How to apply gradient with respect to a vector

Asked 5 years, 1 month ago Modified 3 years, 10 months ago Viewed 2k times

In Deep Learning (adapted from page 108), explaining linear regression as a machine learning algorithm, there is a passage for the solution of this expression:

> To minimize MSE, we can simply solve for where its gradient is 0:
>
> $$\nabla_{\mathbf{w}} MSE = 0$$

In addition, $\hat{\mathbf{y}}$ is defined as the prediction of the linear regression (also defined as $\mathbf{Xw}$, where $\mathbf{X}$ is the matrix of inputs and $\mathbf{w}$ is the weights vector), while $\mathbf{y}$ is defined as the real output value.

The solution follows this path:

$$\nabla_{\mathbf{w}} MSE = 0$$

$$\Rightarrow \nabla_{\mathbf{w}} \frac{1}{m} ||\hat{\mathbf{y}} - \mathbf{y}||_2^2 = 0$$

$$\Rightarrow \frac{1}{m} \nabla_{\mathbf{w}} ||\mathbf{Xw} - \mathbf{y}||_2^2 = 0$$

$$\Rightarrow \nabla_{\mathbf{w}} (\mathbf{Xw} - \mathbf{y})^T (\mathbf{Xw} - \mathbf{y}) = 0$$

$$\Rightarrow \nabla_{\mathbf{w}} (\mathbf{w}^T \mathbf{X}^T \mathbf{Xw} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}) = 0$$

Now, the subsequent step is:

$$\Rightarrow (2\mathbf{X}^T \mathbf{Xw} - 2\mathbf{X}^T \mathbf{y}) = 0$$

I think to understand that it takes the vector derivative with respect to $\mathbf{w}$, however I could not find the exact term of this derivative and consequently its rules to carry out the derivative myself (in particular, how to deal with transposed vectors and matrices).

linear-algebra multivariable-calculus derivatives vector-analysis linear-regression

Well this is a scalar valued function of the weight vector. Look at atmos.washington.edu/~dennis/MatrixCalculus.pdf and in particular propositions 7, 8, 9. Recall that the transpose of a product is equal to the product of the transposes. – James Lea Mar 17, 2017 at 20:36

2 Answers

Sorted by:

Highest score (default)

Using matrix transpose notation with vectors often confuses me. So I prefer to expand the norm using an explicit dot product instead

$$\|z\|_2^2 = z \cdot z$$

In this form, finding the differential and the gradient of the norm is straightforward

$$\begin{aligned} d\|z\|_2^2 &= 2z \cdot dz \\ \frac{\partial \|z\|_2^2}{\partial z} &= 2z \end{aligned}$$

Now repeat the calculation for $z = (X \cdot w - y)$

$$\begin{aligned} d\|z\|_2^2 &= 2z \cdot dz \\ &= 2z \cdot (X \cdot dw) \\ &= 2(X^T \cdot z) \cdot dw \end{aligned}$$

$$\begin{aligned} \frac{\partial \|z\|_2^2}{\partial w} &= 2X^T \cdot z \\ &= 2X^T \cdot (X \cdot w - y) \end{aligned}$$

Greg's answer already solved the problem. However, I want to specifically address the question about the notation $\nabla_w f$.

If you are using "numerator layout" (for me the most logical), $\nabla_w f$ simply refers to $\frac{\partial f}{\partial w^T}$.

$$\begin{gathered} \frac{\partial}{\partial w^T}\left(w^T X^T X w - 2 w^T X^T y + y^T y\right) \\ = \frac{\partial w^T}{\partial w^T} X^T X w + w^T X^T X \frac{\partial w}{\partial w^T} - 2\frac{\partial w^T}{\partial w^T} X^T y \\ = X^T X w + (w^T X^T X)^T - 2X^T y \\ = 2X^T X w - 2X^T y \end{gathered}$$

If you are using "denominator layout" then $\nabla_w f$ is $\frac{\partial f}{\partial w}$, but the result is the same.

You can read more about layouts at [the wikipedia article on matrix calculus.]()
