Algorithmes d'optimisation – L3 MINT/DL/HEC – Université Paris-Sud

TP 2: Régression logistique

Ce TP étudie les aspects numériques de la régression logistique, un outil très utilisé pour des tâches de classification de données. On considère un jeu de données constitué de points $x_a \in \mathbb{R}^d$, $a \in A = \{1, \dots, n\}$, que l'on suppose regroupés en deux catégories disjointes $A = A_0 \sqcup A_1$. On posera $y_a = 0$ si $a \in A_0$ et $y_a = 1$ si $a \in A_1$. L'objectif de la classification supervisée est de construire une fonction $u : \mathbb{R}^d \to \mathbb{R}$ permettant d'estimer la catégorie d'un point: on souhaite que $u(x_a) \simeq y_a$ pour tout $a \in A$. Dans la régression logistique, on cherchera une fonction de la forme[1] $u_w(x) = \sigma(\langle w|x\rangle)$ où $w \in \mathbb{R}^d$ est un vecteur et où $\sigma : t \in \mathbb{R} \mapsto e^t/(1+e^t)$ est la sigmoïde.

Problème d'optimisation Informellement, le problème de régression logistique consiste à trouver $w \in \mathbb{R}^d$ tel que pour tout $a \in A$, $\sigma(\langle w|x_a\rangle) \simeq y_a$. On le formalise par le problème d'optimisation suivant, d'inconnue $w \in \mathbb{R}^d$:

$$\min_{w\in\mathbb{R}^d} F(w)$$
$$\text{où } F(w) = \sum_{a\in A} f_a(w) + \frac{\gamma}{2}\|w\|^2 \text{ et } f_a(w) = \begin{cases} -\log(1-\sigma(\langle w|x_a\rangle) \text{ si } a \in A_0 \\ -\log(\sigma(\langle w|x_a\rangle)) \text{ si } a \in A_1 \end{cases} \tag{1}$$

En pratique, une fois trouvé le minimiseur w^* de (1), on a fini la phase d'apprentissage, et on pourra utiliser $u^* := u_{w^*}$ pour estimer la catégorie de points ne faisant pas partie du jeu de données. La figure 1 représente un exemple de données $(x_a)_{a\in A}$ et la fonction u^* reconstruite.

Q0. Montrer que $\sigma(t) \in]0,1[$ pour tout $t \in \mathbb{R}$, $\lim_{t\to-\infty}\sigma(t) = 0$ et $\lim_{t\to\infty}\sigma(t) = 1$, et interpréter le problème d'optimisation (1) lorsque $\gamma = 0$.

1 Étude théorique du problème régularisé ($\gamma > 0$)

1.1 Existence et unicité

Q1.[Existence de solutions] Montrer que dans (1), $f_a \geq 0$ pour tout $a \in A$. En déduire que $F(w) \geq \frac{\gamma}{2}\|w\|^2$, puis que F admet un minimiseur sur $\mathbb{R}^d$.

Q2.[Convexité et unicité]

1. Soit $g : \mathbb{R} \to \mathbb{R}$ convexe et $x \in \mathbb{R}^d$. Montrer que $f : w \in \mathbb{R}^d \mapsto g(\langle w|x\rangle)$ est convexe.
2. Soit $g : t \mapsto -\log(1-\sigma(t))$. Montrer que $\sigma'(t) = \sigma(t)(1-\sigma(t))$, en déduire que $g''(t) \geq 0$, que g est convexe et que pour tout $a \in A_0$, f_a est convexe.
   (On montrerait de même que f_a est convexe $\forall a \in A_1$)
3. En déduire que F est γ-fortement convexe *(Ceci implique alors: $\forall w \in \mathbb{R}^d$, $\mathrm{D}^2F(w) \succeq \gamma\mathrm{Id}$.)*
4. Démontrer que le minimiseur de F sur $\mathbb{R}^d$ est unique.

On notera désormais w^* l'unique minimiseur de F.

[1] Les fonctions de la forme $x \mapsto u_w(x)$ sont bien adaptées pour approcher nos données $(x_a, y_a)_{a\in A}$ où $y_a \in \{0,1\}$, car elles vérifient automatiquement $u_w(x) \in]0,1[$ (au contraire des polynômes, qui, lorsqu'ils sont non constant sont aussi non-bornés!).

1.2 Calcul du gradient et de la hessienne

Q3.[Calcul du gradient] Montrer que

$$\begin{aligned}\nabla f_a(w) &= (\sigma(\langle w|x_a\rangle) - y_a)x_a \\ \nabla F(w) &= \sum_{a\in A}(\sigma(\langle w|x_a\rangle) - y_a)x_a + \gamma w,\end{aligned} \tag{2}$$

où l'on rappelle que $y_a = 0$ si $a \in A_0$, $y_a = 1$ si $a \in A_1$ et $\sigma'(t) = \sigma(t)(1 - \sigma(t))$.

Q4.[Calcul de la Hessienne]

1. Soit $x \in \mathbb{R}^d$ un vecteur colonne. Montrer que $A = xx^T$ est une matrice carrée symétrique positive dont les entrées sont $(xx^T)_{ij} = x_i x_j$, puis que $\langle Av|v\rangle = \langle x|v\rangle^2$.
2. Soit $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, et $G(w) := (\sigma(\langle w|x\rangle) - y)x$. Montrer que

$$\forall i, j \in \{1, \dots, d\}, \quad \frac{\partial G_i}{\partial w_j}(w) = \sigma(\langle w|x\rangle)(1 - \sigma(\langle w|x\rangle))x_i x_j,$$

   où l'on a noté $G_i(w)$ la $i$ème coordonnée de $G(w)$. En déduire que

$$\begin{aligned}\mathrm{D}^2 f_a(w) &= \sigma(\langle w|x_a\rangle)(1 - \sigma(\langle w|x_a\rangle))x_a x_a^T, \\ \mathrm{D}^2 F(w) &= \sum_{a\in A}\sigma(\langle w|x_a\rangle)(1 - \sigma(\langle w|x_a\rangle))x_a x_a^T + \gamma I_d.\end{aligned} \tag{3}$$

Q5.[Convergence de la descente de gradient]

1. Montrer que, pour $w^{(0)} \in \mathbb{R}^d$, l'ensemble $S = \{w \in \mathbb{R}^d \mid F(w) \leq F(w^{(0)})\}$ est compact.
2. En utilisant(3), montrer que pour tout $v \in \mathbb{R}^d$,

$$\langle \mathrm{D}^2 f_a(w)v|v\rangle \leq \|x_a\|^2 \|v\|^2 .$$

   En déduire que pour tout $w \in \mathbb{R}^d$, $\langle \mathrm{D}^2 F(w)v|v\rangle \leq M \|v\|^2$, où

$$M = \left(\sum_{a\in A}\|x_a\|^2\right) + \gamma.$$

3. Déduire de ce qui précède que $\forall w \in \mathbb{R}^d$, $\gamma \mathrm{Id} \leq \mathrm{D}^2 F(w) \leq M\mathrm{Id}$.
4. En conclure que les itérées de l'algorithme de descente de gradient à pas optimal (resp. avec backtracking d'Armijo) convergent vers w^*.

Q6.[Convergence de la méthode de Newton] La norme d'opérateur d'une matrice carrée A est définie par $|||A||| = \sup_{\|v\|=1} \|Av\|$. On admettra que pour A symétrique, $|||A||| = \sup_{\|v\|=1}\langle Av|v\rangle$.

1. On pose $A = \mathrm{D}^2 F(w_0) - \mathrm{D}^2 F(w_1)$. Démontrer que

$$\langle Av|v\rangle \leq \left(\sum_{a\in A}|\sigma(\langle w_0|x_a\rangle)(1 - \sigma(\langle w_0|x_a\rangle)) - \sigma(\langle w_1|x_a\rangle)(1 - \sigma(\langle w_1|x_a\rangle))|\,\|x_a\|^2\right)\|v\|^2$$

2. Démontrer que la fonction $g_a(w) := \sigma(\langle w|x_a\rangle)(1 - \sigma(\langle w|x_a\rangle))$ est de classe $\mathcal{C}^1$. En déduire qu'il existe une constante $\widetilde{L}_a \geq 0$ telle que g_a est $\widetilde{L}_a$-lipschitzienne sur S.
3. Déduire que la fonction $w \mapsto \mathrm{D}^2 F(w)$ est Lipschitzienne sur S.
4. Conclure à la convergence de la méthode de Newton.

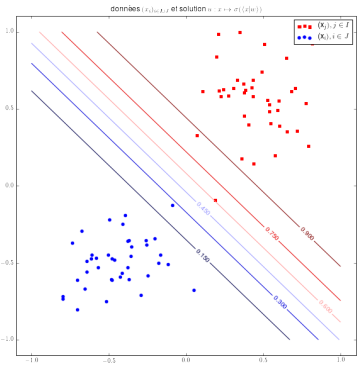

Figure 1: Les points $(x_a)_{a\in A_0}$ sont représentés par des carrés, les points $(x_a)_{a\in A_1}$ par des disques. On trace par dessus des lignes de niveau de la fonction $u_{w^*} : x \mapsto \langle x|w^*\rangle$ où w^* minimise (1).

2 Étude théorique du problème non régularisé ($\gamma = 0$)

On suppose dorénavant que $\gamma = 0$. Il n'est alors plus évident que F est strictement convexe ni même qu'elle tend vers $+\infty$ lorsque $\|w\| \to +\infty$. On doit donc rajouter une hypothèse sur les points $(x_a)_{a\in A}$. Dans la suite, on l'hypothèse que tout point de $\mathbb{R}^d$ peut être écrit comme une combinaison linéaire à coefficients positifs des points $(x_a)_{a\in A_0}$:

$$\forall w \in \mathbb{R}^d, \exists (\lambda_a)_{a\in A_0} \text{ tq } \lambda_a \geq 0 \text{ et } w = \sum_{a\in A_0} \lambda_a x_a \tag{4}$$

On pose $C = F(w^{(0)})$ où $w^{(0)} \in \mathbb{R}^d$ et $S = \{w \in \mathbb{R}^d \mid F(w) \leq C\}$. Pour montrer que le minimum dans (1) est atteint, il suffit donc de démontrer que S est compact.

Q7.[Existence]

1. En utilisant la positivité des f_a, démontrer que si $w \in S$, alors $f_a(w) \leq C$. En déduire que

$$\forall a \in A_0, -\log\left(\frac{1}{1+e^{\langle w|x_a\rangle}}\right) \leq C$$

   En utilisant la décroissance de $-\log$, en déduire que $\forall a \in A_0, \langle w|x_a\rangle \leq C$.
2. (**) Montrer que l'hypothèse (4) implique que

$$\min_{\|w\|=1} \max_{a\in A_0} \langle w|x_a\rangle > 0. \tag{5}$$

   Indication: Poser $\phi(w) = \max_{a\in A_0}\langle w|x_a\rangle$, qui est continue, et raisonner par l'absurde. Montrer que si (5) est fausse, alors il existe $w^* \in \mathbb{R}^d$ tel que $\|w^*\| = 1$ et $\phi(w^*) \leq 0$. En utilisant (4), écrire $w^* = \sum_i \lambda_i x_i$ où $\lambda_i \geq 0$. Démontrer qu'alors $\langle w^*|w^*\rangle \leq 0$, conclure.
3. Déduire de (5) qu'il existe $\kappa > 0$ tel que $\forall w \in \mathbb{R}^d, \exists a \in A, \langle w|x_a\rangle \geq \kappa \|w\|$.
4. Conclure que $S \subseteq \{w \in \mathbb{R}^d \mid \|w\| \leq C/\kappa\}$ est compact, puis que F admet un minimiseur.

Q8.[Forte convexité, $\mathrm{D}^2F > 0$] En utilisant la formule (3) et l'hypothèse (4), démontrer que $\mathrm{D}^2F(w)$ est symétrique définie positive en tout point $w \in \mathbb{R}^d$.

Q9.[Convergence des algorithmes] En déduire la convergence des méthode de descente de gradient et de Newton avec rebroussement d'Armijo.