Examen Master M2A
Analyse numérique et réseaux de neurones

15/04/2021

Tous documents autorisés.
Le sujet est constitué de deux problèmes indépendants.

---

1 Un contrôle théorique de l'overfitting

• Soit une fonction $f : \mathbb{R}^m \to \mathbb{R}^n$ encodée dans un réseau de neurones à p couches cachées

$$f = f_p \circ g_{p-1} \cdots \circ g_2 \circ g_1 \circ g_0, \qquad \text{avec } g_r = R \circ f_r \text{ pour } R = \text{fonction ReLU} . \quad (1)$$

Les fonctions f_r sont linéaires sous la forme

$$f_r(x_r) = W_r x_r + b_r, \quad x_r \in \mathbb{R}^{a_r}, \quad W_r \in \mathcal{M}_{a_{r+1},a_r}(\mathbb{R}).$$

• Pour un vecteur $z = (z_1, z_2, \ldots, z_q) \in \mathbb{R}^q$ de taille arbitraire, on utilisera la norme l^∞: $\|z\| = \max_{1 \leqslant i \leqslant q} |z_i|$. La norme induite pour une matrice rectangulaire $W \in \mathcal{M}_{r,q}$ est

$$\|W\| = \max_{z \neq 0} \frac{\|Wz\|}{\|z\|}.$$

• On dira que fonction $f : \mathbb{R}^m \to \mathbb{R}^n$ est Lipschitz de constante K ssi

$$\|f(x+d) - f(x)\| \leqslant K\|d\|, \quad \forall x, d \in \mathbb{R}^m.$$

Nous allons discuter d'un outil théorique de contrôle de l'overfitting qui s'appuie sur cette constante K.

1. Montrer que la fonction ReLU est Lipschitz de constante $K = 1$, c'est à dire que $\|R(x+d) - R(x)\| \leqslant \|d\|$.

   Indication: on pourra commencer par le cas $x \in \mathbb{R}$, c'est à dire $m = 1$, puis passer au cas général $m > 1$.

2. Soit une fonction sans couche cachée $f_0(x) = W_0 x + b_0$. Montrer que $K \leqslant \|W_0\|$.

3. Soit à présent une fonction avec exactement une couche cachée $f_1(x) = W_1 R(W_0 x + b_0) + b_1$. Montrer que $K \leqslant \|W_1\| \, \|W_0\|$.

4. Soit la fonction f décrite en (1). Montrer que

$$K \leqslant \Pi_{i=0}^{p} \|W_i\|.$$

   Indication: on pourra utiliser une raisonnement par récurrence.

5. Soit une fonction objectif $f^{\text{obj}} : \mathbb{R}^m \to \mathbb{R}^n$ que l'on supposera différentiable avec

$$\|\nabla f^{\text{obj}}\| = \sup_{x \in \mathbb{R}^m} \left\|\nabla f^{\text{obj}}(x)\right\| < \infty.$$

   On construit un dataset idéal sans bruit

$$\mathcal{D} = \{(x_i, y_i), \ 1 \leqslant i \leqslant N\}, \qquad x_i \in [0,1]^m, \ y_i = f^{\text{obj}}(x_i).$$

   Pour la fonction f de (1), on définit la RMSE (root mean square error)

$$\varepsilon = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \|f(x_i) - y_i\|^2}$$

   et l'erreur en norme du maximum $\varepsilon_{\max} = \sup_{i=1}^{N} \|f(x_i) - y_i\|$.
   Montrer les inégalités $\frac{1}{\sqrt{N}} \varepsilon_{\max} \leqslant \varepsilon \leqslant \varepsilon_{\max}$.

6. Soit un point $z \in [0,1]^m$ a priori en dehors du dataset, c'est à dire que $z \neq x_i$ pour tout i.

   Montrer l'estimation

$$\|f(z) - f^{\text{obj}}(z)\| \leqslant \left(K + \|\nabla f^{\text{obj}}\|\right) \|z - x_i\| + N^{\frac{1}{2}} \varepsilon, \quad 1 \leqslant i \leqslant N.$$

7. Faisons l'hypothèse: plus la précision de training est bonne (c'est à dire plus ε est petit) plus la constante K est grande, par exemple avec une loi $K \approx C\varepsilon^{-\alpha}$ avec $\alpha > 0$ et $C > 0$.

   Discuter d'un ordre de grandeur raisonnable en fonction de ε que K ne devrait pas dépasser (pour les petits ε).
   Remarque: ce phénomène correspond à de l'overfitting.

2 Descente de gradient continue avec LASSO

Les méthodes LASSO (Least Absolute Shrinkage and Selection Operator) sont utilisées pour contraindre la recherche du minimum d'un fonction

$$J_0 : \mathbb{R}^p \to \mathbb{R}.$$

Nous supposerons que $J_0 \in C^2(\mathbb{R}^p)$ et que les dérivées secondes sont bornées uniformément

$$\sup_{W\in\mathbb{R}^p} \left\|\nabla^2 J_0\right\| \leqslant C < \infty.$$

Cela revient à dire que ∇J_0 est Lipschitz de constante $C > 0$.

Nous modifions la fonction J_0 avec un LASSO de paramètre $\alpha > 0$. La fonction modifiée J est

$$J(W) = J_0(W) + J_1(W) \text{ où } J_1(W) = \alpha \sum_i |w_i|.$$

La question qui se pose est de donner un sens clair à l'équation différentielle ordinaire

$$\frac{d}{dt}W(t) = -\nabla J(W(t)) \tag{2}$$

qui est la base des méthodes de descente de gradient.

1. Expliquer pourquoi ∇J_1 n'est pas une fonction Lipschitzienne. Le théorème de Cauchy-Lipschitz s'applique-t-il au système (2)?

2. Soit la fonction régularisée (paramètre $\mu > 0$)

$$J_1^\mu(W) = \alpha \sum_i \sqrt{w_i^2 + \mu}, \qquad \mu > 0.$$

   Montrer que $\lim_{\mu\to 0^+} \sqrt{w^2+\mu} = |w|$ et que $\lim_{\mu\to 0^+} J_1^\mu(W) = J_1(W)$.

   Montrer que J_1^μ est une fonction dont les dérivés secondes sont continues. Montrer que J_1^μ est convexe (indication: montrer que $\frac{d^2}{dx^2}\sqrt{x^2+\mu} > 0$).

3. On pose $J^\mu(Y) = J_0(Y) + \alpha J_1^\mu(Y)$ pour tout Y. Soit une solution $W_\mu(t)$ de

$$\frac{d}{dt}W_\mu(t) = -\nabla J_\mu(W_\mu(t)).$$

   Montrer l'inégalité pour tout $Y \in \mathbb{R}^p$

$$\left\langle \frac{d}{dt}W_\mu(t) - \nabla J_0(W_\mu(t)), Y - W_\mu(t) \right\rangle + J_1^\mu(Y) - J_1^\mu(W_\mu(t)) \geqslant 0.$$

4. En passant à la limite formelle $\mu \to 0$, en déduire

$$\left\langle \frac{d}{dt}W(t) - \nabla J_0(W(t)), Y - W(t) \right\rangle + J_1(Y) - J_1(W(t)) \geqslant 0, \quad \forall Y \in \mathbb{R}^p. \tag{3}$$

   Quel pourrait être l'intérêt de cette formulation par rapport à (2)?