TD 7 : Régularisation

Optimisation pour l'apprentissage automatique, M2 Big Data

14 décembre 2021

Dauphine | PSL TUNIS

Exercice 1 : Perte de Huber renversée

On considère un modèle linéaire $\boldsymbol{x} \mapsto \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$ et un jeu de données $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, où $\boldsymbol{x}_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}$.

Dans cet exercice, on renverse l'idée de la perte de Huber, en proposant une fonction de perte qui ressemble à la valeur absolue sur $[-1, 1]$ et à une quadratique partout ailleurs. On définit donc la "*perte de Huber renversée*" comme

$$\begin{array}{rccl} v: & \mathbb{R} & \to & \mathbb{R} \\ & t & \mapsto & v(t) := \left\{ \begin{array}{ll} |t| & \text{if } |t| < 1 \\ \frac{t^2+1}{2} & \text{sinon.} \end{array} \right. \end{array} \tag{1}$$

Cette fonction est convexe mais non lisse (car non dérivable en 0).

a) On s'intéresse tout d'abord au problème non lisse suivant :

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}} \frac{1}{n}\sum_{i=1}^n v(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i). \tag{2}$$

Peut-on appliquer l'algorithme de descente de gradient ? Si non, quel outil peut-on utiliser pour construire des algorithmes pour résoudre (2) ?

b) On considère maintenant le problème

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}} f(\boldsymbol{w}) + \lambda \sum_{i=1}^d v([\boldsymbol{w}]_i), \tag{3}$$

où f est une fonction de perte de classe $\mathcal{C}^1$, et $\lambda > 0$.

i) Comment s'appelle un problème de cette forme ? Quel est le rôle du second terme de l'objectif ?

ii) Écrire l'itération générique de la méthode du gradient proximal pour ce problème.À quelle condition est-il envisageable d'utiliser cette méthode en pratique ?

Exercice 2 : Autour du problème de LASSO

On considère le problème dit de *LASSO généralisé* donné par

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|^2+\lambda\|\boldsymbol{A}\boldsymbol{w}\|_1, \tag{4}$$

où $\boldsymbol{X}\in\mathbb{R}^{n\times d}$ et $\boldsymbol{y}\in\mathbb{R}^n$ représentent un jeu de données, $\lambda>0$ et $\boldsymbol{A}\in\mathbb{R}^{m\times d}$ avec m un entier supérieur ou égal à 1. On rappelle que

$$\forall \boldsymbol{v}\in\mathbb{R}^m, \|\boldsymbol{v}\|_1=\sum_{i=1}^{m}|v_i|,$$

où les v_i sont les coefficients du vecteur $\boldsymbol{v}$.

a) Pourquoi ne peut-on pas appliquer la méthode de descente de gradient à ce problème ?

b) Écrire l'itération de l'algorithme du gradient proximal appliqué à ce problème avec une taille de pas constante. *On donne* $\nabla\psi(\boldsymbol{w})=\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y})$ *pour* $\psi:\boldsymbol{w}\mapsto\frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|^2$.

c) On considère le cas particulier $m=d$ et $\boldsymbol{A}=\boldsymbol{I}\in\mathbb{R}^{d\times d}$.

i) Que représente alors le terme en $\lambda\|\boldsymbol{A}\boldsymbol{w}\|_1$ dans la fonction objectif du problème (4) ? Quelle est son utilité ?

ii) À quel algorithme vu en cours correspond alors la méthode de gradient proximal décrite en question b) ?

Exercice 3 : Perte logistique régularisée

On se donne un problème de régression logistique construit à partir d'un jeu de données $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ pour lequel on considère une perte logistique

$$\ell : (h, y) \mapsto \ln\left(1 + \exp(-y\, h)\right) \tag{5}$$

et un modèle linéaire $\boldsymbol{x} \mapsto \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}$ paramétré par $\boldsymbol{w} \in \mathbb{R}^d$. Le problème d'optimisation associé est ainsi

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}}\; f(\boldsymbol{w}) + \lambda\Omega(\boldsymbol{w}), \tag{6}$$

avec

$$f(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp(-y_i\, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w})\right), \tag{7}$$

$\lambda > 0$ et $\Omega(\boldsymbol{w}) := \frac{1}{2}\|\boldsymbol{w}\|^2$. La fonction Ω est de classe $\mathcal{C}^2$, et on a $\nabla\Omega(\boldsymbol{w}) = \boldsymbol{w}$.

a) Le terme $\lambda\Omega(\boldsymbol{w})$ ne dépend pas des données. Comment appelle-t-on ce terme, et quel est son rôle ?

b) Lorsque $\lambda > 0$, le problème (6) est fortement convexe. Comment cela se traduit-il au niveau de la dérivée seconde de la fonction $f + \lambda\Omega$?

c) Écrire (en pseudo-code) l'itération du gradient proximal appliqué au problème (6), avec un pas générique α_k.

d) Comme Ω est dérivable, on peut appliquer l'algorithme de descente de gradient à (6). Écrire (en pseudo-code) l'itération de la descente de gradient pour ce problème, avec un pas générique α_k.

e) Les solutions de (6) sont généralement de norme euclidienne plus faible que celles du problème non régularisé ($\lambda = 0$). Comment la régularisation influe-t-elle sur cette norme, et quel est le but derrière la réduction de cette norme ?