知乎

矩阵求导术（下）

长躯鬼侠
数学爱好者

已关注

398 人赞了该文章

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 $\nabla_X^2 f=\frac{\partial^2 f}{\partial X^2}=\frac{\partial \nabla_X f}{\partial X}$ (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

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)，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 $\frac{\partial F}{\partial X}$ 中每个m×n子块分别与dX做内积）不够简明，不便于计算和应用。

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： $\mathrm{vec}(A\odot X)=\mathrm{diag}(A)\mathrm{vec}(X)$ ，其中 $\mathrm{diag}(A)$ (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、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。**

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，$\mathrm{vec}(dF)=\frac{\partial F}{\partial Y}^T\mathrm{vec}(dY)=\frac{\partial F}{\partial Y}^T\frac{\partial Y}{\partial X}^T\mathrm{vec}(dX)$，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. $\mathrm{vec}(ab^T)=b\otimes a$ 。
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

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， $\mathrm{vec}(AXB^T)=(B\otimes A)\mathrm{vec}(X)$ ；另一方面，
$\mathrm{vec}(AXB^T)=K_{pm}\mathrm{vec}(BX^TA^T)=K_{pm}(A\otimes B)\mathrm{vec}(X^T)=K_{pm}(A\otimes B)K_{nq}\mathrm{vec}(X)$
。

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：
$\mathrm{vec}(dF)=\mathrm{vec}(AdX)=(I_n\otimes A)\mathrm{vec}(dX)$ ，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， $\boldsymbol{f}=A\boldsymbol{x}$ ，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 $\nabla_X^2 f$ 。

解：使用上篇中的技术可求得 $\nabla_X f=X^{-1T}$ 。为求 $\nabla_X^2 f$ ，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
$\mathrm{vec}(d\nabla_X f)=-K_{nn}\mathrm{vec}(X^{-1}dXX^{-1})=-K_{nn}(X^{-1T}\otimes X^{-1})\mathrm{vec}(dX)$ ，对照导数与微分的联系，得到 $\nabla_X^2 f=-K_{nn}(X^{-1T}\otimes X^{-1})$ ，注意它是对称矩阵。在 X 是对称矩阵时，可简化为 $\nabla_X^2 f=-X^{-1}\otimes X^{-1}$ 。

例3： $F=A\exp(XB)$ ，A是l×m，X是m×n，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：
$\mathrm{vec}(dF)=(I_p\otimes A)\mathrm{vec}(\exp(XB)\odot(dXB))$ ，再用逐元素乘法的技巧：
$\mathrm{vec}(dF)=(I_p\otimes A)\mathrm{diag}(\exp(XB))\mathrm{vec}(dXB)$ ，再用矩阵乘法的技巧：
$\mathrm{vec}(dF)=(I_p\otimes A)\mathrm{diag}(\exp(XB))(B^T\otimes I_m)\mathrm{vec}(dX)$ ，对照导数与微分的联系得到
$\frac{\partial F}{\partial X}=(B\otimes I_m)\mathrm{diag}(\exp(XB))(I_p\otimes A^T)$ 。

例4【一元logistic回归】： $l=-yx^Tw+\log(1+\exp(x^Tw))$，求 $\nabla_w l$ 和 $\nabla_w^2 l$。其中 y 是取值0或1的标量， x,w 是向量。

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

推广：样本 $(x_1,y_1),\dots,(x_n,y_n)$， $l=\sum_{i=1}^N\left(-y_ix_i^Tw+\log(1+\exp(x_i^Tw))\right)$，求 $\nabla_w l$ 和 $\nabla_w^2 l$。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 $X=\begin{bmatrix}x_1^T\\ \vdots\\ x_n^T\end{bmatrix}$，向量 $y=\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}$，将 l 写成矩阵形式 $l=-y^TXw+1^T\log(1+\exp(Xw))$，进而可以求得 $\nabla_w l=X^T(\sigma(Xw)-y)$， $\nabla_w^2 l=X^T\mathrm{diag}(\sigma'(Xw))X$。

例5【多元logistic回归】： $l=-y^T\log\mathrm{softmax}(Wx)=-y^TWx+\log(1^T\exp(Wx))$，求 $\nabla_W l$ 和 $\nabla_W^2 l$。

解：上篇例3中已求得 $\nabla_W l=(\mathrm{softmax}(Wx)-y)x^T$ 。为求 $\nabla_W^2 l$，先求微分：定义 $a=Wx$， $d\mathrm{softmax}(a)=\dfrac{\exp(a)\odot da}{1^T\exp(a)}-\dfrac{\exp(a)(1^T(\exp(a)\odot da))}{(1^T\exp(a))^2}$ ，这里需要化简去掉逐元素乘法，第一项中 $\exp(a)\odot da=\mathrm{diag}(\exp(a))da$ ，第二项中 $1^T(\exp(a)\odot da)=\exp(a)^Tda$，故有 $d\mathrm{softmax}(a)=\mathrm{softmax}'(a)da$，其中 $\mathrm{softmax}'(a)=\dfrac{\mathrm{diag}(\exp(a))}{1^T\exp(a)}-\dfrac{\exp(a)\exp(a)^T}{(1^T\exp(a))^2}$，代入有 $d\nabla_W l=\mathrm{softmax}'(a)dax^T=\mathrm{softmax}'(Wx)dWxx^T$ ，做向量化并使用矩阵乘法的技巧，得到 $\nabla_W^2 l=(xx^T)\otimes\mathrm{softmax}'(Wx)$。

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 $df=\mathrm{tr}(\nabla_X^T f dX)$ ，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 $df=\nabla_x^T f dx$；矩阵对矩阵的导数与微分的联系是

向量的导数与微分的联系是 $d\boldsymbol{f} = \frac{\partial \boldsymbol{f}^T}{\partial \boldsymbol{x}} d\boldsymbol{x}$。

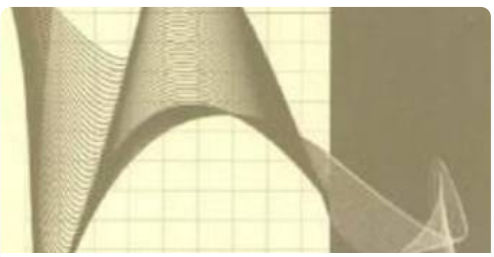
参考资料：

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).

编辑于 2018-08-11

矩阵分析　机器学习　优化

推荐阅读



机器学习之——自动求导

章华燕

如何直观地理解「协方差矩阵」？

协方差矩阵在统计学和机器学习中随处可见，一般而言，可视作方差和协方差两部分组成，即方差构成了对角线上的元素，协方差构成了非对角线上的元素。本文旨在从几何角度介绍我们所熟知的协方...

Xinyu Chen

机器学

“矩阵求
域。虽然
讲的多元
公式以及
必要的，
且易出错

Towser

59 条评论

切换为时间排序

写下你的评论...

知乎

先赞为敬

赞

 方不觉 1 年前

赞

赞

 王赟 Maigo 1 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。

6

元気 李元气 1 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。

7

 Shinji Fujiwara 1 年前

赞

赞

 排骨郎 1 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 1 年前

多谢指正

赞 查看对话

 几许风雨 1 年前

nice

赞

 李作新 1 年前

例3的公式没有渲染出来

赞

求向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 查看对话

仡佬佬 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 仡佬佬 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 查看对话

仡佬佬 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。

赞

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

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

知乎

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

是对称矩阵，转置等于它自己。

赞 查看对话

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。

赞 查看对话

1 2 3 下一页