui lie ces mesures à leur situation (temps, lieu, ...). Cette fonction est souvent prise *polynomiale* en physique car un certain nombre de lois physiques sont polynomiales.



Formellement, si l'on dispose d'une variable $y = (y_1, ..., y_M)$ de M mesures liées à une seconde variable $x = (x_1, ..., x_M)$, on dit que l'on cherche à *expliquer* y par régression par rapport à x. Dans le cadre de la *régression polynomiale* de degré N, il s'agit de trouver les $N+1$ coefficients $\alpha = (\alpha_0, ..., \alpha_N)$ du polynôme $p_\alpha(z) = \alpha_0 + \alpha_1 z + \alpha_2 z^2 + ... + \alpha_N z^N$ qui explique *au mieux* y par rapport à x.

Pour trouver ces coefficients, il s'agit de regarder pour chaque point $i = 1, ..., M$, l'erreur entre la mesure y_i et la valeur du polynôme au même point $p_\alpha(x_i)$: notons la $\varepsilon_i = p_\alpha(y_i) - y_i$.

Matriciellement, cette relation s'écrit :

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & ... & x_1^N \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_i & x_i^2 & ... & x_i^N \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_M & x_M^2 & ... & x_M^N \end{bmatrix}}_{\triangleq X} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} - \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_M \end{bmatrix}$$

c'est-à-dire

$$X\alpha - y = \varepsilon.$$

1) *Donner une condition suffisante entre M et N pour que la régression soit parfaite, c'est-à-dire que l'on puisse trouver α tel que $\varepsilon = 0$.*

II Système exactement déterminé

Supposons que $M = N + 1$, c'est-à-dire que l'on a autant de points que de degrés de liberté sur le polynôme régresseur ; et que les points de mesure sont tous distincts : $\forall i, j$, on a $i \neq j, x_i \neq x_j$.

2) *Montrer que les colonnes de X forment une famille libre. En déduire la taille de l'espace des solutions du système*

$$(\mathcal{S}) \qquad X\alpha = y.$$

Considérons le cas suivant $M = 3$, $N = 2$, $\varepsilon = 0$ (régression quadratique parfaite sur 3 points)

Jeu de données n° 1

x_i	y_i
0	1
2	5
10	8

.

3) *Écrire une fonction `x_to_X` prenant en entrée un vecteur x et un entier N et retournant la matrice X ayant pour colonnes les puissances de x de 0 à N. Que retourne la commande `rref([X y])` ?*

4) *Utiliser le résultat de la question précédente pour trouver les coefficients α du polynôme régresseur pour le jeu de données n° 1.*

5) *Retrouver ces coefficients à l'aide*
a) de l'opération `X \y`,
b) de l'opération `inv(X) * y`,
c) de la fonction `linsolve` (Utiliser l'aide `help` si nécessaire, par ex. `help linsolve`).

6) *Tracer le polynôme régresseur trouvé ainsi que les points du jeu de données. Vérifiez que la régression est parfaite. (On pourra utiliser la fonction `plot`)*

III Système sous-déterminé

Supposons que $M < N+1$, c'est-à-dire que l'on a moins de points que de degrés de liberté sur le polynôme régresseur ; et que les points de mesure sont tous distincts, $\forall i, j, i \neq j, x_i \neq x_j$.

7) *Que peut-on dire sur l'espace des solutions du système $(\mathcal{S})$?*

Considérons le cas suivant $M = 4$, $N = 5$, $\varepsilon = 0$

Jeu de données n° 2

x_i	y_i
0	1
2	5
10	8
11	4

.

8) *Que retournent chacune des méthodes (`rref([X y])` + les trois méthodes de la question 5) de la Section II ? Que retourne la fonction `null(X)` ? Trouver deux valeurs possibles pour α.*

9) *Tracer les deux polynômes régresseurs trouvés à la question précedente ainsi que les points du jeu de données. Vérifiez que la régression est parfaite pour les deux polynômes. A-t-on intérêt à être dans cette situation par rapport au cas précédent ?*

IV Système sur-déterminé

Supposons finalement que $M > N + 1$, c'est-à-dire que l'on a plus de points que de degrés de liberté sur le polynôme régresseur.

10) *Que peut-on dire sur l'espace des solutions du système $(\mathcal{S})$*

Considérons le cas suivant $M = 5$, $N = 1$, (régression linéaire sur 5 points)

Jeu de données n° 3

x_i	y_i
0	1
2	2
5	3.5
8	5.2
10	5.8

.

11) *Que retournent chacune des méthodes (`rref([X y])` + les trois méthodes de la question 5) de la Section II ?*

12) *Entrer la commande `pinv(X)*y`. Tracer le polynôme régresseur obtenu à partir de ce vecteur ainsi que les points du jeu de données. Quelle est la fonction de la commande `pinv` ? De quel système le vecteur `pinv(X)*y` est-il solution ? Conclure.*

V Méthodes itératives

Les méthodes vues précédemment peuvent se révéler coûteuses en temps de calcul lorsque la taille du problème est grande. Typiquement, l'inversion (ou le pivot de Gauss) sur une matrice qui dépasse le millier de lignes est très coûteux. Pour illustrer ce fait, on pourra comparer les temps que mettent les commandes `T = inv(rand(100));` , `T = inv(rand(1000));` , `T = inv(rand(10000));` .

Afin de pas calculer directement l'inverse de matrices de grande dimension, mais aussi de traiter plus facilement les problèmes sur-déterminés, on choisit souvent de remplacer la recherche d'une solution de $(\mathcal{S})$ par une solution du problème d'optimisation suivant :

$$(\mathcal{P}) \qquad \min_{\alpha \in \mathbb{R}^{N+1}} \|X\alpha - y\|^2 = \|\varepsilon\|^2 .$$

13) *Justifier dans les cas sous- et parfaitement déterminés (Sections II et III) que les solutions de $(\mathcal{S})$ sont les mêmes que celles de $(\mathcal{P})$.*

Nous admettrons que l'algorithme suivant est une descente de gradient convergeant vers une solution de $(\mathcal{P})$:

$$\alpha^{k+1} = \alpha^k - \frac{1}{\|X^{\mathrm{T}}X\|}\left(X^{\mathrm{T}}X\alpha^k - X^{\mathrm{T}}y\right).$$

14) *Implémenter cet algorithme sur les jeux de données n° 1 et 3. Examiner $\|\alpha^{k+1} - \alpha^k\|$ ou $\|X\alpha^k - y\|^2$ afin de choisir un nombre d'itérations à effectuer. Quelles différences remarque-t-on sur les vitesses de cet algorithme sur les deux jeux de données ? A quoi sont-elles dues ?*

15) *Tracer les polynômes régresseurs à partir du point final et de quels points intermédiaires obtenus par l'algorithme programmé dans la question précédente, ainsi que les points du jeu de données. Comparer avec le résultats précédents.*

VI Analyse de données réelles

16) *Récupérer le jeu de données n° 4 sur* `http://www.iutzeler.org/docs_teach/MAP35G/data_16.m`*. Il représente*[1] *des mesures de la capacité calorifique (en $cal.mol^{-1}.K^{-1}$) du bromure d'hydrogène solide en fonction de la température (en K). Calculer, avec la méthode de votre choix, les régressions polynomiales de degré 0,1,2 jusqu'au degré de votre choix. Représenter les polynômes régresseurs et les points du jeu de données, comparer les courbes et la norme de l'erreur $\|\varepsilon\|$. Quelle régression choisiriez-vous pour modéliser ce phénomène ?*

17) *Récupérer les jeux de données n° 5.1 et 5.2 sur* `http://www.iutzeler.org/docs_teach/MA035G/data_17_1.m` *et* `data_17_2.m`*. Ils représentent*[2] *les hauteurs respectivement de mousse et de bière lorsque l'on sert une bière dans un verre à 20 degrés Celsius. Représenter les points des deux jeux de données. Quelle type de fonction semblent suivre ces points ? Effectuer une transformation afin de vous ramener à une régression linéaire. Représenter le résultat de la régression.*

1. voir `http://www.stat.ufl.edu/ winner/data/heatcapacity.txt`
2. voir `http://www.stat.ufl.edu/ winner/data/beer_foam.txt`