矩阵求导术（下）

长躯鬼侠
数学爱好者

本文承接上篇 zhuanlan.zhihu.com/p/24...，来讲矩阵对矩阵的求导术。使用小写字母x表示标量，粗体小写字母 $\boldsymbol{x}$ 表示列向量，大写字母X表示矩阵。矩阵对矩阵的求导采用了向量化的思路，常应用于二阶方法求解优化问题。

首先来琢磨一下定义。矩阵对矩阵的导数，需要什么样的定义？第一，矩阵F(p×q)对矩阵X(m×n)的导数应包含所有mnpq个偏导数 $\frac{\partial F_{kl}}{\partial X_{ij}}$ ，从而不损失信息；第二，导数与微分有简明的联系，因为在计算导数和应用中需要这个联系；第三，导数有简明的从整体出发的算法。我们先定义向量 $\boldsymbol{f}$ (p×1)对向量 $\boldsymbol{x}$ (m×1)的导数 $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_p}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_p}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_m} & \frac{\partial f_2}{\partial x_m} & \cdots & \frac{\partial f_p}{\partial x_m} \end{bmatrix}$ (m×p)，有 $d\boldsymbol{f} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}^T d\boldsymbol{x}$ ；再定义矩阵的（按列优先）向量化 $\mathrm{vec}(X) = [X_{11}, \ldots, X_{m1}, X_{12}, \ldots, X_{m2}, \ldots, X_{1n}, \ldots, X_{mn}]^T$ (mn×1)，并定义矩阵F对矩阵X的导数 $\frac{\partial F}{\partial X} = \frac{\partial \mathrm{vec}(F)}{\partial \mathrm{vec}(X)}$ (mn×pq)。导数与微分有联系 $\mathrm{vec}(dF) = \frac{\partial F}{\partial X}^T \mathrm{vec}(dX)$ 。几点说明如下：

1. 按此定义，标量f对矩阵X(m×n)的导数 $\frac{\partial f}{\partial X}$ 是mn×1向量，与上篇的定义不兼容，不过二者容易相互转换。为避免混淆，用记号 $\nabla_X f$ 表示上篇定义的m×n矩阵，则有 $\frac{\partial f}{\partial X} = \mathrm{vec}(\nabla_X f)$ 。虽然本篇的技术可以用于标量对矩阵求导这种特殊情况，但使用上篇中的技术更方便。读者可以通过上篇中的算例试验两种方法的等价转换。
2. 标量对矩阵的二阶导数，又称Hessian矩阵，定义为 $\nabla^2_X f = \frac{\partial^2 f}{\partial X^2} = \frac{\partial \nabla_X f}{\partial X}$ (mn×mn)，是对称矩阵。对向量 $\frac{\partial f}{\partial X}$ 或矩阵 $\nabla_X f$ 求导都可以得到Hessian矩阵，但从矩阵 $\nabla_X f$ 出发更方便。
3. $\frac{\partial F}{\partial X} = \frac{\partial \mathrm{vec}(F)}{\partial X} = \frac{\partial F}{\partial \mathrm{vec}(X)} = \frac{\partial \mathrm{vec}(F)}{\partial \mathrm{vec}(X)}$ ，求导时矩阵被向量化，弊端是这在一定程度破坏了矩阵的结构，会导致结果变得形式复杂；好处是多元微积分中关于梯度、Hessian矩阵的结论可以沿用过来，只需将矩阵向量化。例如优化问题中，牛顿法的更新 ΔX，满足 $\mathrm{vec}(\Delta X) = -(\nabla^2_X f)^{-1}\mathrm{vec}(\nabla_X f)$ 。
4. 在资料中，矩阵对矩阵的导数还有其它定义，比如 $\frac{\partial F}{\partial X} = \left[\frac{\partial F_{kl}}{\partial X}\right]$ (mp×nq)，或是 $\frac{\partial F}{\partial X} = \left[\frac{\partial F}{\partial X_{ij}}\right]$ (mp×nq)，它能兼容上篇中的标量对矩阵导数的定义，但微分与导数的联系（dF等于 $\frac{\partial F}{\partial X}$ 中逐个m×n子块分别与dX做内积）不够简明，不便于计算和应用。文献[5]综述了以上定义，并批判它们是“坏”定义，能配合微分运算的才是“好”定义（the procedure based on differentials is superior to other methods of differentiation）。

然后来建立运算法则。仍然要利用导数与微分的联系 $\mathrm{vec}(dF) = \frac{\partial F}{\partial X}^T \mathrm{vec}(dX)$，求微分的方法与上篇相同，而从微分得到导数需要一些向量化的技巧：

1. 线性： $\mathrm{vec}(A+B) = \mathrm{vec}(A) + \mathrm{vec}(B)$ 。
2. 矩阵乘法： $\mathrm{vec}(AXB) = (B^T \otimes A)\mathrm{vec}(X)$ ，其中 $\otimes$ 表示Kronecker积，A(m×n)与B(p×q)的Kronecker积是 $A \otimes B = [A_{ij}B]$ (mp×nq)。此式证明见张贤达《矩阵分析与应用》第107-108页。
3. 转置： $\mathrm{vec}(A^T) = K_{mn}\mathrm{vec}(A)$ ，A是m×n矩阵，其中 K_{mn} (mn×mn)是交换矩阵(commutation

matrix)。

4. 逐元素乘法： $\text{vec}(A\odot X)=\text{diag}(A)\text{vec}(X)$ ，其中 $\text{diag}(A)$ (mn×mn)是用A的元素（按列优先）排成的对角阵。

观察一下可以断言，**若矩阵函数F是矩阵X经加减乘法、逆、行列式、逐元素函数等运算构成，则使用相应的运算法则对F求微分，再做向量化并使用技巧将其它项交换至vec(dX)左侧，对照导数与微分的联系 $\text{vec}(dF)=\frac{\partial F}{\partial X}^T\text{vec}(dX)$，即能得到导数。**

特别地，若矩阵退化为向量，对照导数与微分的联系 $d\boldsymbol{f}=\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}^T d\boldsymbol{x}$，即能得到导数。

再谈一谈复合：假设已求得 $\frac{\partial F}{\partial Y}$ ，而Y是X的函数，如何求 $\frac{\partial F}{\partial X}$ 呢？从导数与微分的联系入手，$\text{vec}(dF)=\frac{\partial F}{\partial Y}^T\text{vec}(dY)=\frac{\partial F}{\partial Y}^T\frac{\partial Y}{\partial X}^T\text{vec}(dX)$，可以推出链式法则 $\frac{\partial F}{\partial X}=\frac{\partial Y}{\partial X}\frac{\partial F}{\partial Y}$ 。

和标量对矩阵的导数相比，矩阵对矩阵的导数形式更加复杂，从不同角度出发常会得到形式不同的结果。有一些Kronecker积和交换矩阵相关的恒等式，可用来做等价变形：

1. $(A\otimes B)^T=A^T\otimes B^T$ 。
2. $\text{vec}(\boldsymbol{a}\boldsymbol{b}^T)=\boldsymbol{b}\otimes\boldsymbol{a}$ 。
3. $(A\otimes B)(C\otimes D)=(AC)\otimes(BD)$ 。可以对 $F=D^TB^TXAC$ 求导来证明，一方面，直接求导得到 $\frac{\partial F}{\partial X}=(AC)\otimes(BD)$ ；另一方面，引入 $Y=B^TXA$ ，有 $\frac{\partial F}{\partial Y}=C\otimes D, \frac{\partial Y}{\partial X}=A\otimes B$ ，用链式法则得到 $\frac{\partial F}{\partial X}=(A\otimes B)(C\otimes D)$ 。
4. $K_{mn}=K_{nm}^T, K_{mn}K_{nm}=I$ 。
5. $K_{pm}(A\otimes B)K_{nq}=B\otimes A$ ，A是m×n矩阵，B是p×q矩阵。可以对 AXB^T 做向量化来证明，一方面， $\text{vec}(AXB^T)=(B\otimes A)\text{vec}(X)$ ；另一方面，$\text{vec}(AXB^T)=K_{pm}\text{vec}(BX^TA^T)=K_{pm}(A\otimes B)\text{vec}(X^T)=K_{pm}(A\otimes B)K_{nq}\text{vec}(X)$ 。

接下来演示一些算例。

例1： $F=AX$，X是m×n矩阵，求 $\frac{\partial F}{\partial X}$ 。

解：先求微分： $dF=AdX$，再做向量化，使用矩阵乘法的技巧，注意在dX右侧添加单位阵：$\text{vec}(dF)=\text{vec}(AdX)=(I_n\otimes A)\text{vec}(dX)$ ，对照导数与微分的联系得到 $\frac{\partial F}{\partial X}=I_n\otimes A^T$ 。

特例：如果X退化为向量，即 $\boldsymbol{f}=A\boldsymbol{x}$ ，则根据向量的导数与微分的关系 $d\boldsymbol{f}=\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}^T d\boldsymbol{x}$，得到 $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}=A^T$ 。

例2： $f=\log|X|$ ，X是n×n矩阵，求 $\nabla_X f$ 和 $\nabla^2_X f$ 。

解：使用上篇中的技术可求得 $\nabla_X f=X^{-1T}$ 。为求 $\nabla^2_X f$，先求微分： $d\nabla_X f=-(X^{-1}dXX^{-1})^T$ ，再做向量化，使用转置和矩阵乘法的技巧 $\text{vec}(d\nabla_X f)=-K_{nn}\text{vec}(X^{-1}dXX^{-1})=-K_{nn}(X^{-1T}\otimes X^{-1})\text{vec}(dX)$ ，对照导数与微分的联系，得到 $\nabla^2_X f=-K_{nn}(X^{-1T}\otimes X^{-1})$ ，注意它是对称矩阵。在 X 是对称矩阵时，可简化为 $\nabla^2_X f=-X^{-1}\otimes X^{-1}$ 。

例3： $F=A\exp(XB)$ ，A是l×m矩阵，X是m×n矩阵，B是n×p矩阵，exp为逐元素函数，求 $\frac{\partial F}{\partial X}$ 。

解：先求微分： $dF=A(\exp(XB)\odot(dXB))$ ，再做向量化，使用矩阵乘法的技巧：

$\text{vec}(dF) = (I_p \otimes A)\text{vec}(\exp(XB) \odot (dXB))$，再用逐元素乘法的技巧：
$\text{vec}(dF) = (I_p \otimes A)\text{diag}(\exp(XB))\text{vec}(dXB)$，再用矩阵乘法的技巧：
$\text{vec}(dF) = (I_p \otimes A)\text{diag}(\exp(XB))(B^T \otimes I_m)\text{vec}(dX)$，对照导数与微分的联系得到
$\frac{\partial F}{\partial X} = (B \otimes I_m)\text{diag}(\exp(XB))(I_p \otimes A^T)$。

例4【一元logistic回归】：$l = -y\boldsymbol{x}^T\boldsymbol{w} + \log(1 + \exp(\boldsymbol{x}^T\boldsymbol{w}))$，求 $\nabla_{\boldsymbol{w}} l$ 和 $\nabla^2_{\boldsymbol{w}} l$。其中 y 是取值0或1的标量，$\boldsymbol{x}, \boldsymbol{w}$ 是 $n \times 1$ 列向量。

解：使用上篇中的技术可求得 $\nabla_{\boldsymbol{w}} l = \boldsymbol{x}(\sigma(\boldsymbol{x}^T\boldsymbol{w}) - y)$，其中 $\sigma(a) = \frac{\exp(a)}{1 + \exp(a)}$ 为sigmoid函数。为求 $\nabla^2_{\boldsymbol{w}} l$，先求微分：$d\nabla_{\boldsymbol{w}} l = \boldsymbol{x}\sigma'(\boldsymbol{x}^T\boldsymbol{w})\boldsymbol{x}^T d\boldsymbol{w}$，其中 $\sigma'(a) = \frac{\exp(a)}{(1 + \exp(a))^2}$ 为sigmoid函数的导数，对照导数与微分的联系，得到 $\nabla^2_w l = \boldsymbol{x}\sigma'(\boldsymbol{x}^T\boldsymbol{w})\boldsymbol{x}^T$。

推广：样本 $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N)$，$l = \sum_{i=1}^N \left(-y_i\boldsymbol{x}_i^T\boldsymbol{w} + \log(1 + \exp(\boldsymbol{x_i}^T\boldsymbol{w}))\right)$，求 $\nabla_w l$ 和 $\nabla^2_w l$。有两种方法，方法一：先对每个样本求导，然后相加；方法二：定义矩阵 $X = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$，向量 $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$，将 l 写成矩阵形式 $l = -\boldsymbol{y}^T X\boldsymbol{w} + \boldsymbol{1}^T \log(\boldsymbol{1} + \exp(X\boldsymbol{w}))$，进而可以使用上篇中的技术求得 $\nabla_{\boldsymbol{w}} l = X^T(\sigma(X\boldsymbol{w}) - \boldsymbol{y})$。为求 $\nabla^2_{\boldsymbol{w}} l$，先求微分，再用逐元素乘法的技巧：
$d\nabla_{\boldsymbol{w}} l = X^T(\sigma'(X\boldsymbol{w}) \odot (Xd\boldsymbol{w})) = X^T\text{diag}(\sigma'(X\boldsymbol{w}))Xd\boldsymbol{w}$，对照导数与微分的联系，得到
$\nabla^2_w l = X^T\text{diag}(\sigma'(X\boldsymbol{w}))X$。

例5【多元logistic回归】：$l = -\boldsymbol{y}^T \log \text{softmax}(W\boldsymbol{x}) = -\boldsymbol{y}^T W\boldsymbol{x} + \log(\boldsymbol{1}^T \exp(W\boldsymbol{x}))$，求 $\nabla_W l$ 和 $\nabla^2_W l$。其中其中 $\boldsymbol{y}$ 是除一个元素为1外其它元素为0的 $m \times 1$ 列向量，W 是 $m \times n$ 矩阵，$\boldsymbol{x}$ 是 $n \times 1$ 列向量，l 是标量。

解：上篇中已求得 $\nabla_W l = (\text{softmax}(W\boldsymbol{x}) - \boldsymbol{y})\boldsymbol{x}^T$。为求 $\nabla^2_W l$，先求微分：定义 $\boldsymbol{a} = W\boldsymbol{x}$，
$d\nabla_W l = \left(\frac{\exp(\boldsymbol{a}) \odot d\boldsymbol{a}}{\boldsymbol{1}^T \exp(\boldsymbol{a})} - \frac{\exp(\boldsymbol{a})(\boldsymbol{1}^T(\exp(\boldsymbol{a}) \odot d\boldsymbol{a}))}{(\boldsymbol{1}^T \exp(\boldsymbol{a}))^2}\right)\boldsymbol{x}^T = \left(\frac{\text{diag}(\exp(\boldsymbol{a}))}{\boldsymbol{1}^T \exp(\boldsymbol{a})} - \frac{\exp(\boldsymbol{a})\exp(\boldsymbol{a})^T}{(\boldsymbol{1}^T \exp(\boldsymbol{a}))^2}\right)d\boldsymbol{a}\boldsymbol{x}^T$
$= \left(\text{diag}(\text{softmax}(\boldsymbol{a})) - \text{softmax}(\boldsymbol{a})\text{softmax}(\boldsymbol{a})^T\right)d\boldsymbol{a}\boldsymbol{x}^T$，注意这里化简去掉逐元素乘法，第一项中 $\exp(\boldsymbol{a}) \odot d\boldsymbol{a} = \text{diag}(\exp(\boldsymbol{a}))d\boldsymbol{a}$，第二项中 $\boldsymbol{1}^T(\exp(\boldsymbol{a}) \odot d\boldsymbol{a}) = \exp(\boldsymbol{a})^T d\boldsymbol{a}$。定义矩阵
$D(\boldsymbol{a}) = \text{diag}(\text{softmax}(\boldsymbol{a})) - \text{softmax}(\boldsymbol{a})\text{softmax}(\boldsymbol{a})^T$，$d\nabla_W l = D(\boldsymbol{a})d\boldsymbol{a}\boldsymbol{x}^T = D(W\boldsymbol{x})dW\boldsymbol{x}\boldsymbol{x}^T$，做向量化并使用矩阵乘法的技巧，得到 $\nabla^2_W l = (\boldsymbol{x}\boldsymbol{x}^T) \otimes D(W\boldsymbol{x})$。

最后做个总结。我们发展了从**整体**出发的矩阵求导的技术，**导数与微分的联系是计算的枢纽**，标量对矩阵的导数与微分的联系是 $df = \text{tr}(\nabla_X^T f dX)$，先对f求微分，再使用迹技巧可求得导数，特别地，标量对向量的导数与微分的联系是 $df = \nabla_{\boldsymbol{x}}^T f d\boldsymbol{x}$；矩阵对矩阵的导数与微分的联系是 $\text{vec}(dF) = \frac{\partial F}{\partial X}^T \text{vec}(dX)$，先对F求微分，再使用向量化的技巧可求得导数，特别地，向量对向量的导数与微分的联系是 $d\boldsymbol{f} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}^T d\boldsymbol{x}$。

参考资料：

1. 张贤达. *矩阵分析与应用*. 清华大学出版社有限公司, 2004.
2. Fackler, Paul L. "Notes on matrix calculus." *North Carolina State University*(2005).
3. Petersen, Kaare Brandt, and Michael Syskind Pedersen. "The matrix cookbook." *Technical University of Denmark* 7 (2008): 15.
4. HU, Pili. "Matrix Calculus: Derivation and Simple Application." (2012).

5. Magnus, Jan R., and Heinz Neudecker. "Matrix differential calculus with applications to simple, Hadamard, and Kronecker products."*Journal of Mathematical Psychology*29.4 (1985): 474-492.

编辑于昨天 03:04

矩阵分析 机器学习 优化

118 条评论 ⇄ 切换为时间排序

写下你的评论...

James Liu 2 年前
先赞为敬
1

方不觉 2 年前
赞
赞

王赟 Maigo 2 年前
矩阵对矩阵求导的结果中有好多Kronecker积啊……不过也没办法，把一个本来是四维的东西压成两维已经不容易了。
8

知乎用户 2 年前
我只是想吐槽知乎这个狗屎编辑器到现在都不支持 MathJax，真是废物。怕是坦桑尼亚的民族风情网站都支持 MathJax 了。可能知乎的前端一没高分屏，二没大学文凭。
13

Shinji Fujiwara 2 年前
赞
赞

排骨郎 2 年前
例3中，与B的转置做 kronecker product 的单位矩阵应该是 I_m不是I_n

1

长躯鬼侠 (作者) 回复 排骨郎 2 年前

多谢指正

赞

知乎用户 2 年前

nice

赞

李作新 2 年前

例3的公式没有渲染出来

赞

张秦川 1 年前

求问向量对矩阵的求导是否也适用这套方法?

赞

长躯鬼侠 (作者) 回复 张秦川 1 年前

适用啊 向量可以看成矩阵的特例

赞

f8411cjh 回复 长躯鬼侠 (作者) 4 个月前

如果x是向量，是否有VEC(x)=VEC(xT)?

赞

展开其他 1 条回复

DreamYun 1 年前

看完上下篇，是否可以总结出如下：对于复合函数求导，如果是标量函数对矩阵求导，没有链式法则可用；如果是矩阵对矩阵求导，有链式法则可以套用。

赞

长躯鬼侠 (作者) 回复 DreamYun 1 年前

你可以这么理解。不过链式法则就是源自多次求微分，所以只是形式不同，没有本质的区别。

1

DreamYun 1 年前

或者可以这么表达：如果不论标量还是矩阵（包括向量）对矩阵的求导，如果是按照篇二的做法，首先都列向量化（vec），然后求导。那么对于这种形式的求导，是可以适用复合函数的链式求导法则。其它形式的求导方法，可能不适用复合函数求导链式法则。

1

陌烛 1 年前

你好，请问下，为何例二中，f对X的二阶导没有进行转置？在原文（例二中）：“对照导数与微分的关系得到......”后面的那个式子

赞

长躯鬼侠 (作者) 回复 陌烛 1 年前

是对称矩阵，转置等于它自己。

1

陌烛 1 年前

还有，请问下，我怎么确定我的转换矩阵的值是多少啊?

赞

陌烛 1 年前

你好，请问下，原文中有句话：“若矩阵函数F是矩阵X经加减乘法、行列式、逆、逐元素函数等运算构成，则使用相应的运算法则对F求微分，再做向量化并使用技巧将其它项交换至vec(dX)左侧，即能得到导数”，那么如果F是由X卷积操作得到的，那么，对于这个卷积的运算法则是什么呢？👀👀👀👀

赞

长躯鬼侠 (作者) 回复 陌烛 1 年前

对于卷积，你可以自己推导一下，运算法则也可以用卷积来表示，对full、valid模式在细节上有些差异。

1

陌烛 1 年前

你好，我还想知道下，克罗内克积和矩阵乘积，哪个的优先级大啊？😢

赞

长躯鬼侠 (作者) 回复 陌烛 1 年前

我没有指定Kronecker积和矩阵乘积哪个优先级高，所以都加括号了啊。

赞

陌烛 回复 长躯鬼侠 (作者) 1 年前

蟹蟹，我果然不适合推导数学公式😭

赞

查看全部 9 条回复

孙培钦 1 年前

我想问一下，假设我有个等式 S = WX， S是m x 1向量， X是 n x 1向量，W是m x n矩阵，使用上述的求导术去求S向量对X向量的导数， 我得到的是 一个n x n的单位阵与W转置的kronecker积啊，这个积的尺寸应该是nn x mn， 明显不对啊。

赞

长躯鬼侠 (作者) 回复 孙培钦 1 年前

ds = WdxI，x的右面是1×1的单位阵，不是n×n的。

1

暮日落流年 1 年前

收益匪浅，赞一个！

赞

知乎用户 1 年前

能区分一下行列求导就更完美了，同时使用线性变换解释一下导数与微分的关系，可能更加恰当，和利于直觉。

1

neilfvhv 1 年前

问个问题，\frac {\partial F} {\partial X} = \frac {\partial vecF} {\partial vec X} 这一步是怎么得到的？

赞

长躯鬼侠 (作者) 回复 neilfvhv 1 年前

这是定义

赞

张云 1 年前

你好，不怎么懂那个例四推广里的log的和是怎么转化为矩阵形式的，前一半可以理解，后一半就有点懵了

赞

长躯鬼侠 (作者) 回复 张云 1 年前

你是说\sum_i log(1 + exp(\vec{x_i}^T \vec{w})) = \vec{1}^T log(\vec{1} + exp(X\vec{w})) 这个嘛？因为log(1 + exp(\vec{x_i}^T \vec{w}))是向量log(\vec{1} + exp(X\vec{w}))的第i个元素，左边的求和就等于右边的\vec{1}^T乘以向量。

赞

张云 回复 长躯鬼侠 (作者) 1 年前

懂了。那个1的向量的转置是为了把矩阵里面的每个元素都求和。

哇，这些算是本科学的吗？感觉有点酷炫诶😂

赞

展开其他 1 条回复