$\Upsilon\Sigma 19$ ARTIFTICIAL INTELLIGENCE II (DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING). FALL SEMESTER 2020, HOMEWORK 1

National and Kapodistrian University of Athens, DiT. 20/10/2020

Dimitrios Foteinos | sdi1700181

Contents

1 Partial Derivatives of the Ridge Regression Loss Function

Compute the partial derivatives of the ridge regression loss function:

$$\jmath(\boldsymbol{w}) = MSE(\boldsymbol{w}) + \alpha\frac{1}{2}\sum_{i=1}^{n} {w_i}^2 \text{ (1)}$$

In other words, we want to compute the gradient of the above loss function: $\nabla\jmath(\boldsymbol{w})$.

First, we must clarify how we can calculate each partial derivative of this cost function.

For the $MSE(\boldsymbol{w})$, we already know that each partial derivative is equal to:

$$\frac{\partial}{\partial w_j} MSE(\boldsymbol{w}) = \frac{2}{m}\sum_{i=1}^{m}(\boldsymbol{w}\cdot\boldsymbol{x}^{(i)} - y^{(i)}){x_j}^{(i)}$$

And thus, the gradient of this equation (if we compute each partial derivative with respect to w):

$$\nabla_w MSE(\boldsymbol{w}) = \frac{2}{m}(\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})) \text{ (2)}$$

And for the other member of the (1) equation, its partial derivative is equal to:

$$\frac{\boldsymbol{\partial}}{\boldsymbol{\partial w_j}}(\alpha\frac{1}{2}\sum_{i=1}^{n} {w_i}^2) = \alpha\frac{1}{2}\cdot\frac{\boldsymbol{\partial}}{\boldsymbol{\partial w_j}}(\sum_{i=1}^{n} {w_i}^2) = \alpha\frac{1}{2}\cdot\sum_{i=1}^{n}\frac{\boldsymbol{\partial {w_i}^2}}{\boldsymbol{\partial w_j}}$$

And respectively, if we compute its partial derivative, the gradient is:

$$[\alpha\frac{1}{2}\cdot\sum_i\frac{\boldsymbol{\partial {w_i}^2}}{\boldsymbol{\partial w_1}}, \alpha\frac{1}{2}\cdot\sum_i\frac{\boldsymbol{\partial {w_i}^2}}{\boldsymbol{\partial w_2}}, ..., \alpha\frac{1}{2}\cdot\sum_i\frac{\boldsymbol{\partial {w_i}^2}}{\boldsymbol{\partial w_n}}] \text{ (3)*}$$

And now let's assume that, we have $n = 2$ number of features. Then, the **Regularization term** will be equivalent to:

$$\alpha\frac{1}{2} \cdot \sum_{i=1}^{2} {w_i}^2 = \alpha\frac{1}{2} \cdot ({w_1}^2 + {w_2}^2)$$

And lets take for example the computation of the partial derivative $\frac{\partial}{\partial w1}$.

$$\alpha\frac{1}{2} \cdot \frac{\partial}{\partial w1}({w_1}^2 + {w_2}^2) = \alpha\frac{1}{2} \cdot (\frac{\partial {w_1}^2}{\partial w1} + \frac{\partial {w_2}^2}{\partial w1}) = \alpha\frac{1}{2} \cdot 2w1 + 0 = a \cdot w1$$

So, (3)* will be equivalent to:

$$[\alpha \cdot w1, \alpha \cdot w2, ..., \alpha \cdot w_n]] = \alpha \cdot \boldsymbol{w} \text{ .(3)}$$

In conclusion, if we combine the $\mathbf{(2), (3)}$ equations, the gradient of ridge regression's loss function will be:

$$\nabla \jmath(\boldsymbol{w}) = \frac{2}{m}(\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})) + \alpha \cdot \boldsymbol{w}.$$