Optimisation pour l'apprentissage automatique

M2 IASD Apprentissage

Projet 2024/2025

Dauphine | PSL 🔀

Instructions

- Version courante : 25 février 2025.
 - 2025.02.25 : Première version.
- La version la plus récente du projet (mise à jour en cas de changement) est disponible à l'adresse https://www.lamsade.dauphine.fr/~croyer/ensdocs/OAA/ProjOAA.pdf.
- Le travail décrit ci-dessous peut être réalisé en binôme ou individuellement.
- Les réponses aux différentes questions ainsi que l'implémentation utilisée devront être envoyées au format ZIP à clement.royer@lamsade.dauphine.fr. Le nom de l'archive devra comporter le(s) nom(s) de chaque étudiant(e) impliqué(e) dans le projet.
- La date limite de rendu est fixée au 6 avril 2025 AOE (Anywhere On Earth).

Notations et remarques préliminaires

- Les dimensions des vecteurs ou matrices seront toujours supposées supérieures ou égales à 1.
- La notation $\|\cdot\|$ désignera la norme euclidienne d'un vecteur, càd $\|v\| = \sum_{i=1}^{m} [v]_i^2$ pour tout $v \in \mathbb{R}^m$.
- Pour une matrice $A \in \mathbb{R}^{m_1 \times m_2}$, on notera $\|A\|_F^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} [A]_{ij}^2$ la norme dite de Frobenius sur cette matrice.

Introduction

Ce court projet se base sur les notebooks fournis lors des séances de cours, dont il se veut une extension. Cependant, les étudiant(e)s sont cependant libres de proposer leur propre implémentation des algorithmes et méthodes proposées.

1 Variables matricielles et optimisation

Dans cette partie, on considère un problème de régression linéaire basique, pour lequel on se donne $X = [x] \in \mathbb{R}^{n \times 1}$ avec x un vecteur non nul et $y \in \mathbb{R}^n$. Dans sa forme classique, le problème de régression linéaire consiste alors à déterminer $w \in \mathbb{R}^1$ tel que $Xw = wx \approx y$. Cela s'écrit comme le problème de régression linéaire suivant :

$$\underset{w \in \mathbb{R}}{\operatorname{minimiser}} \frac{1}{2n} \|w \boldsymbol{x} - \boldsymbol{y}\|^2.$$
(1)

Question 1

- a) On suppose que le vecteur x est non nul. Donner alors l'ensemble des solutions du problème (1).
- b) Reprendre le code de régression linéaire et de descente de gradient vu durant la première séance de TP, et vérifier que l'algorithme converge vers une solution lorsque x et y sont générés selon une loi normale $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ (cf notebook sur la descente de gradient).

Comme vu en cours, une solution du problème (1) n'est pas nécessairement une solution du système linéaire wx = y. Dans le cas qui nous intéresse, cela ne peut être vrai que si les vecteurs x et y sont colinéaires. Cependant, on peut chercher à surparamétrer le problème en cherchant à calculer une matrice $W \in \mathbb{R}^{n \times n}$ telle que Wx = y. Cela conduit à formuler le problème

$$\min_{\boldsymbol{W} \in \mathbb{R}^{n \times n}} f(\boldsymbol{W}) := \frac{1}{2n} \|\boldsymbol{W}\boldsymbol{x} - \boldsymbol{y}\|^2.$$
(2)

Question 2

- a) Justifier que la valeur du problème (2) est toujours 0.
- b) Adapter le code de descente de gradient de la question 1 au problème (2) et comparer les résultats à ceux de la question 1 pour les mêmes données (x, y).
 NB: On pourra utiliser la formule de la dérivée de f par rapport à W vue dans le notebook sur la différentiation automatique, et considérer que l'itération de descente de gradient s'écrit W_{k+1} = W_k α_kD_Wf(W_k).

2 Factorisation matricielle

Dans cette partie, on prolonge l'étude précédente en considérant des problèmes de factorisation matricielle, pour lesquels le modèle est défini au moyen de deux matrices.

Comme dans l'un des exercices vus en TD, on considère une matrice de données (a priori rectangulaire) $X \in \mathbb{R}^{n_1 \times n_2}$, dont on va chercher une approximation via une matrice de rang 1. En utilisant le fait qu'une matrice de rang 1 s'écrit toujours sous la forme uv^T avec u et v des vecteurs, le problème de déterminer la meilleure approximation de rang 1 peut s'écrire :

$$\min_{\substack{\boldsymbol{u} \in \mathbb{R}^{n_1}\\ \boldsymbol{v} \in \mathbb{R}^{n_2}}} \lim_{\boldsymbol{v} \in \mathbb{R}^{n_2}} \|\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}} - \boldsymbol{X}\|_F^2 = \frac{1}{2n} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} ([\boldsymbol{u}]_i [\boldsymbol{v}]_j - [\boldsymbol{X}]_{ij})^2.$$
(3)

Question 3

a) Générer une matrice X = xz^T ∈ ℝ^{10×20} de rang 1 òu x et z sont des vecteurs générés selon une loi normale N(0, I). puis comparer l'algorithme du gradient stochastique basique et l'algorithme de descente de gradient sur le problème (3).
 NB: Dans l'implémentation, on pourra concaténer les vecteurs u et v dans un unique vecteur w

de taille $n_1 + n_2$.

b) Tester plusieurs valeurs de tailles de fournées (batch size) et en conclure une bonne valeur pour ce problème.

Une généralisation du problème d'approximation de rang 1 consiste à rechercher une approximation de rang (au plus) $r \leq \min(n_1, n_2)$, ce que l'on modélise par le problème

$$\min_{\substack{\boldsymbol{U} \in \mathbb{R}^{n_1 \times r} \\ \boldsymbol{V} \in \mathbb{R}^{n_2 \times r}}} \frac{1}{2n} \| \boldsymbol{U} \boldsymbol{V}^{\mathrm{T}} - \boldsymbol{X} \|_F^2 = \frac{1}{2n} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} ([\boldsymbol{U} \boldsymbol{V}^{\mathrm{T}}]_{ij} - [\boldsymbol{X}]_{ij})^2.$$
(4)

Remarque 2.1 Toute matrice $A \in \mathbb{R}^{n_1 \times n_2}$ de rang au plus r s'écrit sous la forme $A = \sum_{i=1}^r a_i b_i^T$ avec $a_i \in \mathbb{R}^{n_1}$ et $b_i \in \mathbb{R}^{n_2}$ pour tout i = 1, ..., r.

Question 4

- *i)* Adapter le code de la question 3 pour implémenter la descente de gradient et le gradient stochastique sur le problème (4).
- ii) Comparer les deux variantes dans les cas suivants, avec $(n_1, n_2) = (10, 20)$:
 - X est une matrice de rang r = 5, où r est la valeur utilisée dans (4).
 - X est une matrice de rang $min(n_1, n_2)$ et r = 1.

3 Factorisation matricielle régularisée

On considère enfin le cas non pris en compte dans la question 4, où r > 1 dans le problème (4) mais la matrice X est une version bruitée d'une matrice de rang 1. Dans ce cas, et pour promouvoir la robustesse vis, on rajoute un terme de régularisation au problème, et on obtient ainsi :

$$\min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n_1\times r}\\\boldsymbol{V}\in\mathbb{R}^{n_2\times r}}} \frac{1}{2n} \|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}-\boldsymbol{X}\|_F^2 + \frac{\lambda}{2} \|\boldsymbol{U}\|_F^2 + \frac{\lambda}{2} \|\boldsymbol{V}\|_F^2,$$
(5)

où $\lambda > 0$.

Question 5

- *i)* Adapter le code de la question 4 pour implémenter la descente de gradient et le gradient stochastique sur le problème (5).
- ii) Générer une matrice $X = xz^{T} + \epsilon$, où x, z et ϵ sont des vecteurs gaussiens tirés selon une loi $\mathcal{N}(\mathbf{0}, \mathbf{I})$.¹ Comparer le rang des matrices obtenues en résolvant (5) avec ce choix de X pour r = 5 et différentes valeurs de λ (on inclura $\lambda = 0$).
- iii) Peut-on déterminer λ tel que la solution est de rang 1? Si oui, la solution correspond-elle à xz^{T} ?
- *iv)* Les conclusions sont-elles différentes selon que l'on utilise le gradient stochastique ou la descente de gradient ?