矩阵求导

YWJ

2017年7月14日

1 Introduction

矩阵运算在机器学习算法大量出现，其中矩阵求导运算在使用梯度的算法中尤其常见。本文从标量对矩阵求导的定义出发，详细推导了标量，向量，矩阵对矩阵求导的基本公式以及一些运算性质。并在最后讨论了一些尚不明确的规则。

2 矩阵求导的基本定义

2.1 标量对矩阵求导的定义

矩阵对标量求导的意义是明确的，即对矩阵中的每个元素分别对标量求导，导数摆放在被求导元素的原位置组成结果矩阵。

$$\frac{\partial A}{\partial x} = \begin{pmatrix} \frac{\partial a_{11}}{\partial x} & \frac{\partial a_{12}}{\partial x} & \cdots & \frac{\partial a_{1n}}{\partial x} \\ \frac{\partial a_{21}}{\partial x} & \frac{\partial a_{22}}{\partial x} & \cdots & \frac{\partial a_{2n}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial a_{m1}}{\partial x} & \frac{\partial a_{m2}}{\partial x} & \cdots & \frac{\partial a_{mn}}{\partial x} \end{pmatrix}$$

据此可以对称的定义标量对矩阵的求导规则，即标量对矩阵中的每个元素分别求导，导数摆放在被求导元素的原位置组成结果矩阵。

$$\frac{\partial a}{\partial X} = \begin{pmatrix} \frac{\partial a}{\partial x_{11}} & \frac{\partial a}{\partial x_{12}} & \cdots & \frac{\partial a}{\partial x_{1n}} \\ \frac{\partial a}{\partial x_{21}} & \frac{\partial a}{\partial x_{22}} & \cdots & \frac{\partial a}{\partial x_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial a}{\partial x_{m1}} & \frac{\partial a}{\partial x_{m2}} & \cdots & \frac{\partial a}{\partial x_{mn}} \end{pmatrix}$$

2.2 转置或不转置

对矩阵求导定义的一些分歧在于求导结果是否转置，即$\left[\frac{\partial a}{\partial X}\right]_{ij} = \frac{\partial a}{\partial x_{ij}}$或者$\left[\frac{\partial a}{\partial X}\right]_{ij} = \frac{\partial a}{\partial x_{ji}}$,但这并不影响矩阵求导的一般规律，两种定义下的推导过程几乎是完全对称的，即在一种定义下（转置）适用的规则在另一种定义下（不转置）也同样适用。

2.3 约定俗成的规则

一些默认情况下使用的书写规则：

1. 标量对矩阵求导结果不转置
2. 矩阵对矩阵求导结果要转置。
3. 没有转置号的向量默认为列向量，有转置号的向量默认为行向量

第二条实际上是使用第一条默认规则的必然结果。下文中的推导均使用上面的默认规则。

3 矩阵对矩阵求导

3.1 最基本的求导规则

$$\frac{\partial XA}{\partial X} = A^T$$

X是矩阵，A可以是标量，或与X适配的向量，矩阵。

证明：

A是标量时由上一节给出的定义可以直接得到结论，下面讨论A是向量时的情况。XA总是一个列向量，总有一个行向量B^T使B^TXA成为一个标量，那么有：

$$B^TXA = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}^T \cdot \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \sum_{i,j} x_{ij} \cdot (a_j b_i)$$

对这个标量求矩阵X的偏导,由定义有：

$$\frac{\partial B^TXA}{\partial X} = \begin{pmatrix} a_1b_1 & a_2b_1 & \cdots & a_nb_1 \\ a_1b_2 & a_2b_2 & \cdots & a_nb_2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1b_m & a_2b_m & \cdots & a_nb_m \end{pmatrix} = BA^T$$

由链式法则有：

$$\frac{\partial B^TXA}{\partial X} = \frac{\partial B^T(XA)}{\partial (XA)} \cdot \frac{\partial XA}{\partial X} = BA^T$$

注意到(XA)是列向量，不妨记作C，由标量对矩阵求导规则有：

$$B^T(XA) = B^TC = \sum_i b_i c_i$$

$$\frac{\partial B^T(XA)}{\partial (XA)} = \frac{\partial B^TC}{\partial C} = \begin{pmatrix} \frac{\partial \sum_i b_i c_i}{\partial c_1} \\ \frac{\partial \sum_i b_i c_i}{\partial c_2} \\ \vdots \\ \frac{\partial \sum_i b_i c_i}{\partial c_m} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = B$$

带入链式求导等式有：

$$\frac{\partial B^T(XA)}{\partial (XA)} \cdot \frac{\partial XA}{\partial X} = B \cdot \frac{\partial XA}{\partial X} = BA^T \Rightarrow \frac{\partial XA}{\partial X} = A^T$$

注意这个式子仅在B为非零向量时成立。

接下来讨论A是矩阵时的情况，同样总可以找到非零的列向量C^T使AC成为列向量，根据上文已经推导出的A是列向量时的求导规则：

$$\frac{\partial X(AC)}{\partial X} = (AC)^T = C^T A^T$$

注意到A是列向量时的求导规则同样可以对C套用:

$$\frac{\partial (XA)C}{\partial (XA)} = C^T$$

引入链式法则:

$$C^T \frac{\partial XA}{\partial X} = \frac{\partial (XA)C}{\partial (XA)} \cdot \frac{\partial XA}{\partial X} = \frac{\partial X(AC)}{\partial X} = C^T A^T$$

$$\Rightarrow \frac{\partial XA}{\partial X} = A^T$$

至此，A是标量，向量，矩阵情况下的XA对矩阵X的求导规则以全部证毕。

类似的还可以证明：

$$\frac{\partial BX}{\partial X} = B^T$$

3.2 转置的引入

$$\left(\frac{\partial XA}{\partial X}\right)^T = \left(\frac{\partial (AX)^T}{\partial X^T}\right) = A$$

证明：

$$\frac{\partial (AX)^T}{\partial X^T} = \frac{\partial X^T A^T}{\partial X^T} = A$$

$$\left(\frac{\partial Ax}{\partial x}\right)^{\mathrm{T}} = (A^T)^T = A$$

3.3 自变量居中

$$\frac{\partial B^T XA}{\partial X} = BA^T$$

证明：

由链式法则以及简单求导规则：

$$\frac{\partial B^T XA}{\partial X} = \frac{\partial B^T(XA)}{\partial (XA)} \cdot \frac{\partial (XA)}{\partial X} = BA^T$$

3.4 转置对非转置

被求导项为标量时，有

$$\frac{\partial XA}{\partial X^T} = \left(\frac{\partial XA}{\partial X}\right)^T$$

或者说标量对自变量矩阵或自变量矩阵的转置求导的结果互为转置。

证明：

$$\frac{\partial XA}{\partial X^T} = \frac{\partial A^TX^T}{\partial X^T} = \frac{\partial (XA)^T}{\partial (X^T)^T} = \left(\frac{\partial XA}{\partial X}\right)^T$$

但似乎没有办法证明对于被求导项为非标量时这个性质仍然成立。但考虑到机器学习中需要求梯度的损失函数几乎都是标量值，这一点就无关紧要了。

3.5 含多个自变量的项

按标量函数求导规则对每个自变量求导取和：

$$\begin{aligned}
&\tfrac{\partial B^TX^TXA}{\partial X}\\
&= \left(\tfrac{\partial B^TX^T(XA)_C}{\partial X^T}\right)^T + \tfrac{\partial (B^TX^T)_C XA}{\partial X}\\
&= (B(XA)^T)^T + XBA^T\\
&= X(AB^T + BA^T)\\
&\tfrac{\partial X^TAX}{\partial X}\\
&= \tfrac{\partial (X^TA)_C X}{\partial X} + \left(\tfrac{\partial X^T(AX)_C}{\partial X^T}\right)^T\\
&= (X^TA)^T + AX\\
&= (A + A^T)X
\end{aligned}$$

括号下标C意味着括号中的项看作常数（这是一种不规范的记法）。

4 一些后话

关于转置对非转置被求导项为非标量的情况，这里有一些不严谨的推导，可以从侧面反映性质3.4似乎是仍然成立的。例如求：

$$\frac{\partial A^TX}{\partial X^T}$$

被求导项转化为标量：

$$\frac{\partial A^TXB}{\partial X^T} = \frac{\partial B^TX^TA}{\partial X^T} = \frac{\partial B^T(X^TA)}{\partial (X^TA)} \cdot \frac{\partial X^TA}{\partial X^T} = BA^T$$

$$\frac{\partial A^TXB}{\partial X^T} = \frac{\partial X}{\partial X^T} \cdot \frac{\partial A^TX}{\partial X} \cdot \frac{\partial A^TXB}{\partial A^TX} = \frac{\partial X}{\partial X^T} \cdot AB^T = BA^T$$

可以看见

$$\frac{\partial X}{\partial X^T}$$

的作用类似于一个转置算子，一项乘以它相当于把这一项整体做了转置，在其它的计算中这条规则似乎也是适用的，但转置运算不是个初等函数，讨论它的导数似乎是不可行的，因此这里无法构造性的证明对所有运算这条性质都成立。

最后是链式法则的分解问题，实际上分解后的两项是不可以交换位置的，不然会导致求导结果形状改变。上面的推导中有时把子项摆在主项左边，有时摆在右边，这是根据预估求导结果矩阵的形状而做出的调整，例如：

$$\frac{\partial A^T XB}{\partial X^T} = \frac{\partial X}{\partial X^T} \cdot \frac{\partial A^T X}{\partial X} \cdot \frac{\partial A^T XB}{\partial A^T X}$$

如果交换分解的左右位置会得到一个标量，标量对矩阵求导为标量这显然不合理。再例如：

$$\frac{\partial A^T XB}{\partial X^T} = \frac{\partial B^T (X^T A)}{\partial (X^T A)} \cdot \frac{\partial X^T A}{\partial X^T}$$

如果交换子项位置也会得到标量对矩阵求导是标量的结果，实际上这里似乎无法总结出关于特定的式子到底应该左分解还是右分解的一般规律，尝试了很多规则，但都被特例推翻了。因此这里直接考虑求导结果的形状来调整位置可能是最有效的。