Artificial Intelligence II

Nefeli Tavoulari

Fall Semester 2021

1

The Mean Squared Error loss function is defined as:

$\mathcal{MSE} = \frac{1}{m} \sum_{i=1}^{m} (h_w(x_i) - y_i)^2$

Calculating the gradient of MSE, we can find out the parameters which minimize the loss.

$\nabla_{\mathbf{w}} \mathcal{MSE} = \nabla_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} (h_w(x_i) - y_i)^2$

The predicted values y_i and the features x_i can be considered as constants, since the gradient is with respect to w, which is the model's parameter vector.
(scalar multiplication rule)

$\nabla_{\mathbf{w}} \mathcal{MSE} = \frac{1}{m} \nabla_{\mathbf{w}} \sum_{i=1}^{m} (h_w(x_i) - y_i)^2$

Then, assuming we only have one instance (x,y) and using the power rule and the chain rule for derivatives:

$\nabla_{\mathbf{w}} \mathcal{MSE} = \nabla_{\mathbf{w}} (h_w(x) - y)^2$

$$\nabla_{\mathbf{w}}\mathcal{MSE} = 2(h_w(x) - y)\nabla_{\mathbf{w}}(h_w(x) - y)$$

where $h_w(x) = w_0x_0 + ... + w_mx_m = w^Tx$: **(*)**

$$\nabla_{\mathbf{w}}\mathcal{MSE} = 2(h_w(x) - y)\nabla(w_0x_0 + ... + w_mx_m - y)$$

$$\nabla_{\mathbf{w}}\mathcal{MSE} = 2(h_w(x)-y)(\frac{\partial(w_0x_0+...+w_mx_m-y)}{\partial w_0}, ..., \frac{\partial(w_0x_0+...+w_mx_m-y)}{\partial w_m})$$

$$\nabla_{\mathbf{w}}\mathcal{MSE} = 2(h_w(x) - y)(x_0, ..., x_m)$$

$$\nabla_{\mathbf{w}}\mathcal{MSE} = 2(h_w(x) - y)(x)$$

so, from **(*)**:

$$\nabla_{\mathbf{w}}\mathcal{MSE} = 2(w^Tx - y)(x)$$

which equals to:

$$\begin{bmatrix} 2(w^Tx - y)(x_0) \\ ... \\ 2(w^Tx - y)(x_m) \end{bmatrix}$$

Now, for m instances:

$$\nabla_{\mathbf{w}}\mathcal{MSE} = \frac{2}{m}\sum_{i=1}^{m}(h_w(x_i) - y_i)\nabla_{\mathbf{w}}(h_w(x_i) - y_i)$$

$$\nabla_{\mathbf{w}}\mathcal{MSE} = \begin{bmatrix} \frac{2}{m}(w^Tx_1 - y_1)(x_{1,0}) + ... + \frac{2}{m}(w^Tx_m - y_m)(x_{m,0}) \\ ... \\ \frac{2}{m}(w^Tx_1 - y_1)(x_{1,n}) + ... + \frac{2}{m}(w^Tx_m - y_m)(x_{m,n}) \end{bmatrix}$$

So: $\nabla_{\mathbf{w}}\mathcal{MSE} = \frac{2}{m}(\mathrm{X}^T(\mathrm{Xw - y}))$