IFT-3395 Devoir 3

Olivier St-Laurent, Maxime Daigle

2018-11-09

Question 1 Relations et dérivées de quelques fonction de base

1. Montrez que sigmoid$(x) = \frac{1}{2}(tanh(\frac{1}{2}x) + 1)$

$sigmoid(x) = \frac{1}{2}(tanh(\frac{1}{2}x) + 1) \iff 2sigmoid(x) - 1 = tanh(\frac{x}{2})$

$2sigmoid(x) - 1 = \frac{2-1-exp(-x)}{1+exp(-x)} = \frac{1-exp(-x)}{1+exp(x)} = \frac{exp(\frac{x}{2})}{exp(\frac{x}{2})}(\frac{1-exp(-x)}{1+exp(-x)})$

$= \frac{exp(\frac{x}{2})-exp(\frac{x}{2})exp(-x)}{exp(\frac{x}{2})+exp(\frac{x}{2})exp(-x)} = \frac{exp(\frac{x}{2})-exp(\frac{-x}{2})}{exp(\frac{x}{2})+exp(\frac{-x}{2})} = tanh(\frac{x}{2})$

2. Montez que $\ln sigmoid(x) = -softplus(-x)$

$\ln sigmoid(x) = \ln \frac{1}{1+exp(-x)} = \ln(1) - \ln(1+exp(-x)) = 0 + \ln(1+exp(-x)) = -softplus(-x)$

3. Montrez que $\frac{d\ sigmoid}{dx}(x) = sigmoid(x)(1 - sigmoid(x)$

$\frac{d\ sigmoid}{dx}(x) = \frac{((1+exp(-x))^{-1})}{dx} = \frac{-1}{(1+e^{-x})^2}(-e^{-x}) = (\frac{1}{1+e^{-x}})(\frac{e^{-x}}{1+e^{-x}})$

$= sigmoid(x)(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}) = sigmoid(x)(1 - sigmoid(x))$

4. Montrez que la dérivée de tanh est : $tanh'(x) = 1 - tanh^2(x)$

$\frac{d}{dx}(\frac{e^x-e^{-x}}{e^x+e^{-x}}) = \frac{(e^x-e^{-x})'(e^x+e^{-x})-(e^x-e^{-x})(e^x+e^{-x})'}{(e^x e^{-x})^2} = \frac{(e^x+e^{-x})^2-(e^x-e^{-x})^2}{(e^x+e^{-x})^2}$

$= \frac{(e^x+e^{-x})^2}{(e^x+e^{-x})^2} - \frac{(e^x-e^{-x})^2}{(e^x+e^{-x})^2} = 1 - tanh^2(x)$

5. Expliquer la relation entre le gradient du risque empirique et les erreurs du modèle sur l'ensemble d'entraînement

Le gradient du risque empirique permet de trouver le minimum global des erreurs du modèle sur l'ensemble d'entraînement. Autrement dit, le point où le gradient est égal à 0 est le point où les erreurs du modèle sont les plus petites sur l'ensemble d'entraînement.

Question 1.2 Régression linéaire régularisée ("ridge regression")

1. Exprimez le gradient du risque régularisé. En quoi diffère-t-il du gradient du risque empirique non régularisé ?

$$\nabla \widetilde{R}(\theta) = \frac{\delta \widetilde{R}(\theta)}{\delta \theta}$$

$$= \frac{\delta}{\delta \theta}(\hat{R} + \lambda \mathcal{L}(\theta))$$

$$= \frac{\delta}{\delta \theta}(\sum_{i=1}^{n}(w^T x^{(i)} + b - t^{(i)})^2 + \lambda \sum_{k=1}^{d} w_k^2)$$

$$= \frac{\delta}{\delta \theta}(\sum_{i=1}^{n}(w^T x^{(i)} + b - t^{(i)})^2) + \frac{\delta}{\delta \theta}(\lambda \sum_{k=1}^{d} w_k^2)$$

$$= \nabla \hat{R}(\theta) + \frac{\delta}{\delta \theta}\lambda(\sum_{k=1}^{d} w_k^2)$$

Donc, la différence entre $\nabla \widetilde{R}(\theta)$ et $\nabla \hat{R}(\theta)$ est l'addition de $\frac{\delta}{\delta \theta}\lambda(\sum_{k=1}^{d} w_k^2)$

2. Donner le pseudo code détaillé de l'algorithme qui cherche les paramètres optimaux pour minimiser $\widetilde{R}$ par descente de gradient batch.

$$\text{loop } \theta \leftarrow \theta - \eta \frac{\delta \widetilde{R}}{\delta \theta}$$

$$= \theta - \eta [\sum_{i=1}^{n} \frac{\delta}{\delta \theta} (w^T x^{(i)} + b - t^{(i)})^2 + \lambda \frac{\delta}{\delta \theta} (\sum_{k=1}^{d} w_k^2)]$$

L'algorithme est donc :
θ = initialize_randomly()
stop = False
while not stop :
    gradient = $\sum_{i=1}^{n} \frac{\delta}{\delta \theta} (w^T x^{(i)} + b - t^{(i)})^2 + \lambda \frac{\delta}{\delta \theta} (\sum_{k=1}^{d} w_k^2)$
    $\theta = \theta - \eta$.gradient
    stop = || gradient || $< \varepsilon$

3.exprimer le risque empirique et son gradient sous forme matricielle

$$\theta^* = (X^T X)^{-1} X^T t$$

Question 1.3 Régression avec un pré-traitement non-linéaire fixe

1. Écriver de manière détaillée la from paramétrique qu'on obtient pour $\widetilde{f}(x)$ dans le cas une dimension ($x \in \mathbb{R}$) si on utilise $\phi = \phi_{poly^k}$

$$\widetilde{f}(x) = f(\phi_{poly^k}(x)) = w^T \phi_{poly^k}(x) + b = w^T \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^k \end{bmatrix} + b$$

2. Préciser quels sont les paramètres et leur dimensionalité

Les paramètres sont **w** et **b**. Pour $\phi_{poly^k} \in \mathbb{R}^k$, $w \in \mathbb{R}^k$ est un vecteur. $b \in mathbbR$ est un scalaire.