Outils d'optimisation pour les sciences des données et de la décision

29 novembre 2024

Aujourd'hui : Séance-bilan
- Examen 2023-2024 (adapté)
- Questions diverses
- Si temps : petit sujet bonus

Examen : Vendredi 13/12, 9h-11h
Documents : 1 feuille A4 recto-verso de notes manuscrites ou imprimées

Exercice 1 : Un problème non convexe

(Examen ID apprentissage 2023-2024)

Jeu de données $\{(x_i, y_i)\}_{i=1..m}$ $x_i \in \mathbb{R}^d$ et $y_i \in \{0,1\}$ $\forall i$ (classification)

But : Trouver un modèle linéaire des données

Problème d'optimisation

$$\underset{w \in \mathbb{R}^d}{\text{minimiser}} \quad \phi(w) = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \frac{1}{1+e^{-x_i^T w}} \right)^2$$

Rappel : $x_i^T w \gg 1 \Rightarrow \frac{1}{1+e^{-x_i^T w}} \approx 1$

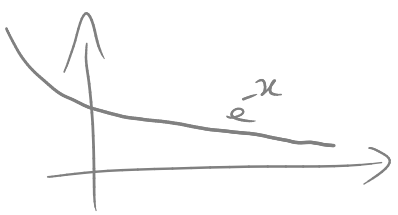
$x_i^T w \ll -1 \Rightarrow \frac{1}{1+e^{-x_i^T w}} \approx 0$

La fonction ϕ est de classe C^2 et est non convexe.

a) Justifier que 0 est un minorant de la fonction ϕ. Est-ce nécessairement sa valeur optimale ?

$$\forall w \in \mathbb{R}^d, \quad \phi(w) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{\left(y_i - \frac{1}{1+e^{-x_i^T w}} \right)^2}_{\geq 0 \ \forall w} \geq 0$$

(Somme de carrés, donc de termes positifs)



$e^{-x} \geq 0 \ \forall x$ mais $e^{-x} > 0$

0 n'est pas nécessairement la valeur optimale du problème.
Ce sera le cas si il existe $\bar{w} \in \mathbb{R}^d$ tel que $\phi(\bar{w}) = 0$

NB: La fonction $t \mapsto e^{-t}$ est un exemple de fonction telle que $e^{-t} \geq 0 \ \forall t$ mais e^{-t} n'est jamais égal à 0.

b) On souhaite appliquer la descente de gradient au problème.

i) Ecrire l'itération de cet algorithme en utilisant une taille de pas quelconque

$$w_{k+1} = w_k - \alpha_k \nabla \phi(w_k) \quad \text{avec } \alpha_k > 0$$

ii) Donner deux choix possibles pour la taille de pas
(Plus de 2 possibilités)

- Taille de pas constante: $\alpha_k = \alpha > 0 \ \forall k$ Ex) $\alpha = 0.1$, $\alpha = 0.01$
- Taille de pas décroissante: $\alpha_k \xrightarrow[k \to \infty]{} 0$ Ex) $\alpha_k = \frac{1}{\sqrt{k+1}}$, $\alpha_k = \frac{\alpha_0}{k+1}$ avec $\alpha_0 > 0$
- Taille de pas adaptative:
  α_k choisie en fonction du point courant, càd en fonction de $x_k, \nabla\phi(x_k), \phi(x_k)$

Ex) Recherche linéaire avec "backtracking"

$\alpha_k = 1$.
Tant que $\left(\phi(x_k - \alpha_k \nabla\phi(x_k)) \geq \phi(x_k) - c\alpha_k \|\nabla\phi(x_k)\|^2 \right)$ ($c > 0$)
$\alpha_k \leftarrow \alpha_k / 2$

Remarque bonus : Utiliser une recherche linéaire est plus coûteux qu'utiliser une taille de pas constante ou décroissante car cela requiert au moins un appel à $\phi(\cdot)$ par itération.

En revanche, les valeurs α_k obtenues par la recherche linéaire sont mieux adaptées à l'itéré correspondant que des valeurs constantes ou décroissantes choisies a priori.

iii) Sous les bonnes hypothèses, on peut obtenir une borne de complexité pour la descente de gradient. Quelle est l'ordre de grandeur de cette borne, et à quelle quantité s'applique-t-elle ?

$\min_{0 \le k \le K-1} \|\nabla \phi(w_k)\| \le \varepsilon$ en au plus $O\left(\frac{1}{\varepsilon^2}\right)$ itérations

$\Leftrightarrow$ L'algorithme calcule un itéré w_k tel que $\|\nabla \phi(w_k)\| \le \varepsilon$ en au plus $O(1/\varepsilon^2)$ itérations (mais ce n'est pas forcément le dernier itéré)

c) Supposons que la descente de gradient calcule un itéré w_k tel que $\|\nabla \phi(w_k)\| = 0$. w_k est-il nécessairement un minimum du problème ?

Non, pas nécessairement car la fonction ϕ est non convexe.

Contre-exemple : $d=1$, $f : w \mapsto w^3$ $\nabla f(0) = 0$ mais 0 est un point selle, pas un minimum.

d) Rappeler la condition d'optimalité nécessaire à l'ordre deux pour le problème. Un point qui vérifie cette condition est-il un minimum ?

Condition nécessaire à l'ordre 2 :

$$\left[\bar{w} \in \mathbb{R}^d \text{ minimum (local) de } \phi\right] \Rightarrow \begin{cases} \nabla\phi(\bar{w}) = 0 \\ \nabla^2\phi(\bar{w}) \succeq 0 \end{cases}$$

"semi-définie positive"

$v^T \nabla^2\phi(\bar{w}) v \geq 0 \quad \forall v \in \mathbb{R}^d$

Un point $\bar{w} \in \mathbb{R}^d$ tel que $\nabla\phi(\bar{w}) = 0$ et $\nabla^2\phi(\bar{w}) \succeq 0$ n'est pas nécessairement un minimum.

Contre-ex : $f : w \mapsto w^3$ et $\bar{w} = 0$

$\nabla f(0) = 0$, $\nabla^2 f(0) = 0 \succeq 0$

Remarques bonus : • Il existe des fonctions non convexes φ pour lesquelles on a

$$\left[\bar{w} \text{ minimum local}\right] \Leftrightarrow \left[\bar{w} \text{ vérifie } \nabla\varphi(\bar{w}) = 0,\ \nabla^2\varphi(\bar{w}) \succeq 0\right]$$

Même

$$\left[\bar{w} \text{ minimum global}\right] \Leftrightarrow \left[\nabla\varphi(\bar{w}) = 0 \text{ et } \nabla^2\varphi(\bar{w}) \succeq 0\right]$$

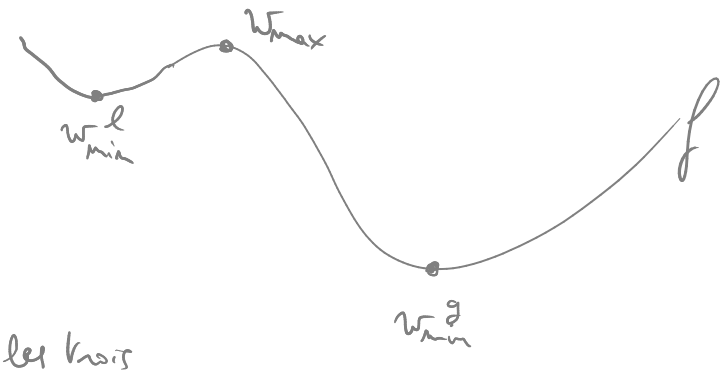
• Si $\nabla\varphi(\bar{w}) = 0$ et $\nabla^2\varphi(\bar{w}) \succ 0$ (définie positive $v^T \nabla^2\varphi(\bar{w}) v > 0$ si $v \neq 0_{\mathbb{R}^d}$)

alors $\bar{w}$ est un minimum local de φ.

e) Supposons que l'on applique la descente de gradient au problème de départ $\left(\min_{w \in \mathbb{R}^d} \phi(w)\right)$ en partant d'un point w_0 tiré aléatoirement. On observe que la méthode converge vers $\bar{w} \in \mathbb{R}^d$ tel que $\nabla \phi(\bar{w}) = 0$ et $\nabla^2 \phi(\bar{w}) \geq 0$. Comment expliquer ce comportement (était-ce attendu) ?

La descente de gradient avec initialisation aléatoire converge presque sûrement (càd en probabilité 1) vers un point stationnaire à l'ordre 2 (càd un point $\bar{w}$ tel que $\nabla \phi(\bar{w}) = 0$ et $\nabla^2 \phi(\bar{w}) \geq 0$). On pouvait donc s'attendre à ce résultat.

Intuition



w_{max} : max local
w_{min}^l : min local
w_{min}^g : min global

) les trois points pour lesquels le gradient est nul

$$\nabla f(w_{max}) = 0, \quad \nabla^2 f(w_{max}) \not\geq 0$$
$$\nabla f(w_{min}^l) = 0, \quad \nabla^2 f(w_{min}^l) \geq 0$$
$$\nabla f(w_{min}^g) = 0, \quad \nabla^2 f(w_{min}^g) \geq 0$$

$w_0 \neq w_{max}$ $\Rightarrow$ La descente de gradient converge vers w_{min}^l ou w_{max}^l

$w_0 = w_{max} \Rightarrow$ la descente de gradient reste en w_{max}

$w_k = w_{max}$

Si w_0 est tiré aléatoirement dans $\mathbb{R}$, alors $\mathbb{P}(w_0 = w_{max}) = 0$

Exercice 2. Gradient proximal (Examen 2023-2024)

Jeu de données : $\{(x_i, y_i)\}_{i=1..n}$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Problème

$$\underset{w \in \mathbb{R}^d}{\text{minimiser}} \quad \underbrace{\frac{1}{2n} \|Xw - y\|^2}_{\text{attache aux données}} + \frac{\lambda_2}{2}\|w\|^2 + \lambda_1 \|w\|_1$$

$\forall v \in \mathbb{R}^d$, $\|v\|^2 = \sum_{j=1}^{d} v_j^2$ et $\|v\|_1 = \sum_{j=1}^{d} |v_j|$

$\Rightarrow$ Problème de régression linéaire avec régularisation "elastic net" avec $\lambda_1 \geq 0$ et $\lambda_2 \geq 0$

a) Quelle est l'utilité générale de la régularisation ?

Utiliser un terme de régularisation permet de favoriser les solutions qui possèdent des propriétés souhaitées. Ces propriétés dépendent de la nature du terme de régularisation.

b) Si $\lambda_1 = 0$ et $\lambda_2 > 0$, quel est le rôle du terme de régularisation ?

$\lambda_1 = 0$, $\lambda_2 > 0$ $\Rightarrow$ Régularisation par $\frac{\lambda_2}{2}\|w\|^2$

"Ridge / ℓ_2 / énétré / ..."

La régularisation ℓ_2 :

- réduit la sensibilité de la solution vis-à-vis des données

sans régularisation : une petite perturbation de (X, y) peut entraîner un grand changement dans la valeur de solution

avec régularisation : une petite perturbation de (X, y) ne change que très peu la valeur de la solution

NB : Si $\lambda_2 >> 1$, alors la solution est quasiment nulle

- transforme un problème convexe en problème fortement convexe (intéressant ici car $w \mapsto \frac{1}{2n}\|Xw - y\|^2$ est convexe)

Fonction convexe : peut avoir une infinité de minimums globaux

Fonction fortement convexe : Unique minimum global

- réduit la norme 1 de la solution

ℓ_2 ($\|.\|$)

c) Si $\lambda_1 > 0$ et $\lambda_2 = 0$, quel est le rôle du terme de régularisation ?

$\lambda_1 > 0$, $\lambda_2 = 0$ $\Rightarrow$ Régularisation $\lambda_1 \|w\|_1$

"en norme ℓ_2/ℓ_1/ LASSO"

Rôle : . Favorise les solutions parcimonieuses (avec beaucoup de coefficients nuls)

. Réduit la norme ℓ_1 $\|.\|_1$ de la solution

. Permet d'identifier les coordonnées les plus importantes pour l'attache aux données

d) On rappelle que $\ell : w \longmapsto \frac{1}{2n}\|Xw - y\|^2$ est C^1 ; et on peut donc appliquer l'algorithme du gradient proximal au problème

$$\underset{w \in \mathbb{R}^d}{\text{minimiser}} \; \ell(w) + \frac{\lambda_2}{2}\|w\|^2 + \lambda_1 \|w\|_1$$

Écrire l'itération du gradient proximal pour ce problème avec une taille de pas quelconque !

$$w_{k+1} \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \ell(w_k) + \nabla \ell(w_k)^T (w - w_k) + \frac{1}{2\alpha_k}\|w - w_k\|^2 + \frac{\lambda_2}{2}\|w\|^2 + \lambda_1 \|w\|_1 \right\}$$

avec $\alpha_k > 0$

Approximation de $\ell(w)$ pertinente lorsque $\|w - w_k\| \ll 1$

Terme proximal : Garantit que $\|w - w_k\|$ est petite

Terme de régularisation repris tel quel

NB: Utile en pratique si le sous-problème à chaque itération est plus facile à résoudre que le problème de départ (c'est le cas pour la régularisation considérée !)

e) Lorsque $\lambda_2 = 0$ et $\lambda_1 > 0$, à quel algorithme le gradient proximal est-il équivalent ?

ISTA (Iterative Soft-Thresholding Algorithm)

→ Dans ce cas, on peut écrire explicitement w_{k+1} en fonction de w_k :

$$\forall j = 1..d, \quad [w_{k+1}]_j = \begin{cases} [w_k - \alpha_k \nabla \ell(w_k)]_j - \alpha_k \lambda_1 & \text{si } [w_k - \alpha_k \nabla \ell(w_k)]_j > \alpha_k \lambda_1 \\ [w_k - \alpha_k \nabla \ell(w_k)]_j + \alpha_k \lambda_1 & \text{si } [w_k - \alpha_k \nabla \ell(w_k)]_j < -\alpha_k \lambda_1 \\ 0 & \text{si } [w_k - \alpha_k \nabla \ell(w_k)]_j \in [-\alpha_k \lambda_1, \alpha_k \lambda_1] \end{cases}$$

f) Si $\lambda_1 = 0$ et $\lambda_2 > 0$, quel algorithme peut-on utiliser au lieu du gradient proximal pour résoudre le problème ?

$$\left(\lambda_1 = 0, \lambda_2 > 0 \Rightarrow \min_{w \in \mathbb{R}^d} \underbrace{\frac{1}{2n} \|Xw - y\|^2}_{C^1} + \underbrace{\frac{\lambda_2}{2} \|w\|^2}_{C^1} \right)$$

→ La fonction $w \mapsto \frac{1}{2n} \|Xw - y\|^2 + \frac{\lambda_2}{2} \|w\|^2$ étant C^1, on peut appliquer la descente de gradient à ce problème

→ Cette fonction est même fortement convexe et quadratique : on peut donc aussi appliquer le gradient accéléré et la

méthode "heavy-ball" / de Polyak

→ En écrivant $\frac{1}{2n}\|Xw-y\|^2+\frac{\lambda_2}{2}\|w\|^2=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}(x_i^Tw-y_i)^2+\frac{\lambda_2}{2}\|w\|^2\right)$

on voit qu'on peut aussi appliquer le gradient stochastique (et toutes ses variantes) au problème, qui s'écrit comme une somme finie de termes qui dépendent chacun d'un point différent du jeu de données

g) Si $\lambda_1>0$ et $\lambda_2>0$, alors le sous-problème vaut

$$\underset{w\in\mathbb{R}^d}{\text{minimiser}}\left\{\ell(w_k)+\nabla\ell(w_k)^T(w-w_k)+\frac{1}{2\alpha_k}\|w-w_k\|^2+\frac{\lambda_2}{2}\|w\|^2+\lambda_1\|w\|_1\right\}$$

Parmi les algorithmes vus en cours, lequel peut être appliqué à la résolution de ce problème ? Justifier.

Comme $w\mapsto\|w\|_1$ n'est pas C^1, on ne peut pas appliquer la descente de gradient, le gradient stochastique ou les méthodes de gradient accéléré / heavy ball.

Par contre, en posant $g(w)=\ell(w_k)+\nabla\ell(w_k)^T(w-w_k)+\frac{1}{2\alpha_k}\|w-w_k\|^2+\frac{\lambda_2}{2}\|w\|^2$

le problème s'écrit

$$\underset{w\in\mathbb{R}^d}{\text{minimiser}}\ g(w)+\lambda_1\|w\|_1$$

avec g C^1

On peut donc appliquer ISTA.

Exercice 3 : Somme finie et optimisation distribuée

(Exam 2023-2024)

Problème : $\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} f_i(w)$ $\quad f_i : \mathbb{R}^d \to \mathbb{R}$ de classe C^1 $\forall i$

a) Solutions du problème

i) Donner la définition d'un minimum global du problème

$w^* \in \mathbb{R}^d$ est un minimum global

si $\sum_{i=1}^{n} f_i(w^*) \leq \sum_{i=1}^{n} f_i(w) \quad \forall w \in \mathbb{R}^d$

ii) On pose $f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$. Justifier que les problèmes

$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} f_i(w)$ et $\min_{w \in \mathbb{R}^d} f(w)$ ont le même ensemble de solutions.

Comme $\frac{1}{n}$ est une constante strictement positive, les deux problèmes ont le même nombre de solutions

optionnel

En revenant à la définition, on voit que

$$\left[w^* \in \operatorname*{argmin}_{w \in \mathbb{R}^d} \sum_{i=1}^{n} f_i(w) \right] \iff \left[\sum_{i=1}^{n} f_i(w^*) \leq \sum_{i=1}^{n} f_i(w) \quad \forall w \in \mathbb{R}^d \right]$$

$$\iff \left[\frac{1}{n} \sum_{i=1}^{n} f_i(w^*) \leq \frac{1}{n} \sum_{i=1}^{n} f_i(w) \quad \forall w \in \mathbb{R}^d \right]$$

$$\iff \left[w^* \in \operatorname*{argmin}_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w) \right]$$

iii) Si $f : w \mapsto \frac{1}{n} \sum_{i=1}^{n} f_i(w)$ est fortement convexe, que peut-on dire des minima globaux du problème ?

Il existe un unique minimum global

b) $\min_{w \in \mathbb{R}^d} f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$

i) Ecrire l'itération de descente de gradient (cf plus haut)

ii) Si toutes les f_i sont des fonctions convexes, quelle est la complexité de la descente de gradient sur ce problème (sous les bonnes hypothèses) ?

$f(w_k) - \min_{w \in \mathbb{R}^d} f(w) \leq \varepsilon$ en au plus $O(\varepsilon^{-1})$ itérations

(NB : Meilleur que le cas non convexe)

iii) Donner un algorithme qui possède une meilleure complexité que la descente de gradient sous les hypothèses de la question ii)

Le gradient accéléré / la méthode de Nesterov

NB : . Pour Nesterov, $f(w_k) - \min_{w \in \mathbb{R}^d} f(w) \leq \varepsilon$ en au plus $O(\varepsilon^{-1/2})$ itérations

Nesterov :

$$w_{k+1} = w_k + \beta_{k+1}(w_k - w_{k-1}) - \alpha_k \nabla f\left(w_k + \beta_{k+1}(w_k - w_{k-1})\right)$$