TD 1 : Introduction à l'optimisation

Optimisation pour l'apprentissage automatique, M2 Big Data

23 novembre 2021

Exercice 1 : Moindres carrés linéaires

On considère un jeu de données de la forme $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, où chaque $\boldsymbol{x}_i$ est un vecteur de $\mathbb{R}^d$ et chaque y_i appartient à $\mathbb{R}$. On cherche un modèle linéaire qui explique les données, que l'on obtient en considérant le problème :

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}}\, f(\boldsymbol{w}) := \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|^2 = \frac{1}{2}\sum_{i=1}^n (\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i)^2, \tag{1}$$

où $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ et $\boldsymbol{y} \in \mathbb{R}^n$ concatènent les données, c'est-à-dire que

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Ce problème est un des plus classiques en analyse de données; sa fonction objectif est de classe $\mathcal{C}^2$, et on peut montrer qu'il possède toujours au moins une solution.

a) Supposons que $\boldsymbol{w}^*$ vérifie $\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{y}$ (c'est donc une solution du système linéaire $\boldsymbol{X}\boldsymbol{w} = \boldsymbol{y}$). Justifier que $\boldsymbol{w}^*$ est alors un minimum global du problème.

b) Le gradient de f en $\boldsymbol{w} \in \mathbb{R}^d$ est donné par $\nabla f(\boldsymbol{w}) = \boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$. Si $\boldsymbol{w}^*$ est un minimum local de f, que vaut $\nabla f(\boldsymbol{w}^*)$?

c) La matrice hessienne de f en $\boldsymbol{w} \in \mathbb{R}^d$ est donnée par $\nabla^2 f(\boldsymbol{w}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$. Elle est donc constante et définie par les données du problème.

   i) On a toujours $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \succeq \boldsymbol{0}$. Quelle propriété sur f cela implique-t-il ?

   ii) On suppose que $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \succeq \mu \boldsymbol{I}_d$ avec $\mu > 0$. Dans ce cas, que peut-on dire de $\nabla^2 f(\boldsymbol{w})$ pour tout $\boldsymbol{w}$? Qu'en déduit-on sur l'ensemble des solutions du problème (1) ?

Exercice 2 : Fonction convexe

Soit la fonction $q : \mathbb{R}^d \to \mathbb{R}$ définie par $q(\boldsymbol{w}) = \frac{1}{4}\|\boldsymbol{w}\|^4$. Cette fonction est de classe $\mathcal{C}^2$, et pour tout $\boldsymbol{w} \in \mathbb{R}^d$, on a :

$$\nabla q(\boldsymbol{w}) = \|\boldsymbol{w}\|^2 \boldsymbol{w}, \qquad \nabla^2 q(\boldsymbol{w}) = 2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}} + \|\boldsymbol{w}\|^2 \boldsymbol{I}_d.$$

a) En utilisant sa matrice hessienne, montrer que la fonction q est convexe. Quelle conséquence cela a-t-il sur ses minima ?

b) Montrer que le vecteur nul $\mathbf{0}_{\mathbb{R}^d}$ est un minimum local. Satisfait-il la condition suffisante à l'ordre 2 ?

c) En fonction de la réponse à la question précédente, la fonction peut-elle alors être fortement convexe ?

Exercice 3 : Fonctions quasi-convexes

Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est dite **quasi-convexe** si

$$\forall \boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d, \ \forall t \in [0,1], \quad f(t\boldsymbol{w} + (1-t)\boldsymbol{v}) \leq \max\{f(\boldsymbol{x}), f(\boldsymbol{y})\}. \tag{2}$$

Toute fonction convexe est quasi-convexe, mais la réciproque est fausse.

On s'intéresse ici aux solutions du problème

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimiser}} \, f(\boldsymbol{w}), \tag{3}$$

où l'on suppose que f est quasi-convexe et de classe $\mathcal{C}^2$.

a) Donner les conditions d'optimalité nécessaires à l'ordre 1 et à l'ordre 2 pour le problème (3).

b) Comme f est quasi-convexe, on a la propriété suivante :

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \ \forall \boldsymbol{v} \in \mathbb{R}^d, \quad \boldsymbol{v}^{\mathrm{T}} \nabla f(\boldsymbol{w}) = 0 \Rightarrow \boldsymbol{v}^{\mathrm{T}} \nabla^2 f(\boldsymbol{w}) \boldsymbol{v} \geq 0. \tag{4}$$

Soit $\boldsymbol{w}^*$ un point stationnaire d'ordre 1. Justifier que $\boldsymbol{w}^*$ est aussi un point stationnaire d'ordre 2.

Solutions des exercices

Solutions de l'exercice 1

a) Si $\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{y}$, alors on a

$$f(\boldsymbol{w}^*) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|^2 = \frac{1}{2}\|\mathbf{0}\|^2 = 0.$$

Or la fonction f est toujours positive ou nulle; on a ainsi

$$\forall \boldsymbol{w} \in \mathbb{R}^d, f(\boldsymbol{w}) \geq 0 = f(\boldsymbol{w}^*).$$

Cette propriété correspond à la définition d'un minimum global, d'où l'on conclut que $\boldsymbol{w}^*$ est bien un minimum global du problème.

b) Si $\boldsymbol{w}^*$ est un minimum local du problème (1) et donc de f, alors on a $\nabla f(\boldsymbol{w}^*) = \mathbf{0}$. C'est la condition d'optimalité nécessaire à l'ordre 1.

i) Si $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \succeq \mathbf{0}$, alors on a $\nabla^2 f(\boldsymbol{w}) \succeq \mathbf{0}$ pour tout $\boldsymbol{w}$: c'est une caractérisation de la convexité pour une fonction de classe $\mathcal{C}^2$, et l'on en conclut donc que f est convexe.

ii) Comme dans la question précédente, le fait que $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \succeq \mu \boldsymbol{I}_d$ signifie que $\nabla^2 f(\boldsymbol{w}) \succeq \mu \boldsymbol{I}_d$ pour tout $\boldsymbol{w} \in \mathbb{R}^d$. C'est une caractérisation de la convexité forte, d'où l'on conclut que f est μ-fortement convexe. Par conséquent, la solution du problème (on sait qu'il en existe au moins une d'après l'énoncé) est unique.

Solutions de l'exercice 2

a) Pour tous $\boldsymbol{w} \in \mathbb{R}^d$ et $\boldsymbol{v} \in \mathbb{R}^d$, on a en utilisant la linéarite des produits scalaires et produits matrice-vecteur :

$$\begin{aligned}
\boldsymbol{v}^{\mathrm{T}}\nabla^2 q(\boldsymbol{w})\boldsymbol{v} &= \boldsymbol{v}^{\mathrm{T}}(2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}} + \|\boldsymbol{w}\|^2\boldsymbol{I}_d)\boldsymbol{v} \\
&= \boldsymbol{v}^{\mathrm{T}}(2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v} + \|\boldsymbol{w}\|^2\boldsymbol{v}) \\
&= 2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v} + \|\boldsymbol{w}\|^2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} \\
&= 2(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v})^2 + \|\boldsymbol{w}\|^2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} \\
&= 2(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v})^2 + \|\boldsymbol{w}\|^2\|\boldsymbol{v}\|^2 \\
&\geq 0.
\end{aligned}$$

Par conséquent, pour tout $\boldsymbol{w} \in \mathbb{R}^d$, la matrice hessienne $\nabla^2 q(\boldsymbol{w})$ est semi-définie positive : on a $\nabla^2 q(\boldsymbol{w}) \succeq \mathbf{0}$. On en déduit que la fonction q est convexe, et donc que tous ses minima locaux sont globaux (elle ne possède ainsi que des minima globaux).

b) Puisque q est convexe, il y a équivalence entre minimum local et minimum global. Or, on a

$$q(\boldsymbol{w}) = \tfrac{1}{4}\|\boldsymbol{w}\|^4 \geq 0 = q(\mathbf{0}_{\mathbb{R}^d}).$$

Le vecteur nul $\mathbf{0}_{\mathbb{R}^d}$ est donc un minimum global de q. Pour satisfaire la condition suffisante d'optimalité à l'ordre 2, il faudrait avoir $\nabla^2 q(\boldsymbol{w}) \succ \mathbf{0}$; or, on trouve en remplaçant dans l'expression donnée dans l'énoncé que

$$\nabla^2 q(\mathbf{0}_{\mathbb{R}^d}) = \mathbf{0},$$

qui n'est pas définie positive mais uniquement semi-définie positive : par conséquent, le vecteur nul ne vérifie pas la condition suffisante d'optimalité. *Remarque : cela ne contredit pas le fait que $\mathbf{0}$ est un minimum global.*

c) Si la fonction était fortement convexe, on aurait $\nabla^2 q(\boldsymbol{w}) \succeq \mu \boldsymbol{I}_d \succ \mathbf{0}$ pour tout $\boldsymbol{w}$, et donc en particulier pour le vecteur nul. Ce n'est pas le cas, et on en conclut donc que cette fonction n'est pas fortement convexe.

Solutions de l'exercice 3

a) *Il s'agit d'une question de cours.* La condition nécessaire d'optimalité à l'ordre 1 s'énonce comme suit : si $\boldsymbol{w}^* \in \mathbb{R}^d$ est un minimum local de d, alors $\nabla f(\boldsymbol{w}^*) = \mathbf{0}$. La condition nécessaire d'optimalité à l'ordre 2 est plus précise encore : si $\boldsymbol{w}^* \in \mathbb{R}^d$ est un minimum local de f, alors

$$\nabla f(\boldsymbol{w}^*) = \mathbf{0} \quad \text{et} \quad \nabla^2 f(\boldsymbol{w}^*) \succeq \mathbf{0}.$$

b) Puisque $\boldsymbol{w}^*$ est un point stationnaire d'ordre 1, il vérifie la condition d'optimalité nécessaire à l'ordre 1 : on a donc $\nabla f(\boldsymbol{w}^*) = \mathbf{0}$. Par conséquent, on a

$$\forall \boldsymbol{v} \in \mathbb{R}^d, \quad \boldsymbol{v}^{\mathrm{T}} \nabla f(\boldsymbol{w}^*) = \boldsymbol{v}^{\mathrm{T}} \mathbf{0} = 0.$$

La première partie de l'implication (4) est donc vraie pour $\boldsymbol{w}^*$ et pour tout vecteur $\boldsymbol{v}$. On en déduit donc que la seconde partie de l'implication l'est aussi, c'est-à-dire que l'on a :

$$\boldsymbol{v}^{\mathrm{T}} \nabla^2 f(\boldsymbol{w}^*) \boldsymbol{v} \geq 0 \ \forall \boldsymbol{v} \in \mathbb{R}^d,$$

qui correspond à $\nabla^2 f(\boldsymbol{w}^*) \succeq \mathbf{0}$. Par conséquent, $\boldsymbol{w}^*$ vérifie la condition nécessaire d'optimalité à l'ordre 2 : c'est donc bien un point stationnaire d'ordre 2.