Modèles Linéaires

Devoir Maison
Licence 3 MIASHS (2025 - 2026)

Guillaume Metzler
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr

Résumé

Il n'est pas demandé de réaliser l'ensemble des exercices, je vous demande simplement de faire ce que vous pouvez.
Ce devoir est composé de deux parties, une première porte sur l'étude des fonctions à plusieurs variables et la deuxième partie porte sur la régression linéaire.

Fonctions à plusieurs variables et conditionnement d'une matrice

Dans cette première partie, on va cherche à étudier une fonction de plusieurs variables à valeurs réelles.
On commencera par étudier une telle fonction et on va ensuite regarder comment optimiser cette fonction à travers un algorithme que l'on appelle **la descente** de gradient.
Ce problème s'inscrit dans un cadre général que l'on appelle, le **conditionnement d'une matrice**, qui joue un rôle important dans la résolution de problèmes de façon numérique

Contexte Soit γ un nombre réel. L'objectif de cet exercice est d'étudier la fonction $f_\gamma : \mathbb{R}^2 \to \mathbb{R}$ définie par :

$$f_\gamma(x, y) = \frac{1}{2}(x^2 + \gamma y^2 + 2xy) + 2x + 2y.$$

Etude de la fonction f_γ. Cette première partie est consacrée à l'étude de la fonction f_γ.

1. Etudier la convexité de la fonction f_γ.
2. Donner les solutions de l'*équation d'Euler*, *i.e.*, les solutions du système linéaire $\nabla f_\gamma(x, y) = (0, 0)$ pour toutes les valeurs de γ.
3. Donner la nature des extrema de la fonction f_γ en fonction de la valeur de γ.
4. Montrer que la fonction f_γ peut s'écrire sous la forme

$$f_\gamma(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T\mathbf{A}\mathbf{u} - \mathbf{b}^T\mathbf{u},$$

où $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & \gamma \end{pmatrix}$, $\mathbf{b} \in \mathbb{R}^2$ est un vecteur à déterminer et $\mathbf{u} = (x, y)^T$.

Descente de gradient à pas optimal Dans cette section uniquement, on supposera que $\gamma = 2$.

1. Que peut-on dire de la convexité de la fonction f_2 et des extrema de cette fonction ?

On cherche maintenant à trouver une procédure algorithmique qui permet d'atteindre le minimum de cette fonction en partant de n'importe quel point $\mathbf{u}_0 = (x_0, y_0)$. Une telle procédure s'appelle une *descente de gradient*. Dans les grandes lignes l'idée est de construire une suite $(\mathbf{u}_k)_{k\in\mathbb{N}}$ qui converge vers $\mathbf{u}^\star$, le minimum de notre fonction, en utilisant le principe suivant :

- choisir une valeur $\mathbf{u}_0$: c'est le point de départ de notre algorithme d'optimisation.
- $\mathbf{u}_k \to \mathbf{u}_{k+1}$: on choisit une direction $\mathbf{d}_k$ et on minimise la fonction objective f le long de cette direction.
- on résout $\arg\min_{\rho>0} f(\mathbf{u}_k - \rho\mathbf{d}_k) = \rho_k$: on cherche à quel point on doit se déplacer dans la direction donnée pour minimiser la fonction f. Cette constante ρ est souvent appelé **pas d'apprentissage**.
- $\mathbf{u}_{k+1} = \mathbf{u}_k - \rho_k\mathbf{d}_k$: on met à jour la valeur de notre suite.

Dans toute cette procédure, la direction $\mathbf{d}_k$ que l'on va choisir est $\nabla f(\mathbf{u}_k)$ et on va d'abord se concentrer sur cet algorithme que l'on appelle descente de gradient à pas constant.

Définition: Descente de gradient à pas constant

gradcst Soit f défini sur un sous-ensemble D de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}$ et soit ρ, ε des nombres réels positifs.
Alors la **descente de gradient à pas constant** est décrit par

- choisir $\mathbf{u}_0$ pour initialiser notre algorithme,
- tant que $\|\nabla f(\mathbf{u}_k)\| \geq \varepsilon$
  1. calculer $\nabla f(\mathbf{u}_k)$
  2. poser $\mathbf{u}_{k+1} = \mathbf{u}_k - \rho \nabla f(\mathbf{u}_k)$, $\rho > 0$

On peut alors montrer que pour un bon choix de ρ, la suite des valeurs de $\mathbf{u}_k$ va se rapprocher petit à petit du point qui minimise notre fonction (la solution de l'équation d'Euler).

2. On suppose que $\rho = 1$ et on pose $\mathbf{u}_0 = (0, 0)$. Calculer les valeurs de $\mathbf{u}_k$ pour $k = 1, 2$ et 3 à l'aide d'un algorithme de descente de gradient à pas constant.

Dans la suite, on va s'intéresser à une autre version dite **à pas optimale** ou **de plus profonde descente**. L'idée est identique, la procédure est la même, mais, cette fois, le pas de descente ρ est choisi de façon à minimiser $f(\mathbf{u}_k - \rho \nabla f(\mathbf{u}_k))$.

Taux de convergence de la descente de gradient à pas optimal. A partir de maintenant, on suppose que $\gamma > 1$ de telle sorte que la fonction f soit bien strictement convexe. L'objectif est d'étudier la vitesse de convergence de la descente de gradient à pas optimal. Cette vitesse de convergence dépend de ce que l'on appelle **le conditionnement de la matrice A**, notée $Cond(\mathbf{A})$, et il est défini par

$$Cond(\mathbf{A}) = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})},$$

où $\lambda_{\max}(\mathbf{A})$ et $\lambda_{\min}(\mathbf{A})$ sont respectivement la plus grande et la plus petite valeur propre de la matrice $\mathbf{A}$.

1. Déterminer les valeurs propres de la matrice $\mathbf{A}$.
2. Donner une expression du conditionnement de la matrice $\mathbf{A}$ en fonction de γ et donner un équivalent asymptotique de ce conditionnement, *i.e.* pour de grandes valeurs de γ.
   Indice : on utilisera le fait que pour de grandes valeurs de γ on a $(\gamma-1)^2+4 \simeq (\gamma-1)^2$.
3. Notons $\mathbf{u}^\star$ le point pour lequel la fonction f_γ atteint son minimum et $\mathbf{u}_0$ le point initial de notre algorithme de descente de gradient.

On définit le *taux de convergence* de notre algorithme par la nombre $\eta = 1 - Cond(\mathbf{A})^{-1}$ et on

$$\|\mathbf{u}_{k+1} - \mathbf{u}^\star\|_{\mathbf{A}} \leq \eta^k \|\mathbf{u}_0 - \mathbf{u}^\star\|_{\mathbf{A}}. \tag{1}$$

A l'aide de cette définition dire pour quelles valeurs de γ la convergence de la l'algorithme est la plus rapide. [1]

4. On cherche maintenant à démontrer l'inégalité donnée en Equation (1). On note ρ_k le pas optimal de notre algorithme à l'itération k.

   (a) Montrer que

   $$\|\mathbf{u}_{k+1} - \mathbf{u}^\star\|_{\mathbf{A}}^2 \leq \|(\mathbf{I} - \rho_k \mathbf{A})(\mathbf{u}_k - \mathbf{u}^\star)\|_{\mathbf{A}}^2.$$

   Indice : on se rappelle que si $\mathbf{u}^\star$ *est un minimum de* f_γ*, alors* $\mathbf{A}\mathbf{u}^\star = \mathbf{b}$ *où* $\mathbf{A}$ *et* $\mathbf{b}$ *ont été définis dans la précédente partie.*

   (b) Maintenant, on admet que pour tout $k \in \mathbb{N}$, nous avons :

   $$\|\mathbf{u}_{k+1} - \mathbf{u}^\star\|_{\mathbf{A}}^2 \leq \|(\mathbf{I} - \rho_k \mathbf{A})\|_2^2 \|(\mathbf{u}_k - \mathbf{u}^\star)\|_{\mathbf{A}}^2.$$

   Montrer que η^2 est une borne supérieure de $\|\mathbf{I} - \rho_k \mathbf{A}\|_2^2$, *i.e.*,

   $$\|\mathbf{I} - \rho_k \mathbf{A}\|_2^2 \leq \eta^2 = \left(1 - \frac{\lambda_{\min}(\mathbf{A})}{\lambda_{\max}(\mathbf{A})}\right)^2.$$

   (c) Conclure quant à la convergence.

Autour du modèle linéaire gaussien

On suppose que l'on dispose d'un échantillon $S = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$ où $y_i \in \mathbb{R}$ et $\mathbf{x}_i \in \mathbb{R}^p$, où $p > 1$ représente la dimension de notre jeu de données. Notre objectif est de déterminer une relation linéaire entre les valeurs observées y_i et les caractéristiques des individus $\mathbf{x}_i$. Pour cela, on considère le modèle suivant :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ est la matrice de *design*, $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ et $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ est notre vecteur des résidus ou erreurs du modèle. On suppose que les nos erreurs suivent une distribution normale de moyenne nulle et de variance inconnue σ^2.

1. Le conditionnement d'une matrice joue également un rôle important dans la stabilité des solutions numériques données par notre ordinateur lorsque la précision numérique est limitée. Pour des problèmes dits *mal conditionnés*, une faible perturbation des données peut engendrer une modification radicale de la solution, *i.e.*, une multiplication de l'erreur.

On rappelle que le vecteur $\boldsymbol{\beta}$ est solution du problème suivant :

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Variantes du modèles gaussien

Dans cette section on va regarder deux variantes du modèle linéaire gaussien : *(i)* on remet en cause l'hypothèse d'homoscédasticité et *(ii)* en supposant que les individus $\mathbf{x}_i$ n'ont pas le même poids lors de l'estimation des paramètres du modèle.

(i) Remise en cause de l'homoscédasticité On suppose que l'hypothèse $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}$ n'est plus vérifiée mais que l'on cette fois ci $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{\Sigma}$, où la matrice $\boldsymbol{\Sigma} \in \mathbb{R}n \times n$ est connue.

6. Déterminer l'estimateur obtenu par **MCO** en tenant compte de cette nouvelle hypothèse.

(ii) Pondération des individus On suppose maintenant que chque individu a un poids différent dans l'estimation des paramètres du modèle. On notera w_i la pondération de l'exemple $\mathbf{x}_i$. Notre prolbème de minimisation peut alors se réécrire

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} w_i(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2.$$

7. Déterminer l'estimateur obtenu par **MCO** en tenant compte de cette nouvelle hypothèse.