OPTIMISATION POUR L'APPRENTISSAGE

23 janvier 2025

Ce matin : Régularisation (cours, illustration numérique)
Exercice du TD 3 en ligne

RÉGULARISATION

Problème régularisé

$\lambda \geq 0$: poids du terme en $\Omega(w)$ par rapport au terme en $f(w)$

$$\underset{w \in \mathbb{R}^d}{\text{minimiser}} \quad f(w) + \lambda \Omega(w)$$

$\Omega(w)$ → Terme de régularisation $\Omega : \mathbb{R}^d \to \mathbb{R}$

⇒ Modélise des propriétés souhaitées pour la solution

$f(w)$: Terme d'attache aux données

(typiquement $f(w) = \frac{1}{m} \sum_{i=1}^{m} f_i(w)$)

⇒ Exprime la tâche d'apprentissage à réaliser (régression, classification, etc)

Exemples en régression linéaire

$X \in \mathbb{R}^{m \times d}$, $y \in \mathbb{R}^m$ But : Trouver $w \in \mathbb{R}^d$ tq $Xw \approx y$

$X = \begin{bmatrix} x_1^T \\ | \\ x_m^T \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ | \\ y_m \end{bmatrix}$

Problème non régularisé : $\underset{w \in \mathbb{R}^d}{\text{minimiser}} \quad \frac{1}{2m} \|Xw - y\|_2^2$

$$\|Xw-y\|_2^2 = \sum_{i=1}^{m} \left(x_i^T w - y_i\right)^2$$

→ Régularisation ℓ_2 (aka "ridge", "Tychonov", ...)

$$\underset{w \in \mathbb{R}^d}{\text{minimiser}} \quad \frac{1}{2m} \|Xw - y\|_2^2 + \lambda \|w\|_2^2$$

$$\|w\|_2^2 = \sum_{j=1}^{d} w_j^2$$

$\lambda = 0$ ⇒ Problème de départ

$\lambda \ll 1$: proche du problème de départ

$\lambda >> 1$: proche du problème

$$\left.\underset{w \in \mathbb{R}^d}{\text{minimiser}} \; \lambda \|w\|_2^2 \right) \Rightarrow \text{unique solution } w^* = 0_{\mathbb{R}^d}$$

Pourquoi régulariser avec la norme ℓ_2 ?

→ Améliorer la généralisation du modèle

$\Rightarrow$ Erreur $\|X^{test} w^* - y^{test}\|_2^2$

avec $w^* \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2n}\|Xw - y\|_2^2 + \lambda\|w\|_2^2$

et (X^{test}, y^{test}) provient de la même distribution de données que (X, y)

→ Rendre le problème fortement convexe

$w \mapsto \lambda \|w\|_2^2$ est 2λ-fortement convexe

$w \mapsto \frac{1}{2n}\|Xw - y\|_2^2$ est convexe

⊕ Garantie d'unicité de la solution

⊕ Convergence plus rapide des algorithmes type descente de gradient / gradient stochastique

→ Réduit la sensibilité de la solution par rapport aux données

→ On souhaite que la solution varie peu si les données (X, y) sont légèrement perturbées

(Ex) $n = d = 2$ $\quad X = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$$\underset{w \in \mathbb{R}^2}{\text{minimiser}} \; \frac{1}{2n}\|Xw - y\|_2^2 = \frac{1}{4}\left[(w_1 - 1)^2 + 1\right]$$

$$\underset{w \in \mathbb{R}^2}{\operatorname{argmin}} \ \frac{1}{2n}\|Xw-y\|^2 = \left\{ w = \begin{bmatrix} 1 \\ w_2 \end{bmatrix} \,\middle|\, w_2 \in \mathbb{R} \right\}$$

$\forall \lambda > 0,$

$$\underset{w \in \mathbb{R}^2}{\operatorname{minimiser}} \left\{ \frac{1}{2n}\|Xw-y\|_2^2 + \lambda \|w\|_2^2 \right\}$$

$$\shortparallel$$

$$\frac{1}{4}\left((w_1-1)^2+1 \right) + \lambda w_1^2 + \lambda w_2^2 \Big)$$

$\frac{1}{4}w_1^2 - \frac{1}{2}w_1 + \lambda w_1^2$

$\frac{1}{2}w_1 + 2\lambda w_1 - \frac{1}{2} = 0$

$w_1 = \frac{1}{1+4\lambda}$

$$\underset{w \in \mathbb{R}^2}{\operatorname{argmin}} \{ \text{---} \} = \left\{ \begin{bmatrix} \frac{1}{1+4\lambda} \\ 0 \end{bmatrix} \right\}$$

__Remarque__ : Dans le cas général $\underset{w \in \mathbb{R}^d}{\operatorname{minimiser}} \ f(w) + \lambda \|w\|_2^2$

avec f C^1, l'itération de descente de gradient sur ce problème donne

$$w_{k+1} = w_k - \alpha_k \left[\nabla f(w_k) + 2\lambda w_k \right]$$

$$\nabla\left(\lambda \|\cdot\|_2^2\right)(w_k) = 2\lambda w_k$$

$$= (1 - 2\lambda\alpha_k) w_k - \alpha_k \nabla f(w_k)$$

Pour $\alpha_k < \frac{1}{2\lambda}$, "weight decay" : on réduit l'amplitude des coefficients de w_k en plus de faire un pas de gradient

→ Régularisation ℓ_1 / LASSO

Pour la régression linéaire,

$$\underset{w \in \mathbb{R}^d}{\text{minimiser}} \ \frac{1}{2n}\|Xw - y\|_2^2 + \lambda \|w\|_1$$

$$\|w\|_1 = \sum_{j=1}^{d} |w_j|$$

Pourquoi la norme ℓ_1 ?

→ Favorise les solutions avec des coordonnées nulles
(cas extrême $\lambda >> 1$: $\underset{w \in \mathbb{R}^d}{\text{minimiser}} \ \lambda \|w\|_1$)
$\downarrow$
$w^* = 0_{\mathbb{R}^d}$

→ Lorsque λ augmente, la solution du problème a de plus en plus de coordonnées nulles, et les coordonnées sont "mises à 0" par ordre croissant d'importance

Ex)

$$\underset{w \in \mathbb{R}^2}{\text{minimiser}} \ \left[(w_1 - 1)^2 + (w_2 - 2)^2\right] \Rightarrow w^* = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$\underset{w \in \mathbb{R}^2}{\text{minimiser}} \ \left[(w_1 - 1)^2 + (w_2 - 2)^2\right] + \lambda \|w\|_2^2$$

$2w_1 - 2 + 2\lambda w_1 = 0$

$w_1 = \frac{1}{1+\lambda}$

$$\Rightarrow w^* = \begin{bmatrix} \frac{1}{1+\lambda} \\ \frac{2}{1+\lambda} \end{bmatrix} \xrightarrow[\lambda \to +\infty]{} \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\underset{w \in \mathbb{R}^2}{\text{minimiser}} \left[(w_1-1)^2 + (w_2-2)^2 \right] + \lambda \|w\|_1$$

$(w_1-1)^2 + \lambda|w_1|$

$2(w_1-1) \in \lambda \partial|\cdot|(w_1)$

$0 \leq \lambda < 2 \qquad w^* = \begin{bmatrix} 1 - \frac{\lambda}{2} \\ 2 - \frac{\lambda}{2} \end{bmatrix}$

$2 \leq \lambda < 4 \qquad w^* = \begin{bmatrix} 0 \\ 2 - \lambda/2 \end{bmatrix}$

$4 \leq \lambda \qquad w^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

Q) Existe-t-il un algorithme qui s'applique à tout problème de la forme $\underset{w \in \mathbb{R}^d}{\text{minimiser}}\ f(w) + \lambda \Omega(w)$?

<u>Hypothèses</u> :

- f C^1 (mais pas forcément convexe)
- Ω convexe (mais pas forcément dérivable)

Itération du gradient proximal :

Terme proximal : garantit que w est "proche" de w_k

$$w_{k+1} \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \underbrace{f(w_k) + \nabla f(w_k)^T (w - w_k)}_{\approx f(w) \text{ autour de } w_k} + \frac{1}{2\alpha_k} \|w - w_k\|_2^2 + \lambda \Omega(w) \right\}$$

Régularisation (inchangée)

avec $\alpha_k > 0$

Définition implicite : w_{k+1} est défini comme la solution d'un problème d'optimisation

Remarques importantes :

- Si $\Omega \equiv 0$ $\quad (\Omega(w) = 0 \ \forall w)$

  alors

  $$\underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(w_k) + \nabla f(w_k)^T (w - w_k) + \frac{1}{2\alpha_k} \|w - w_k\|_2^2 \right\}$$

  $$= w_k - \alpha_k \nabla f(w_k)$$

- Dans le cas général, l'algorithme du gradient proximal n'est intéressant que si le sous-problème résolu à chaque itération est plus facile à résoudre que le problème de départ !

  $\Rightarrow$ Pour les régularisations classiques $(\ell_2, \ell_1, \dots)$, on peut résoudre le sous-problème de manière explicite

  $\Rightarrow$ Cas $\Omega(w) = \|w\|_1$: l'algorithme du gradient proximal correspond à l'algorithme ISTA (Iterative Soft-Thresholding Algorithm)

Exo 2 TD 3

a) La régularisation sert à favoriser les vecteurs/modèles qui vérifient une propriété souhaitée (parcimonie, par exemple).

$\Leftrightarrow$ ______ pénaliser les vecteurs/modèles qui ne vérifient une propriété souhaitée.

d)

$$w_{k+1} \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \phi(w_k) + \nabla\phi(w_k)^T (w - w_k) + \frac{1}{2\alpha_k} \|w - w_k\|_2^2 + \frac{\lambda_2}{2}\|w\|_2^2 + \lambda_1 \|w\|_1 \right\}$$

$$\phi(w) = \frac{1}{2n} \|Xw - y\|_2^2$$

$$\Leftrightarrow w_{k+1} \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|Xw_k - y\|_2^2 + \frac{1}{n}(Xw_k - y)^T X (w - w_k) \right.$$

$$\left. + \frac{1}{2\alpha_k}\|w - w_k\|_2^2 + \frac{\lambda_2}{2}\|w\|_2^2 + \lambda_1\|w\|_1 \right\}$$

e) $\lambda_2 = 0$

$$w_{k+1} \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \phi(w_k) + \nabla\phi(w_k)^T (w - w_k) + \frac{1}{2\alpha_k}\|w - w_k\|_2^2 \right.$$

$$\left. + \lambda_1 \|w\|_1 \right\}$$

ISTA = gradient proximal pour ce problème avec formule explicite pour w_{k+1}

g) $\underset{w \in \mathbb{R}^d}{\text{minimiser}} \quad \underbrace{\phi(w_k) + \nabla\phi(w_k)^T(w - w_k) + \frac{1}{2\alpha_k}\|w - w_k\|_2^2 + \frac{\lambda_2}{2}\|w\|_2^2}_{f(w)} + \lambda_1 \|w\|_1$

$\lambda_2 > 0 \quad \lambda_1 > 0$

$w_k \in \mathbb{R}^d, \alpha_k > 0$

→ Sous-gradient ✓

→ ISTA ✓

↳ Problème : $\underset{w \in \mathbb{R}^d}{\text{minimiser}} \; f(w) + \lambda_1 \|w\|_1$

ISTA

$w^{(0)}, \ldots, w^{(j)}, w^{(j+1)}, \ldots$

↳ Sous-problème :

$$w^{(j+1)} \in \underset{\bar{w} \in \mathbb{R}^d}{\text{argmin}} \left\{ f(w^{(j)}) + \nabla f(w^{(j)})^T(\bar{w} - w^{(j)}) + \frac{1}{2\beta^{(j)}}\|\bar{w} - w^{(j)}\|_2^2 + \lambda_1 \|\bar{w}\|_1 \right\}$$

avec $\beta^{(j)} > 0$