本文整理自李建平机器学习中的矩阵向量求导系列和长躯鬼侠的矩阵求导术。

1. 符号说明

默认符号：

- x ： 标量
- $\mathbf{x}$: n 维列向量
- $\mathbf{y}$: m 维列向量
- X ： $m \times n$矩阵
- Y ： $p \times q$矩阵

2. 矩阵向量求导布局

自变量\因变量	标量y	向量$\mathbf{y}$	矩阵$\mathbf{Y}$
标量x	/	$\frac{\partial \mathbf{y}}{\partial x}$ 分子布局：m维列向量（默认） 分母布局：m维行向量	$\frac{\partial \mathbf{Y}}{\partial x}$ 分子布局：p × q矩阵（默认） 分母布局：$q \times p$矩阵
向量$\mathbf{x}$	$\frac{\partial y}{\partial \mathbf{x}}$ 分子布局：n维行向量 分母布局：n 维列向量（默认）	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ 分子布局：m × n 雅克比矩阵（默认） 分母布局：$n \times m$ 梯度矩阵	/
矩阵$\mathbf{X}$	$\frac{\partial y}{\partial \mathbf{X}}$ 分子布局：$n \times m$ 矩阵 分母布局：m × n 矩阵（默认）	/	$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$ 分母布局：mn × pq 矩阵

3. 矩阵向量求导大全

自变量\因变量	标量y	向量$\mathbf{y}$	矩阵$\mathbf{Y}$
标量x	$\frac{\partial y}{\partial x}$ 大学微积分知识	$\frac{\partial \mathbf{y}}{\partial x}$ 定义法求导	$\frac{\partial \mathbf{Y}}{\partial x}$ 定义法求导
向量$\mathbf{x}$	$\frac{\partial y}{\partial \mathbf{x}}$ 1. 定义法求导 2. 基本法则：线性法则、乘法法则、除法法则 3. 矩阵微分：$df = tr\left(\left(\frac{\partial f}{\partial \mathbf{x}}\right)^T d\mathbf{x}\right)$ 4. 链式法则：$\frac{\partial z}{\partial \mathbf{x}} = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)^T \frac{\partial z}{\partial \mathbf{y}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ 1. 定义法求导 2. 链式法则： $\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{z}}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	——
矩阵$\mathbf{X}$	$\frac{\partial y}{\partial \mathbf{X}}$ 1. 定义法求导 2. 矩阵微分：$df = tr\left(\left(\frac{\partial f}{\partial \mathbf{X}}\right)^T d\mathbf{X}\right)$ 3. 矩阵微分性质 4. 迹技巧 5. 链式求导法则： $\frac{\partial z}{\partial x_{ij}} = \sum_{k,l} \frac{\partial z}{\partial \mathbf{Y}_{kl}} \frac{\partial \mathbf{Y}_{kl}}{\partial \mathbf{X}_{ij}} = tr\left(\left(\frac{\partial z}{\partial \mathbf{Y}}\right)^T \frac{\partial \mathbf{Y}}{\partial \mathbf{X}_{ij}}\right)$	——	$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$ 1. 定义：$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}} = \frac{\partial vec(\mathbf{Y})}{\partial vec(\mathbf{X})}$ 2. 微分法： $vec(d\mathbf{Y}) = \frac{\partial \mathbf{Y}}{\partial \mathbf{X}}^T vec\left(d\mathbf{X}\right)$ 3. 运算法则

4. 标量对向量求导

已知：

$$y = f(\mathbf{x})$$

求：

$$\frac{\partial y}{\partial \mathbf{x}} = ?$$

4.1 定义法求导

所谓标量对向量的求导，其实就是标量对向量里的每个分量分别求导，最后把求导的结果排列在一起，按一个向量表示而已。那么我们可以将实值函数对向量的每一个分量来求导，最后找到规律，得到求导的结果向量。

例1： $y = \mathbf{a}^T\mathbf{x}$

$$\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial x_i} = \frac{\partial \sum_{j=1}^{n} a_j x_j}{\partial x_i} = \frac{\partial a_i x_i}{\partial x_i} = a_i$$

所以，将求导结果组成向量：

$$\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$$

4.2 标量对向量求导基本法则

在我们寻找一些简单的方法前，我们简单看下标量对向量求导的一些基本法则，这些法则和标量对标量求导的过程类似。

1） 常量对向量的求导结果为0。

2）线性法则：如果f, g都是实值函数，c_1, c_2为常数，则：

$$\frac{\partial(c_1 f(\mathbf{x}) + c_2 g(\mathbf{x})}{\partial \mathbf{x}} = c_1 \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + c_2 \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}$$

3) 乘法法则：如果f, g都是实值函数，则：

$$\frac{\partial f(\mathbf{x}) g(\mathbf{x})}{\partial \mathbf{x}} = f(\mathbf{x}) \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} + \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} g(\mathbf{x})$$

要注意的是如果不是实值函数，则不能这么使用乘法法则。

4) 除法法则：如果f, g都是实值函数，且$g(\mathbf{x}) \neq 0$，则：

$$\frac{\partial f(\mathbf{x})/g(\mathbf{x})}{\partial \mathbf{x}} = \frac{1}{g^2(\mathbf{x})}(g(\mathbf{x}) \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} - f(\mathbf{x}) \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}})$$

4.3 通过向量微分求导

利用导数和微分之间的关系：

$$df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^T d\mathbf{x}$$

例： $y = \mathbf{x}^T \mathbf{x}$

$$\begin{aligned} dy &= d\left(\mathbf{x}^T\right) \mathbf{x} + \mathbf{x}^T d\mathbf{x} \\ &= (d\mathbf{x})^T \mathbf{x} + \mathbf{x}^T d\mathbf{x} \\ &= \mathbf{x}^T d\mathbf{x} + \mathbf{x}^T d\mathbf{x} \\ &= 2\mathbf{x}^T d\mathbf{x} \end{aligned}$$

所以，根据导数与微分的联系：

$$\frac{\partial \mathbf{x}^T \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$$

4.4 标量对多个向量的链式求导法则

若标量z 和向量 $\mathbf{x}, \mathbf{y}$之间的依赖关系为：$\mathbf{x} \rightarrow \mathbf{y} \rightarrow z$，则：

$$\frac{\partial z}{\partial \mathbf{x}} = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)^T \frac{\partial z}{\partial \mathbf{y}}$$

推广到多个向量的情况，$\mathbf{y}_1 \rightarrow \mathbf{y}_2 \rightarrow \cdots \mathbf{y}_n \rightarrow z$，则有：

$$\frac{\partial z}{\partial \mathbf{y}_1} = \left(\frac{\partial \mathbf{y}_n}{\partial \mathbf{y}_{n-1}} \frac{\partial \mathbf{y}_{n-1}}{\partial \mathbf{y}_{n-2}} \cdots \frac{\partial \mathbf{y}_2}{\partial \mathbf{y}_1}\right)^T \frac{\partial z}{\partial \mathbf{y}_n}$$

5. 标量对矩阵求导

已知：

$$y = f(X)$$

求：

$$\frac{\partial y}{\partial X} = ?$$

5.1 定义法求导

与标量对向量求导类似，标量对矩阵里的每个分量分别求导，最后把求导的结果排列在一起，用一个矩阵表示而已。

例： $y = \mathbf{a}^T X \mathbf{b}$，求 $\frac{\partial y}{\partial X}$

先对矩阵X的任意一个位置的X_{ij}求导：

$$\frac{\partial \mathbf{a}^T X \mathbf{b}}{\partial X_{ij}} = \frac{\partial \sum_{p=1}^{m} \sum_{q}^{n} a_p X_{pq} b_q}{\partial X_{ij}} = \frac{\partial a_i X_{ij} b_j}{\partial X_{ij}} = a_i b_j$$

将所有位置求导结果排成$m \times n$矩阵，得：

$$\frac{\partial \mathbf{a}^T X \mathbf{b}}{\partial X} = \mathbf{a}\mathbf{b}^T$$

5.2 通过矩阵微分求导

一元微积分中的导数（标量对标量的导数）与微分之间的关系：

$$df = f'(x)dx$$

多元微积分中的梯度（标量对向量的导数）与微分之间的关系：

$$df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i = \frac{\partial f}{\partial \mathbf{x}}^T d\mathbf{x}$$

矩阵导数与微分之间的关系：

$$df = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial f}{\partial X_{ij}} dX_{ij} = \operatorname{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$$

利用导数与微分，以及迹技巧，可以求得标量函数f 对于矩阵 X 的导数：

- 对标量函数 f 求微分，需用到矩阵微分运算法则；
- 使用迹技巧，对 df 套上迹，再将其它项移至dX 左侧；

- 对照导数与微分之间的联系，可求得$\frac{\partial f}{\partial X}$

注：标量对矩阵的求导不能随意沿用标量的链式求导法则。

5.3 矩阵微分运算法则

矩阵加减法	法则	示例
矩阵加减法	$d(X \pm Y) = dX \pm dY$	
矩阵乘法	$d(XY) = (dX)Y + X(dY)$	
矩阵转置	$d(X^T) = (dX)^T$	
矩阵的迹	$d\mathrm{tr}(X) = \mathrm{tr}(dX)$	
矩阵的逆	$dX^{-1} = -X^{-1}dXX^{-1}$	
行列式	$d\|X\| = \mathrm{tr}(X^{\#}dX)$	$X^{\#}$是X的伴随矩阵
逐元素相乘	$d(X \odot Y) = dX \odot Y + X \odot dY$	
逐元素函数	$d\sigma(X) = \sigma'(X) \odot dX$	$d\sin(X) = \cos(X) \odot dX$

5.4 迹技巧（trace trick）

运算	法则	备注
标量套上迹	$a = \mathrm{tr}(a)$	
转置	$\mathrm{tr}(A^T) = \mathrm{tr}(A)$	
线性	$\mathrm{tr}(A \pm B) = \mathrm{tr}(A) \pm \mathrm{B}$	
矩阵乘法交换	$\mathrm{tr}(AB) = \mathrm{tr}(BA)$	A与B^T维度相同，两侧都等于$\sum_{ij} A_{ij}B_{ji}$
矩阵乘法/逐元乘法交换	$\mathrm{tr}(A^T(B \odot C)) = \mathrm{tr}((A \odot B)^T C)$	A, B, C尺寸相同，两侧都等于$\sum_{ij} A_{ij}B_{ij}C_{ij}$

5.5 标量对多个矩阵的链式求导法则

假设有这样的依赖关系：$X \to Y \to z$，很难给出矩阵基于矩阵整体的链式求导法则，可以给出关于X中某一标量的链式求导：

$$\frac{\partial z}{\partial X_{ij}} = \sum_{k,l} \frac{\partial z}{\partial Y_{kl}} \frac{\partial Y_{kl}}{\partial X_{ij}} = tr\left(\left(\frac{\partial z}{\partial \mathbf{Y}}\right)^T \frac{\partial \mathbf{Y}}{\partial \mathbf{X}_{ij}}\right)$$

5.6 计算示例

例1：$f = \mathbf{a}^T X \mathbf{b}$ ，求$\frac{\partial f}{\partial X}$ 。其中$\mathbf{a}$ 是$m \times 1$ 列向量，X是$m \times n$ 矩阵，$\mathbf{b}$ 是$n \times 1$ 列向量。

Step1: 使用矩阵乘法法则求微分：

$$df = d\mathbf{a}^T X \mathbf{b} + \mathbf{a}^T dX \mathbf{b} + \mathbf{a}^T X d\mathbf{b} = \mathbf{a}^T dX \mathbf{b}$$

这里因为$\mathbf{a}, \mathbf{b}$是常量，所以$d\mathbf{a} = 0, d\mathbf{b} = 0$

Step2: 套上迹，并做矩阵乘法交换：

$$df = \operatorname{tr}(\mathbf{a}^T dX \mathbf{b}) = \operatorname{tr}(\mathbf{b}\mathbf{a}^T dX) = \operatorname{tr}\left((\mathbf{a}\mathbf{b}^T)^T dX\right)$$

Step3：对照导数与微分之间的联系：

$$\frac{\partial f}{\partial X} = \mathbf{a}\mathbf{b}^T$$

例2：$f = \mathbf{a}^T \exp(X\mathbf{b})$，求$\frac{\partial f}{\partial X}$ 。

Step1: 使用矩阵乘法法则求微分：

$$df = \mathbf{a}^T \left(\exp(X\mathbf{b}) \odot (dX\mathbf{b})\right)$$

Step2: 套上迹，并做矩阵乘法交换：

$$\begin{aligned} df &= \operatorname{tr}\left(\mathbf{a}^T \left(\exp(X\mathbf{b}) \odot (dX\mathbf{b})\right)\right) \\ &= \operatorname{tr}\left((\mathbf{a} \odot \exp(X\mathbf{b}))^T dX\mathbf{b}\right) \\ &= \operatorname{tr}\left(\mathbf{b}(\mathbf{a} \odot \exp(X\mathbf{b}))^T dX\right) \\ &= \operatorname{tr}\left(\left((\mathbf{a} \odot \exp(X\mathbf{b}))\, \mathbf{b}^T\right)^T dX\right) \end{aligned}$$

Step3：对照导数与微分之间的联系：

$$\frac{\partial f}{\partial X} = (\mathbf{a} \odot \exp(X\mathbf{b}))\, \mathbf{b}^T$$

例3【线性回归】$l = \|X\mathbf{w} - \mathbf{y}\|^2$，求$\mathbf{w}$ 的最小二乘估计。其中$\mathbf{y}$是$m \times 1$ 列向量，X 是$m \times n$矩阵，$\mathbf{w}$ 是$n \times 1$ 列向量，l是标量。

Step1: 将向量模平方改成向量与内积形式：

$$l = (X\mathbf{w} - \mathbf{y})^T (X\mathbf{w} - \mathbf{y})$$

Step1: 使用矩阵乘法法则求微分：

$$
\begin{aligned}
dl &= (Xd\mathbf{w})^T(X\mathbf{w}-\mathbf{y}) + (X\mathbf{w}-\mathbf{y})^T(Xd\mathbf{w}) \\
&= 2(X\mathbf{w}-\mathbf{y})^T(Xd\mathbf{w})
\end{aligned}
$$

$$
\begin{aligned}
&\because Xd\mathbf{w}\ X\mathbf{w}-\mathbf{y} \text{ are coloum vector} \\
&\therefore (Xd\mathbf{w})^T(X\mathbf{w}-\mathbf{y}) = (X\mathbf{w}-\mathbf{y})^T(Xd\mathbf{w})
\end{aligned}
$$

Step2: 套上迹，并做矩阵乘法交换：

$$
dl = \operatorname{tr}\left(\left(2X^T(X\mathbf{w}-\mathbf{y})\right)^T d\mathbf{w}\right)
$$

Step3：对照导数与微分之间的联系：

$$
\frac{\partial l}{\partial \mathbf{w}} = 2X^T(X\mathbf{w}-\mathbf{y})
$$

Step3：求$\mathbf{w}$的最小二乘估计

$$
\begin{aligned}
\frac{\partial l}{\partial \mathbf{w}} &= 2X^T(X\mathbf{w}-\mathbf{y}) = 0 \\
\mathbf{w} &= (X^TX)^{-1}X^T\mathbf{y}
\end{aligned}
$$

6. 向量对向量求导

向量对向量的求导比较麻烦，m 为列向量$\mathbf{y}$ 对n 维列向量$\mathbf{x}$ 求导，那么一共有mn 个标量对标量的求导。

分子布局（numerator layout）：

> 求导结果矩阵的第一个维度以分子为准，结果是一个$m \times n$矩阵，一般叫作**雅克比矩阵**。

分母布局（denominator layout）：

> 求导结果矩阵的第一个维度以分母为准，结果是一个$n \times m$矩阵，一般叫作**梯度矩阵**。

对于机器学习算法原理中的推导，究竟是采用什么布局一般是隐含的，需自己推导。本文以分子布局的雅克比矩阵为主。

6.1 定义法求导

例： $\mathbf{y} = A\mathbf{x}$，其中A 是$n \times m$ 的矩阵，$\mathbf{x}, \mathbf{y}$ 分别是m, n 维例向量。

Step1: 先求$\mathbf{y}$ 的第i 个分量对$\mathbf{x}$ 的第j 个分量的导数：

$$\frac{\partial A_i\mathbf{x}}{\partial \mathbf{x}_j} = \frac{\partial A_{ij}x_j}{\partial \mathbf{x}_j} = A_{ij}$$

Step2: 将每个标量求导结果排列成矩阵，这里用分子布局：

$$\frac{\partial A\mathbf{x}}{\partial \mathbf{x}_j} = A$$

6.2 向量对向量求导的链式法则

若向量之间有这样的依赖关系：$\mathbf{x} \to \mathbf{y} \to \mathbf{z}$，则有下面的链式求导法则：

$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{z}}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$$

7. 矩阵对矩阵求导

7.1 矩阵对矩阵求导定义

$\frac{\partial F}{\partial X}$， $F : p \times q$， $X : m \times n$ 矩阵F中的pq 个元素要分别对矩阵 X 中的mn 个元素求导，那么求导结果一共会有$mnpq$ 个元素。求导结果的排列有很多种。这里只介绍目前主流的做法。

目前主流的矩阵对矩阵求导定义是对矩阵先做向量化，然后再使用向量对向量的求导。而这里的向量化一般是使用列向量化。也就是说，现在我们的矩阵对矩阵求导可以表示为：

$$\frac{\partial F}{\partial X} = \frac{\partial vec\,(F)}{\partial vec\,(X)}$$

$vec(F)$ 的维度是$pq \times 1$ 列向量，$vec(X)$的维度是$mn \times 1$ 列向量。结果使用分母布局，得到一个$mn \times pq$的矩阵。

7.2 微分求导法

利用导数与微分的联系：

$$vec(dF) = \frac{\partial F}{\partial X}^T vec(dX)$$

求解步骤：

1. 使用矩阵微分运算法则对矩阵F 求微分；
2. 做向量化并使用迹技巧将其它项交换至$vec(dX)$的左侧；
3. 根据导数与微分关系，得到矩阵对矩阵的微分结果。

7.3 矩阵向量化的运算法则

性质	法则
线性性质	$vec(A+B)=vec(A)+vec(B)$
矩阵乘法	$vec(AXB)=\left(B^T\otimes A\right)vec(X)$
矩阵转置	$vec(A^T)=K_{mn}vce(A)$
逐元素乘法	$vec(A\odot X)=diag(A)vec(X)$

注：

> 1. $\otimes$：克罗内克(Kronecker)积，$A(m\times n)$ 与 $B(p\times q)$ 的克罗内克积是 $A\otimes B=[A_{ij}B](mp\times nq)$；
> 2. K_{mn}：交换矩阵。若$vec(A)$是$mn\times 1$ 的列向量，则$K_{mn}(mn\times mn)$，将按列优先的向量化变为按行优先的向量化；
> 3. $diag(A)(mn\times mn)$ 是用A 的元素（列优先）排成的对角阵。

7.4 克罗内克积运算法则

1. $(A\otimes B)^T=A^T\otimes B^T$
2. $vec(\mathbf{ab}^T)=\mathbf{b}\otimes\mathbf{a}$
3. $(A\otimes B)(C\otimes D)=AC\otimes BD$
4. $K_{mn}=K_{nm}^T,\quad K_{mn}K_{nm}=I$

7.5 计算实例

参见矩阵求导术（下）

参考

1. 机器学习中的矩阵向量求导(一) 求导定义与求导布局
2. 机器学习中的矩阵向量求导(二) 矩阵向量求导之定义法
3. 机器学习中的矩阵向量求导(三) 矩阵向量求导之微分法
4. 机器学习中的矩阵向量求导(四) 矩阵向量求导链式法则
5. 机器学习中的矩阵向量求导(五) 矩阵对矩阵的求导
6. 矩阵求导术（上）
7. 矩阵求导术（下）