TD 06 : Gradient stochastique

Outils d'optimisation pour les sciences des données et de la décision, M2 MIAGE

22 novembre 2024

Dauphine | PSL
UNIVERSITÉ PARIS

Exercice 1 : Perte de Huber

On considère un jeu de données $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, où $n \geq 1$, $\boldsymbol{x}_i \in \mathbb{R}^d$ avec $d \geq 1$ et $y_i \in \mathbb{R}$. On cherche un modèle linéaire qui prédise au mieux chaque y_i à partir du $\boldsymbol{x}_i$ correspondant. On définit donc une famille de modèles paramétrée par $\boldsymbol{w} \in \mathbb{R}^d$ comme suit :

$$\begin{array}{rccl} h_{\boldsymbol{w}} : & \mathbb{R}^d & \to & \mathbb{R} \\ & \boldsymbol{x} & \mapsto & \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} = \sum_{i=1}^d [\boldsymbol{x}]_i [\boldsymbol{w}]_i. \end{array}$$

Pour un modèle $h_{\boldsymbol{w}}$, on considèrera que ce modèle prédit parfaitement y_i à partir de $\boldsymbol{x}_i$ si on a $\ell\left(h_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i\right) = \ell\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i\right) = 0$, où $\ell : \mathbb{R} \to \mathbb{R}$ est la fonction de perte de Huber définie par :

$$\ell(t) = \begin{cases} \frac{1}{2}t^2 & \text{if } |t| < 1 \\ |t| - \frac{1}{2} & \text{otherwise.} \end{cases} \tag{1}$$

Cette fonction se comporte comme $t \mapsto \frac{t^2}{2}$ pour $|t| < 1$ et comme $t \mapsto |t|$ lorsque $|t|$ est très grand. Contrairement à ce que son expression peut suggérer, ℓ est continûment dérivable (ou de classe $\mathcal{C}^1$)

L'expression $\ell\left(h_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i\right)$ représente l'erreur du modèle en $(\boldsymbol{x}_i, y_i)$, et on cherche un modèle (c'est-à-dire un vecteur $\boldsymbol{w} \in \mathbb{R}^d$) tel que la somme de ces erreurs soit minimale. On considère donc :

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i). \tag{2}$$

a) Justifier que 0 est un minorant de (2). Est-ce sa valeur minimale ?

b) Le gradient de f en $\boldsymbol{w} \in \mathbb{R}^d$ est donné par

$$\nabla f(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n \ell'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i)\boldsymbol{x}_i, \tag{3}$$

avec

$$\ell'(t) = \begin{cases} 1 & \text{si} \quad t > 1 \\ t & \text{si} \quad |t| \leq 1 \\ -1 & \text{si} \quad t < -1. \end{cases}$$

Écrire (en pseudo-code) l'itération de descente de gradient avec une taille de pas constante α et en utilisant la formule (3). Que devient cette itération si le point courant est un minimum local ?

c) Une constante de Lipschitz pour ∇f est $L = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2$. Comment utiliser cette constante pour définir la longueur de pas ? Lorsque L est inconnue, donner deux choix possibles pour la taille de pas.

d) La fonction f s'écrit $f = \frac{1}{n}\sum_{i=1}^{n} f_i$, où $f_i(\boldsymbol{w}) = \ell(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i)$. Le gradient f_i en $\boldsymbol{w}$ est

$$\nabla f_i(\boldsymbol{w}) = \ell'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} - y_i)\boldsymbol{x}_i.$$

Écrire (en pseudo-code) l'itération du gradient stochastique pour ce problème sans choix particulier de taille de pas.

e) *On considère ici que notre unité de coût est un accès à un $\boldsymbol{x}_i$.* Quel est le coût d'une itération de descente de gradient, et celui d'une itération de gradient stochastique ?

f) Quand on applique le gradient stochastique avec une longueur de pas fixe, on peut parfois observer que la méthode génère des itérés de norme de plus en plus grande, ce qui conduit à un dépassement de mémoire pour l'algorithme. Fournir une justification à ce phénomène.

g) On considère une variante par fournées (*batch*) du gradient stochastique, dans laquelle on choisit un sous-ensemble de n_b composantes dans la somme finie de (2).

i) Écrire l'itération correspondante (en pseudo-code).

ii) Si n_b correspond au nombre de processeurs disponibles pour les calculs, quel peut être l'intérêt de choisir n_b comme taille de fournée ?

iii) Donner un autre intérêt plus général des méthodes par fournées en comparaison avec l'algorithme du gradient stochastique basique.

iv) Supposons que l'on utilise plusieurs tailles de fournées et que l'on observe une amélioration en termes de convergence quand n_b augmente pour $1 \leq n_b \leq \frac{n}{10}$. Supposons que l'on observe aussi qu'augmenter n_b au-delà de $n/10$ conduise à une dégradation de la performance. Comment expliquer ces observations ?

Exercice 2 : Tirage par importance

On considère un problème en somme finie de la forme

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}}\, f(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w}), \tag{4}$$

où, pour chaque $i = 1, \dots, n$, la fonction f_i est de classe $\mathcal{C}^1$et son gradient ∇f_i est L_i-lipschitzien. On suppose également que la fonction f est μ-fortement convexe.

On considère l'itération du gradient stochastique, où l'on suppose que l'indice i_k correspondant au gradient stochastique est tiré selon son importance (*importance sampling*), que l'on définit en fonction des quantités $c_i = \frac{nL_i}{\sum_{j=1}^n L_j}$. On a ainsi

$$\forall i \in \{1, \dots, n\}, \qquad \mathbb{P}\left(i_k = i\right) \;=\; \frac{c_i}{\sum_{j=1}^n c_j}. \tag{5}$$

On remplace alors l'itération du gradient stochastique telle que vue en cours par

$$\boldsymbol{w}_{k+1} \leftarrow \boldsymbol{w}_k - \frac{\alpha_k}{c_{i_k}} \nabla f_{i_k}(\boldsymbol{w}_k). \tag{6}$$

a) Montrer que

$$\mathbb{P}\left(i_k = i\right) = \frac{L_i}{\sum_{j=1}^n L_j}.$$

Interpréter alors le concept de tirage par importance en fonction de ce résultat : quelles valeurs de i ont le plus de chances d'être tirées ?

b) Montrer que l'on a $\mathbb{E}_{i_k}\left[\frac{1}{c_{i_k}} \nabla f_{i_k}(\boldsymbol{w}_k)\right] = \nabla f(\boldsymbol{w}_k)$.

c) On peut montrer que ∇f est L-lipschitzien avec $L = \frac{1}{n}\sum_{i=1}^n L_i$. Supposons que l'on se fixe une longueur de pas constante $\alpha_k = \frac{1}{L}$ pour tout k. Pour un même indice i_k, on souhaite comparer l'itération k du gradient stochastique classique à l'itération (6).

i) Montrer que $\frac{\alpha_k}{c_{i_k}} = \frac{1}{L_{i_k}}$.

ii) Quand peut-on alors avoir $\frac{\alpha_k}{c_{i_k}} \geq \alpha_k$? Que cela signifie-t-il sur l'itération (6) ?