Linear Algebra

Stephen Huan

October 6, 2020

1 Introduction

Suppose we have a scalar field (a function of possibly multiple variables that returns a single scalar value). We know from multivariable calculus that we can take derivatives of a function of multiple variables with respect to each variable, and encapsulate all the derivatives into a ***gradient*** vector.

Now we will see what will happen if we take the derivative of a scalar field *with respect to a vector*. We will approach this from two angles, a standard multivariate approach and a more tensor-theoretic approach.

2 Examples

Suppose we have the scalar field

$$f(\vec{\beta}) = \vec{z}^T \vec{\beta}$$

This is a function of multiple variables which returns a single number ($\vec{z}^T \vec{\beta} = \vec{z} \cdot \vec{\beta}$). If we explicitly write it out, we get $\vec{z}^T \vec{\beta} = z_1\beta_1 + z_2\beta_2 + \dots$ Taking the partial derivative with respect to β_1, we get z_1, with respect to β_2, we get z_2, and so on. Since the gradient of f, denoted $\nabla_{\vec{\beta}} f$ is $\left\langle \frac{\partial f}{\partial \beta_1}, \frac{\partial f}{\partial \beta_2}, \dots \right\rangle$, the gradient is just $\vec{z}$.

We can come to the same conclusion with a different method. Recall that the definition for the derivative in singlevariate calculus is:

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Intuitively, the change in the function over the infinitesimal change in x. If we try to extend this definition to vectors, and replace zero with the zero vector, we run into a problem—division by a vector isn't well-defined. So we need a different conceptual basis to define the derivative.

The derivative is a ***linear transformation***, that is, it fulfills two properties:

1. $T(x+y) = T(x) + T(y)$ for any x, y
2. $T(cx) = cT(x)$ for any scalar.

The derivative of $f+g$, for two functions f and g is the derivative of f plus the derivative of g, scalars can be taken out of differentiation and added back in later. We can also think of the derivative giving us a way to estimate the change in a function as a function of changing x, e.g. $df = f'(x)dx$. df can then be thought of as a linear transformation of dx. So we have two forms of linearity: the derivative is a linear transformation in its operator sense, that is, a linear transformation from functions to their derivatives, as well as a linear transformation from a infinitesimal change in x to its corresponding infinitesimal change in $f(x)$. If we think of the derivative as a linear transformation of differentials, that gives the alternative definition we are looking for.

$$\begin{aligned} df &= f(\vec{\beta} + d\vec{\beta}) - f(\vec{\beta}) && \text{Definition} \\ &= \vec{z}^T\vec{\beta} + \vec{z}^T d\vec{\beta} - \vec{z}^T\vec{\beta} && \text{Expanding} \\ &= \vec{z}^T d\vec{\beta} \end{aligned}$$

which is a linear transformation of $d\vec{\beta}$, and matches the result derived earlier. Except for the transpose, which I don't have a good way of explaining. I guess we need to transpose our answer at the end. For a more complicated example, suppose we have the field

$$f(\vec{\beta}) = \vec{\beta}^T\boldsymbol{\sigma}\vec{\beta}$$

where $\boldsymbol{\sigma}$ is a symmetric matrix. For convenience, let $\boldsymbol{\sigma}_i$ be the ith *row* of $\boldsymbol{\sigma}$.

$$\begin{aligned} \vec{\beta}^T\boldsymbol{\sigma}\vec{\beta} &= \vec{\beta} \cdot \left\langle \boldsymbol{\sigma}_1 \cdot \vec{\beta}, \boldsymbol{\sigma}_2 \cdot \vec{\beta}, \ldots \right\rangle && \text{Definition of matrix-vector product} \\ &= \vec{\beta}_1 \cdot \boldsymbol{\sigma}_1 \cdot \vec{\beta} + \vec{\beta}_2 \cdot \boldsymbol{\sigma}_2 \cdot \vec{\beta} + \ldots && \text{Expanding the dot product} \\ &= \vec{\beta}_1(\boldsymbol{\sigma}_{11}\vec{\beta}_1 + \boldsymbol{\sigma}_{12}\vec{\beta}_2 + \boldsymbol{\sigma}_{13}\vec{\beta}_3 + \ldots + \boldsymbol{\sigma}_{21}\vec{\beta}_2 + \boldsymbol{\sigma}_{31}\vec{\beta}_3 + \ldots) && \text{Collecting } \vec{\beta}_1 \text{ terms} \end{aligned}$$

We now compute the partial with respect to $\vec{\beta}_1$

$$\begin{aligned} \frac{\partial f}{\partial \vec{\beta}_1} &= 2\boldsymbol{\sigma}_{11}\vec{\beta}_1 + [2\boldsymbol{\sigma}_{12}\vec{\beta}_2 + \boldsymbol{\sigma}_{13}\vec{\beta}_3 + \ldots + \boldsymbol{\sigma}_{21}\vec{\beta}_2 + \boldsymbol{\sigma}_{13}\vec{\beta}_3 + \ldots] \\ &= 2\boldsymbol{\sigma}_{11}\vec{\beta}_1 + [2\boldsymbol{\sigma}_1 \cdot \vec{\beta} - 2\boldsymbol{\sigma}_{11}\vec{\beta}_1] && \text{By symmetry of } \boldsymbol{\sigma} \\ &= 2\boldsymbol{\sigma}_1 \cdot \vec{\beta} \end{aligned}$$

Since $\vec{\beta}_1$ is symmetric to every index in $\vec{\beta}$, $\nabla_{\vec{\beta}} f = 2\boldsymbol{\sigma}\vec{\beta}$. We can also do this with our alternative definition.

$$\begin{aligned}
df &= f(\vec{\beta} + d\vec{\beta}) - f(\vec{\beta}) && \text{Definition} \\
&= (\vec{\beta} + d\vec{\beta})^T \boldsymbol{\sigma} (\vec{\beta} + d\vec{\beta}) - \vec{\beta}^T \boldsymbol{\sigma}\vec{\beta} && \text{Expanding} \\
&= (\vec{\beta} + d\vec{\beta})(\boldsymbol{\sigma}\vec{\beta} + \boldsymbol{\sigma} d\vec{\beta}) - \vec{\beta}^T \boldsymbol{\sigma}\vec{\beta} \\
&= \vec{\beta}^T \boldsymbol{\sigma}\vec{\beta} + \vec{\beta}^T \boldsymbol{\sigma} d\vec{\beta} + d\vec{\beta}^T \boldsymbol{\sigma}\vec{\beta} + d\vec{\beta}^T \boldsymbol{\sigma} d\vec{\beta} - \vec{\beta}^T \boldsymbol{\sigma}\vec{\beta}
\end{aligned}$$

First, we can discard $d\vec{\beta}^T \boldsymbol{\sigma} d\vec{\beta}$ since it is not a linear transformation of $d\vec{\beta}$ (intuitvely, it is a higher order differential term)

$$= \vec{\beta}^T \boldsymbol{\sigma}\vec{\beta} + \vec{\beta}^T \boldsymbol{\sigma} d\vec{\beta} + d\vec{\beta}^T \boldsymbol{\sigma}\vec{\beta} - \vec{\beta}^T \boldsymbol{\sigma}\vec{\beta}$$

Taking advantage of the fact that $(\vec{\beta}^T \boldsymbol{\sigma} d\vec{\beta})^T = d\vec{\beta}^T \boldsymbol{\sigma}^T \vec{\beta} = d\vec{\beta}^T \boldsymbol{\sigma}\vec{\beta}$, and the fact that both are scalars, so if their transpose is equal they are equal,

$$= 2\vec{\beta}^T \boldsymbol{\sigma} d\vec{\beta}$$

If we transpose $2\vec{\beta}^T \boldsymbol{\sigma}$, we get $2\boldsymbol{\sigma}\vec{\beta}$, which is our answer.

3 Least-squares

The least squares problem is the following: we have a matrix of features $\boldsymbol{X}$, and a list of prediction values $\vec{y}$. We suspect there is a linear relationship between the features and the target value, so we are trying to find a set of weights $\vec{\beta}$ such that the predictions generated by $\hat{y} = \boldsymbol{X}\vec{\beta}$ are as close to $\vec{y}$ as possible, i.e. $\|\vec{y} - \hat{y}\|$ is minimized. First, we can minimize $\|\vec{y} - \hat{y}\|^2$ instead since squaring is monotonic, and that avoids having to take a pesky square root. To minimize a function, we take the gradient and set equal to the zero vector.

$$\begin{aligned}
f(\vec{\beta}) &= \|\vec{y} - \hat{y}\|^2 \\
&= (\vec{y} - \hat{y}) \cdot (\vec{y} - \hat{y}) && \text{Definition of magnitude} \\
&= \vec{y} \cdot \vec{y} - 2\vec{y} \cdot \hat{y} + \hat{y} \cdot \hat{y} \\
&= \vec{y}^T \vec{y} - 2\vec{y}^T \boldsymbol{X}\vec{\beta} + (\boldsymbol{X}\vec{\beta})^T \boldsymbol{X}\vec{\beta} && \text{Definition of } \hat{y} \\
&= \vec{y}^T \vec{y} - \underbrace{2\vec{y}^T \boldsymbol{X}}_{\vec{z}} \vec{\beta} + \vec{\beta}^T \underbrace{\boldsymbol{X}^T \boldsymbol{X}}_{\boldsymbol{\sigma}} \vec{\beta} && \boldsymbol{X}^T \boldsymbol{X} \text{ is symmetric}
\end{aligned}$$

Using the gradients dervied above, and the fact that $\vec{y} \cdot \vec{y}$ is a constant,

$$\begin{aligned}
\nabla_{\vec{\beta}} f = -2\boldsymbol{X}^T \vec{y} + 2\boldsymbol{X}^T \boldsymbol{X}\vec{\beta} = \vec{0} \\
\boldsymbol{X}^T \boldsymbol{X}\vec{\beta} = \boldsymbol{X}^T \vec{y} \\
\boxed{\vec{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \vec{y}}
\end{aligned}$$