Algorithmes d'optimisation – L3 MINT/DL 2017/2018 – Université Paris-Sud

TP 3: Régression logistique

Motivation Ce TP étudie les aspects numériques de la régression logistique, un outil statistique très utilisé en apprentissage automatique. On considère un jeu de données constitué de points $x_i \in \mathbb{R}^d$, $1 \leq i \leq n$, que l'on suppose découpé en deux classes

$$\{1, \dots, n\} = I \cup J,$$

où l'union est disjointe (cf Figure 1). Le problème de l'apprentissage est de construire ("apprendre") une fonction $u : \mathbb{R}^d \to \mathbb{R}$ permettant de "deviner" la classe d'un point: on souhaite que $u(x_i) = 0$ si $i \in I$ et $u(x_i) = 1$ si $i \in J$. Pour rendre le problème intéressant, il faut choisir une classe de fonctions dans laquelle choisir u. Dans la régression logistique, on suppose que la fonction u est de la forme $u(x) = \sigma(\langle w|x\rangle)$ où $w \in \mathbb{R}^d$ et où σ est la sigmoïde:

$$\sigma(t) = \frac{e^t}{1+e^t} \tag{1}$$

Elle vérifie $\sigma(t) \in]0,1[$ pour tout $t \in \mathbb{R}$, $\lim_{t\to-\infty} \sigma(t) = 0$ et $\lim_{t\to\infty} \sigma(t) = 1$. Le problème de régression logistique consiste donc à trouver $w \in \mathbb{R}^d$ tel que $\sigma(\langle w|x_i\rangle) \simeq 0$ si $i \in I$ et $\sigma(\langle w|x_i\rangle) \simeq 1$ si $i \in J$, et pour cela on passe par un problème d'optimisation.

$$\max_{w\in\mathbb{R}^d} \left(\prod_{i\in I}(1-\sigma(\langle w|x_i\rangle))\right)\left(\prod_{i\in J}\sigma(\langle w|x_i\rangle)\right). \tag{2}$$

L'idée est la suivante. Pour $i \in I$, on souhaite que $\sigma(\langle w|x_i\rangle) \simeq 0$, ou de manière équivalente, que $1 - \sigma(\langle w|x_i\rangle)$ soit aussi grand que possible. Réciproquement, pour $j \in J$, on veut que $\sigma(\langle w|x_j\rangle) \simeq 1$, donc aussi grand que possible. En passant au logarithme, on réécrit ce problème sous la forme d'un problème d'optimisation sans contrainte (3).

Problème d'optimisation On en vient finalement au problème d'optimisation suivant: où l'inconnu est $w \in \mathbb{R}^d$ et les données sont $(x_i)_{i\in I\cup J}$:

$$\min_{w\in\mathbb{R}^d} F(w)$$
$$\text{où } F(w) = \sum_{i\in I\cup J} f_i(w) + \frac{\gamma}{2}\|w\|^2 \text{ et } f_i(w) = \begin{cases} -\log(1-\sigma(\langle w|x_i\rangle)) \text{ si } i \in I \\ -\log(\sigma(\langle w|x_i\rangle)) \text{ si } i \in J \end{cases} \tag{3}$$

Une fois trouvé le minimum w^* de (3), on a fini la phase d'apprentissage, et on peut utiliser $u(x) := \sigma(\langle x|w^*\rangle)$ pour classer de nouveaux points. On espère que $u(x_i) \simeq 0$ si $i \in I$ et $u(x_j) \simeq 1$ si $i \in J$. (Si ça n'est pas le cas, alors la classe de fonction utilisée n'est pas assez riche pour traiter ce problème). On peut alors se servir de cette fonction pour "deviner" la classe d'un point $x \in \mathbb{R}^d$ ne faisant pas partie de l'échantillon (voir Figure 1). Nous verrons en TP comment se servir de cette approche pour un problème d'apprentissage supervisé en reconnaissance d'écriture (distinguer des chiffres 1 et 0 manuscrits).

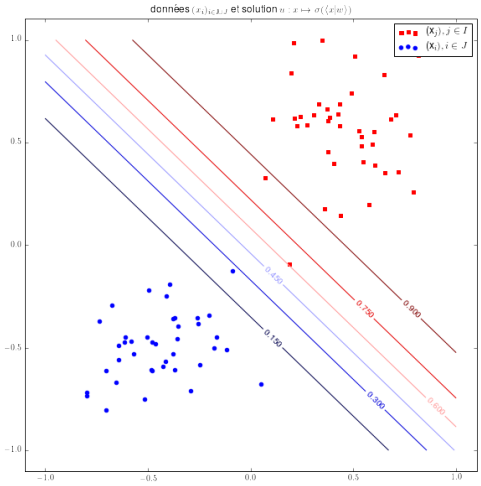

Figure 1: Les points $(x_i)_{i\in I}$ sont représentés par des carrés, les points $(x_i)_{i\in J}$ par des disques. On trace par dessus quelques lignes de niveau de la fonction $u : x \mapsto \langle x|w^*\rangle$ où w^* minimise (3).

Remarque Dans la définition de F dans (3), on a rajouté un terme $\gamma \|\cdot\|^2$. Ce terme rend l'étude théorique un peu plus simple, et peut aussi être utile en pratique lorsque $d \gg 1$.

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

1.1 Existence et unicité

Q1.[Existence de solutions] Montrer que dans (3), $f_i \geq 0$ pour tout $i \in I \cup J$. En déduire que $F(w) \geq \frac{\gamma}{2}\|w\|^2$, puis que F atteint son minimum sur $\mathbb{R}^d$.

Q2.[Convexité et unicité]

1. Montrer que la composition d'une fonction convexe (de $\mathbb{R}^k \to \mathbb{R}$) et d'une fonction affine (de $\mathbb{R}^d \to \mathbb{R}^k$) est convexe.
2. Soit $g : t \mapsto -\log(1-\sigma(t))$. Montrer que $g''(t) \geq 0$, en déduire que pour tout $i \in I$, f_i est convexe. (Et on montrerait exactement pareil que f_j est convexe pour $j \in J$)
3. En déduire que pour tout $w \in \mathbb{R}^d$, $\mathrm{D}^2F(w) \geq \gamma$ (ce qui, comme toujours dans ce cours, signifie $\forall v \in \mathbb{R}^d, \langle \mathrm{D}^2F(w)v|v\rangle \geq \gamma\|v\|^2$.)
4. Démontrer que le minimiseur de F est unique. *(Indication: utiliser une propriété du cours sur les fonctions vérifiant* $\mathrm{D}^2F \geq m > 0$.*)*

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

1.2 Gradient et hessienne

Q3.[Calcul du gradient] Montrer que $\sigma'(t) = \sigma(t)(1-\sigma(t))$ et en déduire que

$$\begin{aligned} &\forall i \in I,\ \nabla f_i(w) = \sigma(\langle w|x_i\rangle)x_i \qquad \forall j \in J,\ \nabla f_j(w) = (\sigma(\langle w|x_j\rangle) - 1)x_j \\ &\nabla F(w) = \sum_{i\in I} \sigma(\langle w|x_i\rangle)x_i + \sum_{j\in J}(\sigma(\langle w|x_j\rangle) - 1)x_j + \gamma w \end{aligned} \tag{4}$$

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 et de rang 1 dont les entrées sont $(xx^T)_{kl} = x_k x_l$. Calculer $\langle Av|v\rangle$.
2. Soit $G(w) = \sigma(\langle w|x_i\rangle)x_i$. Montrer que $\mathrm{D}G(w) = \sigma(\langle w|x_i\rangle)(1-\sigma(\langle w|x_i\rangle))x_i x_i^T$ et que

$$\begin{aligned} &\forall i \in I \cup J, \quad \mathrm{D}^2 f_i(w) = \sigma(\langle w|x_i\rangle)(1-\sigma(\langle w|x_i\rangle))x_i x_i^T, \\ &\mathrm{D}^2 F(w) = \sum_{i\in I\cup J} \sigma(\langle w|x_i\rangle)(1-\sigma(\langle w|x_i\rangle))x_i x_i^T + \gamma I_d. \end{aligned} \tag{5}$$

1.3 Convergence de deux méthodes de descente

On rappelle la méthode de descente de gradient (6) et de Newton (7) avec backtracking d'Armijo:

$$\begin{cases} d^{(k)} = -\nabla F(w^{(k)}) \\ t^{(k)} = \arg\max\{t \mid \exists k \in \mathbb{N},\ t = \beta^k, f(w^{(k)} + td^{(k)}) \leq f(w^{(k)}) + \alpha\langle d^{(k)}|\nabla f(w^{(k)})\rangle \\ w^{(k+1)} = w^{(k)} + t^{(k)}d^{(k)} \end{cases} \tag{6}$$

$$\begin{cases} d^{(k)} = -D^2F(w^{(k)})^{-1}\nabla F(w^{(k)}) \\ t^{(k)} = \arg\max\{t \mid \exists k \in \mathbb{N},\ t = \beta^k, f(w^{(k)} + td^{(k)}) \leq f(w^{(k)}) + \alpha\langle d^{(k)}|\nabla f(w^{(k)})\rangle \\ w^{(k+1)} = w^{(k)} + t^{(k)}d^{(k)} \end{cases} \tag{7}$$

La norme d'opérateur d'une matrice carrée A est définie par $|||A||| = \sup_{\|v\|=1} \|Av\|$. Si A est symétrique, alors on a $|||A||| = \sup_{\|v\|=1}\langle Av|v\rangle$.

Théorème 1. *Soit $\Omega \subseteq \mathbb{R}^d$ ouvert convexe, $F \in \mathcal{C}^2(\Omega)$, et $w^{(0)} \in \Omega$ vérifiant*

1. *$S = \{w \in \Omega \mid f(w) \leq f(w^{(0)})\}$ est compact.*
2. *$\exists M \geq m > 0$ tel que $\forall w \in S, \quad m \leq \mathrm{D}^2F(w) \leq M$.*

Alors la suite définie par (6) converge vers l'unique minimum de F sur Ω. Si de plus

3. *$\exists L \geq 0$ tel que l'application $w \in S \mapsto \mathrm{D}^2F(w)$ est L-lipschitzienne, i.e.*

$$\forall w_0, w_1 \in S, \left|\left|\left|\mathrm{D}^2F(w_0) - \mathrm{D}^2F(w_1)\right|\right|\right| \leq L\|w_0 - w_1\|$$

Alors la suite définie par (7) converge également vers l'unique minimum de F sur Ω.

Q5.[Convergence de la descente de gradient]

1. Montrer que, pour tout $w^{(0)} \in \mathbb{R}^d$, l'ensemble $S = \{w \in \mathbb{R}^d \mid F(w) \leq F(w^{(0)})\}$ est compact.
2. En utilisant(5), montrer que pour tout $v \in \mathbb{R}^d$,

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

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

$$M = \left(\sum_{i\in I\cup J} \|x_i\|^2\right) + \gamma.$$

3. En déduire que $\gamma \leq \mathrm{D}^2F(w) \leq M$, puis la convergence de la suite (6) vers w^*.

Q6.[Convergence de la méthode de Newton]

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

$$\langle Av|v\rangle \leq \left(\sum_{i\in I\cup J} |\sigma(\langle w_0|x_i\rangle)(1-\sigma(\langle w_0|x_i\rangle)) - \sigma(\langle w_1|x_i\rangle)(1-\sigma(\langle w_1|x_i\rangle))| \, \|x_i\|^2 \right) \|v\|^2$$

2. Démontrer que la fonction $g_i(w) := \sigma(\langle w|x_i\rangle)(1-\sigma(\langle w|x_i\rangle))$ est de classe $\mathcal{C}^1$. En déduire qu'il existe une constante $\widetilde{L}_i \geq 0$ telle que g_i est $\widetilde{L}_i$-lipschitzienne sur S.
3. Conclure à la convergence de la méthode de Newton.

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 $(x_i)_{i\in I\cup J}$. Dans la suite, on *fera l'hypothèse* que tout point de $\mathbb{R}^d$ peut être écrit comme une combinaison linéaire à coefficients positifs des points $(x_i)_{i\in I}$:

$$\forall w \in \mathbb{R}^d, \exists (\lambda_i)_{i\in I} \text{ tq } \lambda_i \geq 0 \text{ et } w = \sum \lambda_i x_i \tag{8}$$

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 (3) est atteint, il suffit donc de démontrer que S est compact.

Q7.[Existence]

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

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

   En utilisant la décroissance de $-\log$, en déduire que $\forall i \in I, \langle w|x_i\rangle \leq C$.
2. (**) Montrer que l'hypothèse (8) implique que

$$\min_{\|w\|=1} \max_{i\in I} \langle w|x_i\rangle > 0. \tag{9}$$

   Indication: Poser $\phi(w) = \max_{i\in I}\langle w|x_i\rangle$, qui est continue, et raisonner par l'absurde. Montrer que si (9) est fausse, alors il existe $w^* \in \mathbb{R}^d$ tel que $\|w^*\| = 1$ et $\phi(w^*) \leq 0$. En utilisant (8), 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 (9) qu'il existe $\kappa > 0$ tel que $\forall w \in \mathbb{R}^d, \exists i \in I, \langle w|x_i\rangle \geq \kappa \|w\|$.
4. Conclure que $S \subseteq \{w \in \mathbb{R}^d \mid \|w\| \leq C/\kappa\}$ est compact, puis montrer que F atteint son minimum sur $\mathbb{R}^d$.

Q8.[Forte convexité, $\mathrm{D}^2F > 0$] En utilisant la formule (5) et l'hypothèse (8), démontrer que $\mathrm{D}^2F(w)$ est symétrique définie positive en tout point $w \in \mathbb{R}^d$, puis qu'il existe $m \geq 0$ tel que

$$\forall w \in S, \quad \mathrm{D}^2F(w) \geq m.$$

Q9.[Convergence des algorithmes] En déduire comme précédemment la convergence des méthode de descente de gradient et de Newton avec backtracking.