Régression Linéaire et Généralisation

14 mars 2017

1 Contexte

1.1 Régression Linéaire, Erreur de Généralisation

Soit $f : E \subset [0,1]^d \rightarrow [0,1]$ une fonction inconnue avec E **fini**. On suppose que l'on connaît f en quelques points : On dispose de $D = \{(x_i, y_i) \,|\, y_i = f(x_i)\}$ fini avec $m = |D| << |E|$. Le but de la **régression linéaire** est d'approcher f au mieux avec f_θ, modèle linéaire, défini par :

$$f_\theta(x) = \theta^T x$$

Avec $\theta \in \mathbb{R}$ le paramètre de f_θ. Plus formellement on veut trouver θ qui minimise **l'erreur absolue** :

$$J(\theta) = \sum_{x \in E} (f(x) - f_\theta(x))^2$$

Cependant comme on ne connaît pas E, on doit se contenter d'approcher J par $\hat{J}$ **l'erreur empirique** sur D :

$$\hat{J}(\theta) = \sum_i (y_i - f_\theta(x_i))^2 = \sum_i (y_i - \theta^T x_i)^2$$

1) Que signifie $J(\theta) = 0$? $\hat{J}(\theta) = 0$?
2) Montrer

$$\hat{J}(\theta) \leq J(\theta)$$

3) Minimiser $\hat{J}(\theta)$ revient-il à minimiser $J(\theta)$?
Dans l'idéal on voudrait donc minimiser $J(\theta)$, cependant c'est $\hat{J}(\theta)$ que l'on

peut minimiser en pratique. Ce qui nous intéresserait ce serait trouver un terme $M > 0$ tel que :

$$J(\theta^*) \leq \hat{J}(\theta^*) + M$$

Avec θ^* qui minimise $\hat{J}$.
4) En quoi peut-on qualifier M **d'erreur de généralisation** ?

1.2 Régression Linéaire Ridge-Régularisée

Pour des raisons passées sous silence, on considère ici les minimisations de :

$$J_r(\theta) = J(\theta) + \lambda||\theta||_2^2$$
$$\hat{J}_r(\theta) = \hat{J}(\theta) + \lambda||\theta||_2^2$$

Avec λ un scalaire donné. On parle alors de **régression linéaire ridge-régularisée**. On admet qu'il existe un algorithme capable de fournir θ^* qui minimise $\hat{J}_r$. On cherche toujours une inégalité du type :

$$J_r(\theta^*) \leq \hat{J}_r(\theta^*) + M$$

2 Stabilité Uniforme

Dans la suite on note :

— θ qui minimise $\hat{J}_r$ sur D

— $\theta^{\backslash i}$ qui minimise $\hat{J}_r$ sur $D \backslash \{(x_i, y_i)\}$

Pour trouver la borne M on dispose du théorème suivant :
Théorème : S'il existe β tel que :

$$\forall i \in \{1 \dots m\}, \quad sup_{x \in E}|l(\theta, x) - l(\theta^{\backslash i}, x)| \leq \frac{\beta}{m}$$

Avec $l(\theta, x) = (\theta^T x - f(x))^2$ et $l(\theta^{\backslash i}, x) = (\theta^{\backslash} i^T x - f(x))^2$
Et si $l(\theta, x)$ **est bornée** par M
Alors :
Avec probabilité $1 - \delta$:

$$J_r(\theta) \leq \hat{J}_r(\theta) + 2\frac{\beta}{m} + (4\beta + M)\sqrt{\frac{ln(\frac{1}{\delta})}{2m}}$$

5) Que se passe-t-il lorsque m est grand ? Comment l'interprétez vous ?

3 Trouvons β

Nous avons toujours
— θ qui minimise $\hat{J}_r$ sur D
— $\theta^{\setminus i}$ qui minimise $\hat{J}_r$ sur $D\setminus\{(x_i, y_i)\}$

6) Montrer que $||\theta||_2^2 \leq \frac{1}{\lambda}$
7) En déduire $l(\theta, x) \leq 1$

On remplit donc la deuxième condition du théorème avec M=1. Passons à la première.

8) Montrer que l est convexe en son premier argument
9) Montrer qu'il existe $\sigma > 0$ tel que :

$$\forall \theta_1, \theta_2 \, \forall x \quad |l(\theta_1, x) - l(\theta_2, x)| \leq \sigma |(\theta_1 - \theta_2)^T x|$$

10) En déduire ce **lemme** :

$$\forall t \in [0,1] \quad ||\theta||_2^2 - ||\theta + t\Delta\theta||_2^2 + ||\theta^{\setminus i}||_2^2 - ||\theta^{\setminus i} - t\Delta\theta||_2^2 \leq \frac{t\sigma}{\lambda m} |\Delta\theta^T x_i|$$

Avec $\Delta\theta = \theta^{\setminus i} - \theta$.
11) Constater que pour $t = \frac{1}{2}$ on a $||\theta||_2^2 - ||\theta + t\Delta\theta||_2^2 + ||\theta^{\setminus i}||_2^2 - ||\theta^{\setminus i} - t\Delta\theta||_2^2 = \frac{1}{2}||\Delta\theta||_2^2$

12) En déduire une majoration de :

$$|l(\theta, x) - l(\theta^{\setminus i}, x)|$$

13) Conclure

4 Correction

Coming soon

5 Sources

— Sujet (adapté) : $http : //perso.univ{-}st{-}etienne.fr/habrarda/ENSML/td_uniform_stab$

— Correction : $http : //perso.univ{-}st{-}etienne.fr/habrarda/ENSML/recap_uniformStab$

— Papier d'origine : Olivier Bousquet, André Elisseeff : Stability and Generalization. Journal of Machine Learning Research 2 : 499-526 (2002)