函数对向量求导
通俗易懂的描述

Dezeming Family

2021 年 6 月 5 日

DezemingFamily 系列书和小册子因为是电子书，所以可以很方便地进行修改和重新发布。如果您获得了 DezemingFamily 的系列书，可以从我们的网站 [https://dezeming.top/] 找到最新版。对书的内容建议和出现的错误欢迎在网站留言。

2020 年 6 月 5 日，完成第一版。

2020 年 6 月 25 日，增加了一些细节性的描述。

目录

一　函数对向量的偏导

我们可以把一个向量简单理解为多元变量：

$$\boldsymbol{x}^T = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_n \end{bmatrix} \tag{一.1}$$

现在我们有一个函数，它的输入变量是该向量，则函数对该向量的导数其实就是对多元变量的偏导：

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}^T} = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} & \frac{\partial f(\boldsymbol{x})}{\partial x_2} & \dots & \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \tag{一.2}$$

而输入向量的函数也可以理解为是一个多元函数，每个向量分量都是函数中的一元，这样求导以后的结果其实就是多元函数偏导数，构成一个向量。这个多元偏导构成的向量其实对于向量 $\boldsymbol{x}$ 来说就是一个函数组了，因此，函数对向量的二阶导其实就是一个函数组对向量再求一次导数。

矩阵也可以理解为一个函数组，即一个矩阵包含了多个函数。我们假设有一个矩阵 A 作用于向量 $\boldsymbol{x}$：

$$f(\boldsymbol{x}) = \begin{bmatrix} a_{1,1} & a_{1,2} & . & . & . & a_{1,n} \\ a_{2,1} & a_{2,2} & . & . & . & a_{2,n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ a_{m,1} & a_{m,2} & . & . & . & a_{m,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix} \tag{一.3}$$

该矩阵作用于向量后生成另一个向量，我们可以看到：

$$Ax = \begin{bmatrix} a_{1,1}x_1 + a_{1,2}x_2 + ... + a_{1,n}x_n \\ a_{2,1}x_1 + a_{2,2}x_2 + ... + a_{2,n}x_n \\ . \\ . \\ . \\ a_{m,1}x_1 + a_{m,2}x_2 + ... + a_{m,n}x_n \end{bmatrix} \tag{一.4}$$

需要注意的是，$A\boldsymbol{x}$ 其实是一个向量，对于 $\boldsymbol{x}$ 来说，也是一个函数组（单值向量函数求导以后也是一个函数组）。

我们首先先看向量 $\boldsymbol{x}$ 其中一个元素的偏导：

$$\frac{\partial A\boldsymbol{x}}{\partial x_i} = \begin{bmatrix} \frac{\partial A\boldsymbol{x}_1}{\partial x_i} : a_{1,i} \\ \frac{\partial A\boldsymbol{x}_2}{\partial x_i} : a_{2,i} \\ . \\ . \\ . \\ \frac{\partial A\boldsymbol{x}_m}{\partial x_i} : a_{m,i} \end{bmatrix} \tag{一.5}$$

之后整个求导以后的偏导我们可以构成一个向量为：

$$\frac{\partial A\boldsymbol{x}}{\partial \boldsymbol{x}} = [\frac{\partial A\boldsymbol{x}}{\partial x_1}, \frac{\partial A\boldsymbol{x}}{\partial x_2}, ..., \frac{\partial A\boldsymbol{x}}{\partial x_n}] \tag{一.6}$$

$$= \left[[a_{1,1}, a_{2,1}, ..., a_{m,1}]^T, [a_{1,2}, a_{2,2}, ..., a_{m,2}]^T, ..., [a_{n,1}, a_{n,2}, ..., a_{n,m}]^T\right] \tag{一.7}$$

那么现在有个问题，求导以后为什么要这么表示？或者更进一步，求导以后为什么构造成这样一个矩阵？为什么不能是一个 $m \times n$ 维的向量？其实，这只是一种表示形式罢了。

其实这个问题并没有那么直观，首先我们需要明白的是，微分是一种线性操作（满足线性运算的基本性质），矩阵是线性运算的一种表示形式，也就是说我们是可以将对向量的导数描述为一个矩阵。

我们把输出向量 $A\boldsymbol{x}$ 定义为 $\boldsymbol{f} = [f_1, f_2, ..., f_m]$，然后定义矩阵 $\boldsymbol{f'}$：

$$\boldsymbol{f'} = \begin{bmatrix} \frac{\partial \boldsymbol{f}}{\partial x_1} & \frac{\partial \boldsymbol{f}}{\partial x_2} & ... & \frac{\partial \boldsymbol{f}}{\partial x_n} \end{bmatrix} \tag{一.8}$$

$$= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & . & . & . & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & . & . & . & \frac{\partial f_2}{\partial x_n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \frac{f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & . & . & . & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \tag{一.9}$$

以上介绍的形式为行向量形式，至于表示成什么形式，其实也只是为了方便计算罢了。而一般我们会将偏导写为列向量的形式，这个时候称为梯度算子：

$$\nabla_{\boldsymbol{x}}^{A\boldsymbol{x}} = [\frac{\partial A\boldsymbol{x}}{\partial \boldsymbol{x}}]^T = [\frac{\partial A\boldsymbol{x}}{\partial x_1}, \frac{\partial A\boldsymbol{x}}{\partial x_2}, ..., \frac{\partial A\boldsymbol{x}}{\partial x_n}]^T \tag{一.10}$$

$$= \begin{bmatrix} [a_{1,1}, a_{2,1}, ..., a_{m,1}] \\ [a_{1,2}, a_{2,2}, ..., a_{m,2}] \\ ...... \\ [a_{n,1}, a_{n,2}, ..., a_{n,m}] \end{bmatrix} = A^T \tag{一.11}$$

我们下一节用两个例子来描述。

二　两个偏导求解的例子

示例 1

定义：

$$f(\boldsymbol{x}) = \boldsymbol{x}^T A \boldsymbol{x} \tag{二.1}$$

首先我们需要明确的是，能这么计算的矩阵 A 是一个方阵。

我们写出来 f 的结果：

$$\begin{aligned} f(\boldsymbol{x}) &= a_{1,1}x_1x_1 + a_{1,2}x_1x_2 + ... + a_{1,n}x_1x_n & (二.2)\\ &+ a_{2,1}x_2x_1 + a_{2,2}x_2x_2 + ... + a_{2,n}x_2x_n & (二.3)\\ &+ ...... & (二.4)\\ &+ a_{n,1}x_nx_1 + a_{n,2}x_nx_2 + ... + a_{n,n}x_nx_n & (二.5)\\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} a_{j,k}x_jx_k & (二.6)\end{aligned}$$

因此，导数可以计算为：

$$\left[\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}\right]_i = \left[\frac{\partial\left(\sum_{j=1}^{n}\sum_{k=1}^{n} a_{j,k}x_jx_k\right)}{\partial x_i}\right]_i = \sum_{j=1}^{n} x_j a_{j,i} + \sum_{k=1}^{n} x_k a_{i,k} \tag{二.7}$$

插入内容：导数计算

说一下这种 $\sum$ 形式怎么计算导数。

$$\left(\sum_{j=1}^{n}\sum_{k=1}^{n} a_{j,k}x_jx_k\right) \tag{二.8}$$

首先令 $j = i$，得到 $\sum_{k=1}^{n} a_{i,k}x_ix_k$，对 x_i 求导可得 $\sum_{k=1}^{n} x_k a_{i,k}$，然后令 $k = i$，计算方法同理。

导数计算结束

现在怎么表示它是你的自由，我们可以把它表示为一个向量（上式只表示了其中一个元素的值），因此我们将该向量可以写为（我们第三章再讲解具体原因）：

$$(A^T + A)\boldsymbol{x} \tag{二.9}$$

注意我们这么写的时候，在参与运算中要与其他的表示方法一致，我们见下例。

示例 2

我们看一下最小二乘法中的误差描述（暂时不用管这是怎么来的），向量表示为上加箭头 $\overrightarrow{x}$：

$$e = \left(\overrightarrow{x}^{*T} A^T A \overrightarrow{x}^{*} - 2\overrightarrow{x}^{*T} A^T \overrightarrow{b} + \overrightarrow{b}^T \overrightarrow{b}\right) \tag{二.10}$$

我们需要 e 对 $\overrightarrow{x}^*$ 进行求导，让导数为 0 时则说明达到了极值点：

$$\frac{\partial}{\partial \overrightarrow{x}^*} e = 0 \tag{二.11}$$

因为 $\overrightarrow{b}$ 是一个无关 $\overrightarrow{x}^*$ 的向量，我们可以不管后面这一项，只研究前面两项的求导。

设 A 是 m 行 n 列的矩阵。由此我们知道 $\overrightarrow{x}^*$ 是 n 维向量，A^TA 是 n 行 n 列的矩阵。

可以计算得到：

$$A^TA = \begin{bmatrix} \sum_{i=1}^{m} a_{i,1}a_{i,1} & \sum_{i=1}^{m} a_{i,1}a_{i,2} & ... & \sum_{i=1}^{m} a_{i,1}a_{i,n} \\ \sum_{i=1}^{m} a_{i,2}a_{i,1} & \sum_{i=1}^{m} a_{i,2}a_{i,2} & ... & \sum_{i=1}^{m} a_{i,2}a_{i,n} \\ ... & ... & ... & ... \\ \sum_{i=1}^{m} a_{i,n}a_{i,1} & \sum_{i=1}^{m} a_{i,n}a_{i,2} & ... & \sum_{i=1}^{m} a_{i,n}a_{i,n} \end{bmatrix} \tag{二.12}$$

$$A^T A\overrightarrow{x}^* = \begin{bmatrix} \sum_{i=1}^m a_{i,1}a_{i,1}x_1^* + \sum_{i=1}^m a_{i,1}a_{i,2}x_2^* + ... + \sum_{i=1}^m a_{i,1}a_{i,n}x_n^* \\ \sum_{i=1}^m a_{i,2}a_{i,1}x_1^* + \sum_{i=1}^m a_{i,2}a_{i,2}x_2^* + ... + \sum_{i=1}^m a_{i,2}a_{i,n}x_n^* \\ ... \\ \sum_{i=1}^m a_{i,n}a_{i,1}x_1^* + \sum_{i=1}^m a_{i,n}a_{i,2}x_2^* + ... + \sum_{i=1}^m a_{i,n}a_{i,n}x_n^* \end{bmatrix} \tag{二.13}$$

$$= \begin{bmatrix} \sum_{j=1}^n \sum_{i=1}^m a_{i,1}a_{i,j}x_j^* \\ \sum_{j=1}^n \sum_{i=1}^m a_{i,2}a_{i,j}x_j^* \\ ... \\ \sum_{j=1}^n \sum_{i=1}^m a_{i,n}a_{i,j}x_j^* \end{bmatrix} \tag{二.14}$$

进一步：

$$\overrightarrow{x}^{*T} A^T A\overrightarrow{x}^* = \left[\sum_{k=1}^n x_k^* \sum_{j=1}^n \sum_{i=1}^m a_{i,k}a_{i,j}x_j^*\right] \tag{二.15}$$

然后再求导就好了，但是这样太麻烦。我们把 A^TA 当成一个整体。因为我们知道 $\left(A^TA\right)^T = A^TA$，所以其实我们可以这么来求：

$$\frac{\partial \overrightarrow{x}^{*T}\left(A^TA\right)\overrightarrow{x}^*}{\partial \overrightarrow{x}^*}\left(\left(A^TA\right)^T + A^TA\right)\overrightarrow{x}^* = 2A^TA\overrightarrow{x}^* \tag{二.16}$$

对于第二部分，我们把 $A^T\overrightarrow{b}$ 作为一个整体，这是一个 n 维向量，我们设为 $\overrightarrow{C}$。

$$\overrightarrow{x}^{*T} A_b = \left[x_1^*C_1 + x_2^*C_2 + ... + x_n^*C_n\right] \tag{二.17}$$

求导以后要写为列向量的形式，才能与上面的结果一致，因此就可以表示为：$2A^T\overrightarrow{b}$。最终我们得到导数为：

$$\frac{\partial}{\partial \overrightarrow{x}^*}e = 2A^T\left(A\overrightarrow{x}^* - \overrightarrow{b}\right) \tag{二.18}$$

三 关于表示形式的解析

现在有了一个新的问题：$\boldsymbol{x}^TA$ 和 $A\boldsymbol{x}$ 进行求导（假设 A 是一个 n 维方阵）应该得到同样的结果吗？答案肯定是否定的。

对向量求偏导以后的行向量矩阵形式应该满足的格式为：

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}^T} = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} & \frac{\partial f(\boldsymbol{x})}{\partial x_2} & ... & \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \tag{三.1}$$

其中，$\frac{\partial f(\boldsymbol{x})}{\partial x_i}$ 是根据 $f(\boldsymbol{x})$ 是列向量推断出它也是一个列向量。也就是说，无论函数 $f(\boldsymbol{x})$ 长什么样，最后都应该这么排列。因此第二章的示例 1：

$$\left[\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}\right]_i = \sum_{j=1}^n x_j a_{j,i} + \sum_{k=1}^n x_k a_{i,k} \tag{三.2}$$

显然这里的列向量 $\frac{\partial f(\boldsymbol{x})}{\partial x_i}$ 是 1 维向量，整个 $\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}$ 可以写为 $\boldsymbol{x}(A+A^T)$，转换为梯度偏导即为 $(A^T+A)\boldsymbol{x}$。

因此，对于 $\boldsymbol{x}^TA$ 求导以后的行向量矩阵形式中，设 $\boldsymbol{x}$ 是 m 维向量，A 是 $m\times n$ 维矩阵，$\frac{\partial f(\boldsymbol{x})}{\partial x_i}$ 是一个行向量，因此 $\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}$ 就是一个 1 行 mn 列的矩阵，转换为梯度偏导后为 mn 行 1 列的矩阵。

对于 $A\boldsymbol{x}$ 则如第一章所述，这里就不再赘述了。也就是说，我们在运算（比如最小二乘求导）的时候，一定要统一形式，要么都表示成行向量形式，要么都表示成列向量形式。

参考文献

[1] 张贤达. 矩阵分析与应用 [M]. 清华大学出版社, 2013.

[2] https://zhuanlan.zhihu.com/p/29742646