Bureau d’étude : Optimisation différentiable 2

MICHELET Lucie / PEDONI Chloé / DIETTE Timo

Classe : 3TS2

Professeur : M. COUFFIGNAL

Date de soumission : 05/05/2022

1- Les moindres carrées multi-classes avec régularisation

Nous noterons dans tout cette partie : *Dat*a $\in M_{m,d}(\mathbb{R})$ une matrice de m données où chaque donnée est donnée par un vecteur ligne de taille d de la matrice *Data*.

On suppose que ces données sont fournies avec leurs catégories. Notons $C \in \mathbb{N}^*$ le nombre de catégorie des données *Data*

1.1. Version linéaire

On cherche à apprendre une fonction linéaire:

$$f: \mathbb{R}^{d+1} \to \mathbb{R}^C$$

De la forme :

$$f(x) = x^T W$$

Où : $x = (Data_{ligne,i}, 1)^T$ avec $Data_{ligne,i}$ le i ème vecteur ligne de *Data* et qui renvoie un vecteur ligne de taille C de la forme :

$$b := \begin{cases} b_j = 1 & si\ le\ label\ de\ x\ est\ j \in [\![1, C]\!] \\ 0 & sinon \end{cases}$$

Déterminons la taille de W que l'on appellera la matrice de paramètres.

On sait que Data $\in M_{m,d}(\mathbb{R})$ donc $Data_{ligne,i} \in M_{1,d}(\mathbb{R})$.

On sait que $x = (Data_{ligne,i}, 1)^T$ donc $x \in M_{d+1,1}(\mathbb{R})$ ➔ $x^T \in M_{1,d+1}(\mathbb{R})$

On sait également que f renvoie un vecteur ligne de taille C

Ainsi : on cherche dim1 et dim2 les dimensions de W tel que :

$$[\ M_{1,C}\ ] = [\ M_{1,d+1}\ ] @ [\ M_{dim1,dim2}\ ]$$

Logiquement on en déduit que $W \in M_{d+1,C}$

Notre problème est sous la forme suivante : $f(x) = x^T W = b$

On va chercher à minimiser $\|f(x_i) - b_i\|_F = \left\| \begin{matrix} x_1^T - b_1 \\ \vdots \\ x_m^T - b_m \end{matrix} \right\|_F$

➔ $\|f(x_i) - b_i\|_F = \|DW - B\|_F$

Avec D la matrice par blocs : D :=(Data 1_m)

Et B la matrice des coefficient b de la forme :

$$B_{ij} := \begin{cases} 1 & si\ le\ label\ de\ Data_{ligne,i}\ est\ j \in [\![1, C]\!] \\ 0 & sinon \end{cases}$$

Avec les moindres carrés on obtient un problème de la forme :

$$\arg min_w \|DW - B\|_F$$

$$\Leftrightarrow \arg min_w \frac{1}{2} \|DW - B\|_F^2$$

Intéressons-nous à la forme dite régularisée du problème d'apprentissage :

$$\text{(P)} \qquad \arg min_w \frac{1}{2} \|DW - B\|_F^2 + \frac{\rho}{2}\|W\|_F^2$$

Avec ρ ∈ $\mathbb{R}^*_+$ le facteur de régulation.
L'intérêt d'ajouter le terme $\frac{\rho}{2}\|W\|_F^2$ est de permettre d'éliminer les valeurs trop extrêmes de nos paramètres afin d'optimiser la résolution de notre problème.
On transforme (P) pour l'écrire sous une forme quadratique matricielle,
On définit que le produit scalaire est donné par :

$$\langle X, Y\rangle \coloneqq trace(X^T Y)$$

Ainsi :

$$\text{(P)} \qquad \arg min_w \frac{1}{2} \|DW - B\|_F^2 + \frac{\rho}{2}\|W\|_F^2$$

$$\Leftrightarrow \quad \arg min_w \frac{1}{2}\, trace((W^T D^T - B^T)(WD - B)) + \frac{\rho}{2}\text{trace}(W^T W)$$

$$\Leftrightarrow \quad \arg min_w \frac{1}{2}\, trace(W^T D^T DW - 2B^T DW + B^T B) + \frac{1}{2}\text{trace}(\rho W^T W)$$

$$\Leftrightarrow \quad \arg min_w \frac{1}{2}\, trace(W^T D^T DW + \rho W^T W - 2B^T DW) \qquad \{B^T B=0\}$$

$$\Leftrightarrow \quad \arg min_w \frac{1}{2}\, trace(W^T (D^T D + \rho I_d) W) - trace(B^T DW)$$

On retrouve ainsi la forme :

$$\text{(P1)} \qquad \arg min_w \frac{1}{2}\langle W, AW\rangle - \langle C, W\rangle$$

Avec $A = D^T D + \rho I_d$ et $C = BD^T$

1.2. Version non-linéaire avec l'astuce des noyaux

Dans la partie précédente nous cherchions à apprendre une fonction linéaire :

$$f: \mathbb{R}^{d+1} \to \mathbb{R}^C$$

de la forme par blocs :

$$f(x) \coloneqq x^T W = (x^T w_1 \quad x^T w_2 \quad \dots \quad x^T w_C)$$

Nous allons utiliser l'astuce des noyaux sur la forme précédente. On suppose fixée un noyau de type positif kern : $\mathbb{R}^{d+1} * \mathbb{R}^{d+1} \to \mathbb{R}$, et un ensemble de taille P∈$\mathbb{N}^*$:

$$P_k := \{x \in data\}$$

On sait alors d'après le théorème de Mercer qu'il existe une fonction de re description $\Phi : P_k \to H$ et un produit scalaire $\langle \bullet, \bullet \rangle$ sur H tel que :

$$kern(a, b) = \langle \Phi(a), \Phi(b)\rangle$$

Pour tout (a,b) ∈ P^2_K

Pour tout w ∈ H et x :=$(d,1)^T$ avec d vecteur ligne de Data :
On part de l'expression du produit scalaire :

$$\langle \Phi(x), w\rangle = \langle \Phi(x), \sum a_i \Phi(x_i)\rangle$$

$$\Leftrightarrow = \sum a_i \langle \Phi(x), \Phi(x_i)\rangle$$

$$\Leftrightarrow = \sum a_i\, kern(x, x_i)$$

$$\rightarrow \langle \Phi(x), w\rangle = (kern(x, x_1) \quad \dots \quad kern(x, x_P)) \begin{pmatrix} a_1 \\ \vdots \\ a_P \end{pmatrix}$$

On pose une matrice D^K de taille m*P , une matrice de paramètres W^K de tille P*C et K la matrice de Gram associée à P_K et kern. On garde la matrice B de la partie précédente.
Il est possible de transformer la version linéaire du problème en une version non linéaire:

$$(P) \quad \arg min_w \frac{1}{2} \|DW - B\|_F^2 + \frac{\rho}{2}\|W\|_F^2$$

$$\Leftrightarrow \quad \arg min_w \frac{1}{2} \|DW - B\|_F^2 + \frac{\rho}{2}\langle W, W\rangle$$

$$\Leftrightarrow \quad \arg min_w \frac{1}{2} \|D^K W^K - B\|_F^2 + \frac{\rho}{2}\langle W^K, W^K\rangle$$

$$\rightarrow (P^K) \quad \arg min_w \frac{1}{2} \|D^K W^K - B\|_F^2 + \frac{\rho}{2} trace((W^K)^T K W^K$$

2- Un algorithme « alternatif » pour les SVM

On a l'ensemble C tel que :

$$\mathcal{C} =: \left\{(w,b) \in \mathbb{R}^n \times \mathbb{R} \,|\, w^T u_i - b \geq 1, -w^T v_j + b \geq 1, \forall i \in [\![1,p]\!], \forall j \in [\![1,q]\!]\right\}$$

On remarque immédiatement que l'ensemble est fermé, borné et non-vide et composé de contraintes de fonctions linéaires, il est donc convexe.

On résume la formulation de SVM au problème P2 suivant

$$(P_2) : \begin{cases} \min \dfrac{1}{2}\|w\|^2 \\ \text{Sous les contraintes :} \begin{cases} w^T u_i - b \geq 1 & \forall i \in [\![1,p]\!] \\ -w^T v_j + b \geq 1 & \forall j \in [\![1,q]\!] \end{cases} \end{cases}$$

On a $J(w,b) = \frac{1}{2}\|w\|^2$ d'où $\nabla J(w) = w$

On pose $(w, w') \in \mathbb{R}^n \times \mathbb{R}^n$ d'où $\langle \nabla J(w) - \nabla J(w'), w - w' \rangle = \|w - w'\|_2^2$

On a donc J qui est K-lipschitzienne pour K=1 donc 1-lipschitzienne.

De fait le problème est K-lipschitzienne, 1-lipschitzienne et fait parti d'un ensemble de contraintes convexe, il existe donc une solution au problème P2.

On peut réécrire les contraintes sous forme matricielle telle que

$$\Leftrightarrow \begin{pmatrix} u_1 & -1 \\ \vdots & \vdots \\ u_p & -1 \\ -v_1 & 1 \\ \vdots & \vdots \\ -v_q & 1 \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \geq \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

On peut donc identifier la matrice contrainte C telle que C = $\begin{pmatrix} -u_1 & 1 \\ \vdots & \vdots \\ -u_p & 1 \\ v_1 & -1 \\ \vdots & \vdots \\ v_q & -1 \end{pmatrix}$ et de fait réécrire les contraintes sous la forme suivante :

$$\Leftrightarrow C \begin{pmatrix} w \\ b \end{pmatrix} + 1_{p+q} = O_{\mathbb{R}^{p+q}}$$

On pose les expressions des matrices suivantes A = (I_{p+q} $O_{\mathbb{R}^{p+q}}$) et x = $\binom{w}{b}$, on peut alors réécrire le lagrangien de la forme :

$$\mathcal{L}(x, z, \lambda) = \frac{1}{2}\|Ax\|^2 + \lambda^T\left(Cx + z + 1_{p+q} \right) + \frac{\rho}{2}\left\| Cx + z + 1_{p+q} \right\|_2^2$$

On détermine les dérivées partielles du lagrangien suivantes en posant d = 1_{p+q} :

- $\frac{\partial}{\partial x}\mathcal{L}(x, z, \lambda) = Ax + \lambda C^T + \rho C^T(C_x + z + d) = (A + \rho C^T C)x + C^T(\lambda + \rho(z + d))$

- $\frac{\partial}{\partial z}\mathcal{L}(x, z, \lambda) = \lambda + 2 \times \frac{\rho}{2}(Cx + z + d)$