矩阵求导术（上）

长躯鬼侠
数学爱好者

矩阵求导的技术，在统计学、控制论、机器学习等领域有广泛的应用。鉴于我看过的一些资料或言之不详、或繁乱无绪，本文来做个科普，分作两篇，上篇讲标量对矩阵的求导术，下篇讲矩阵对矩阵的求导术。本文使用小写字母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。

首先来琢磨一下定义，标量f对矩阵X的导数，定义为 $\frac{\partial f}{\partial X} = \left[\frac{\partial f}{\partial X_{ij}}\right]$ ，即f对X逐元素求导排成与X尺寸相同的矩阵。然而，这个定义在计算中并不好用，实用上的原因是对函数较复杂的情形难以逐元素求导；哲理上的原因是逐元素求导破坏了**整体性**。试想，为何要将f看做矩阵X而不是各元素 X_{ij} 的函数呢？答案是用矩阵运算更整洁。所以在求导时不宜拆开矩阵，而是要找一个从整体出发的算法。

为此，我们来回顾，一元微积分中的导数（标量对标量的导数）与微分有联系： $df = f'(x)dx$ ；多元微积分中的梯度（标量对向量的导数）也与微分有联系： $df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}dx_i = \frac{\partial f}{\partial \boldsymbol{x}}^T d\boldsymbol{x}$ ，这里第一个等号是全微分公式，第二个等号表达了梯度与微分的联系：全微分 df 是梯度向量 $\frac{\partial f}{\partial \boldsymbol{x}}$ (n×1)与微分向量 $d\boldsymbol{x}$ (n×1)的内积；受此启发，我们将矩阵导数与微分建立联系： $df = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial f}{\partial X_{ij}}dX_{ij} = \text{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$ 。其中tr代表迹(trace)是方阵对角线元素之和，满足性质：对尺寸相同的矩阵A,B， $\text{tr}(A^TB) = \sum_{i,j} A_{ij}B_{ij}$ ，即 $\text{tr}(A^TB)$ 是矩阵A,B的**内积**。与梯度相似，这里第一个等号是全微分公式，第二个等号表达了矩阵导数与微分的联系：全微分 df 是导数 $\frac{\partial f}{\partial X}$ (m×n)与微分矩阵 dX (m×n)的内积。

运用这些法则。故而，我们来创立常用的矩阵微分的运算法则：

1. 加减法：$d(X \pm Y) = dX \pm dY$；矩阵乘法：$d(XY) = (dX)Y + XdY$；转置：$d(X^T) = (dX)^T$；迹：$d\text{tr}(X) = \text{tr}(dX)$。
2. 逆：$dX^{-1} = -X^{-1}dXX^{-1}$。此式可在 $XX^{-1} = I$ 两侧求微分来证明。
3. 行列式：$d|X| = \text{tr}(X^{\#}dX)$，其中 $X^{\#}$ 表示X的伴随矩阵，在X可逆时又可以写作 $d|X| = |X|\text{tr}(X^{-1}dX)$。此式可用Laplace展开来证明，详见张贤达《矩阵分析与应用》第279页。
4. 逐元素乘法：$d(X \odot Y) = dX \odot Y + X \odot dY$，$\odot$ 表示尺寸相同的矩阵X,Y逐元素相乘。
5. 逐元素函数：$d\sigma(X) = \sigma'(X) \odot dX$，$\sigma(X) = [\sigma(X_{ij})]$ 是逐元素标量函数运算，$\sigma'(X) = [\sigma'(X_{ij})]$ 是逐元素求导数。例如 $X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}, d\sin(X) = \begin{bmatrix} \cos X_{11} dX_{11} & \cos X_{12} dX_{12} \\ \cos X_{21} dX_{21} & \cos X_{22} dX_{22} \end{bmatrix} = \cos(X) \odot dX$。

我们试图利用矩阵导数与微分的联系 $df = \text{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$，在求出左侧的微分 df 后，该如何写成右侧的形式并得到导数呢？这需要一些迹技巧(trace trick)：

1. 标量套上迹：$a = \text{tr}(a)$
2. 转置：$\text{tr}(A^T) = \text{tr}(A)$。
3. 线性：$\text{tr}(A \pm B) = \text{tr}(A) \pm \text{tr}(B)$。
4. 矩阵乘法交换：$\text{tr}(AB) = \text{tr}(BA)$，其中 A 与 B^T 尺寸相同。两侧都等于 $\sum_{i,j} A_{ij}B_{ji}$。
5. 矩阵乘法/逐元素乘法交换：$\text{tr}(A^T(B \odot C)) = \text{tr}((A \odot B)^T C)$，其中 A, B, C 尺寸相同。两侧都等于 $\sum_{i,j} A_{ij}B_{ij}C_{ij}$。

观察一下可以断言，**若标量函数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，再使用迹技巧给df套上迹并将其它项交换至dX左侧，对照导数与微分的联系 $df = \text{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$，即能得到导数。**

在建立法则的最后，来谈一谈复合：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？在微积分中有标量求导的链式法则 $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y}\frac{\partial y}{\partial x}$ ，但这里我们**不能随意沿用标量的链式法则**，因为矩阵对矩阵的导数 $\frac{\partial Y}{\partial X}$ 截至目前仍是未定义的。于是我们继续追本溯源，链式法则是从何而来？源头仍然是微分。我们直接从微分入手建立复合法则：先写出 $df = \text{tr}\left(\frac{\partial f}{\partial Y}^T dY\right)$ ，再将dY用dX表示出来代入，并使用迹技巧将其他项交换至dX左侧，即可得到 $\frac{\partial f}{\partial X}$ 。

最常见的情形是 $Y = AXB$，此时
$df = \text{tr}\left(\frac{\partial f}{\partial Y}^T dY\right) = \text{tr}\left(\frac{\partial f}{\partial Y}^T AdXB\right) = \text{tr}\left(B\frac{\partial f}{\partial Y}^T AdX\right) = \text{tr}\left((A^T\frac{\partial f}{\partial Y}B^T)^T dX\right)$ ，可得到 $\frac{\partial f}{\partial X} = A^T\frac{\partial f}{\partial Y}B^T$ 。注意这里 $dY = (dA)XB + AdXB + AXdB = AdXB$，由于 A, B 是常量，$dA = 0, dB = 0$，以及我们使用矩阵乘法交换的迹技巧交换了 $\frac{\partial f}{\partial Y}^T AdX$ 与 B。

接下来演示一些算例。特别提醒要依据已经建立的运算法则来计算，不能随意套用微积分中标量导数的结论，比如认为AX对X的导数为A，这是没有根据、意义不明的。

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

解：先使用矩阵乘法法则求微分， $df = d\boldsymbol{a}^T X\boldsymbol{b} + \boldsymbol{a}^T dX\boldsymbol{b} + \boldsymbol{a}^T Xd\boldsymbol{b} = \boldsymbol{a}^T dX\boldsymbol{b}$，注意这里的 $\boldsymbol{a}, \boldsymbol{b}$ 是常量， $d\boldsymbol{a} = \boldsymbol{0}, d\boldsymbol{b} = \boldsymbol{0}$。由于df是标量，它的迹等于自身， $df = \text{tr}(df)$ ，套上迹并做矩阵乘法交换： $df = \text{tr}(\boldsymbol{a}^T dX\boldsymbol{b}) = \text{tr}(\boldsymbol{b}\boldsymbol{a}^T dX) = \text{tr}((\boldsymbol{a}\boldsymbol{b}^T)^T dX)$ ，注意这里我们根据 $\text{tr}(AB) = \text{tr}(BA)$ 交换了 $\boldsymbol{a}^T dX$ 与 $\boldsymbol{b}$。对照导数与微分的联系 $df = \text{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$ ，得到 $\frac{\partial f}{\partial X} = \boldsymbol{a}\boldsymbol{b}^T$ 。

合法的）。有些资料在计算矩阵导数时，会略过求微分这一步，这是逻辑上解释不通的。

例2：$f = \boldsymbol{a}^T \exp(X\boldsymbol{b})$，求 $\frac{\partial f}{\partial X}$。其中$\boldsymbol{a}$是$m \times 1$列向量，$X$是$m \times n$矩阵，$\boldsymbol{b}$是$n \times 1$列向量，exp表示逐元素求指数，$f$是标量。

解：先使用矩阵乘法、逐元素函数法则求微分：$df = \boldsymbol{a}^T(\exp(X\boldsymbol{b}) \odot (dX\boldsymbol{b}))$，再套上迹并做交换：$df = \text{tr}(\boldsymbol{a}^T(\exp(X\boldsymbol{b}) \odot (dX\boldsymbol{b}))) = \text{tr}((\boldsymbol{a} \odot \exp(X\boldsymbol{b}))^T dX\boldsymbol{b})$ $= \text{tr}(\boldsymbol{b}(\boldsymbol{a} \odot \exp(X\boldsymbol{b}))^T dX) = \text{tr}(((\boldsymbol{a} \odot \exp(X\boldsymbol{b}))\boldsymbol{b}^T)^T dX)$，注意这里我们先根据 $\text{tr}(A^T(B \odot C)) = \text{tr}((A \odot B)^T C)$ 交换了$\boldsymbol{a}$、$\exp(X\boldsymbol{b})$与$dX\boldsymbol{b}$，再根据 $\text{tr}(AB) = \text{tr}(BA)$ 交换了$(\boldsymbol{a} \odot \exp(X\boldsymbol{b}))^T dX$与$\boldsymbol{b}$。对照导数与微分的联系 $df = \text{tr}\left({\frac{\partial f}{\partial X}}^T dX\right)$，得到 $\frac{\partial f}{\partial X} = (\boldsymbol{a} \odot \exp(X\boldsymbol{b}))\boldsymbol{b}^T$。

例3：$f = \text{tr}(Y^T M Y), Y = \sigma(WX)$，求 $\frac{\partial f}{\partial X}$。其中$W$是$l \times m$矩阵，$X$是$m \times n$矩阵，$Y$是$l \times n$矩阵，$M$是$l \times l$对称矩阵，$\sigma$是逐元素函数，$f$是标量。

解：先求 $\frac{\partial f}{\partial Y}$，求微分，使用矩阵乘法、转置法则：

$df = \text{tr}((dY)^T M Y) + \text{tr}(Y^T M dY) = \text{tr}(Y^T M^T dY) + \text{tr}(Y^T M dY) = \text{tr}(Y^T(M + M^T)dY)$，对照导数与微分的联系，得到 $\frac{\partial f}{\partial Y} = (M + M^T)Y = 2MY$，注意这里M是对称矩阵。

为求 $\frac{\partial f}{\partial X}$，写出 $df = \text{tr}\left({\frac{\partial f}{\partial Y}}^T dY\right)$，再将dY用dX表示出来代入，并使用矩阵乘法/逐元素乘法交换：

$df = \text{tr}\left({\frac{\partial f}{\partial Y}}^T (\sigma'(WX) \odot (WdX))\right) = \text{tr}\left(\left(\frac{\partial f}{\partial Y} \odot \sigma'(WX)\right)^T WdX\right)$，对照导数与微分的联系，得到

$\frac{\partial f}{\partial X} = W^T\left(\frac{\partial f}{\partial Y} \odot \sigma'(WX)\right) = W^T((2M\sigma(WX)) \odot \sigma'(WX))$。

$m\times 1$列向量，X是$m\times n$矩阵，$\boldsymbol{w}$是$n\times 1$列向量，l是标量。

解：这是标量对向量的导数，不过可以把向量看做矩阵的特例。先将向量模平方改写成向量与自身的内积：$l=(X\boldsymbol{w}-\boldsymbol{y})^T(X\boldsymbol{w}-\boldsymbol{y})$，求微分，使用矩阵乘法、转置等法则：$dl=(Xd\boldsymbol{w})^T(X\boldsymbol{w}-\boldsymbol{y})+(X\boldsymbol{w}-\boldsymbol{y})^T(Xd\boldsymbol{w})=2(X\boldsymbol{w}-\boldsymbol{y})^TXd\boldsymbol{w}$，注意这里$Xd\boldsymbol{w}$和$X\boldsymbol{w}-\boldsymbol{y}$是向量，两个向量的内积满足$\boldsymbol{u}^T\boldsymbol{v}=\boldsymbol{v}^T\boldsymbol{u}$。对照导数与微分的联系$dl=\frac{\partial l}{\partial \boldsymbol{w}}^T d\boldsymbol{w}$，得到$\frac{\partial l}{\partial \boldsymbol{w}}=2X^T(X\boldsymbol{w}-\boldsymbol{y})$。$\frac{\partial l}{\partial \boldsymbol{w}}=\boldsymbol{0}$即$X^TX\boldsymbol{w}=X^T\boldsymbol{y}$，得到$\boldsymbol{w}$的最小二乘估计为$\boldsymbol{w}=(X^TX)^{-1}X^T\boldsymbol{y}$。

例5【方差的最大似然估计】：样本$\boldsymbol{x}_1,\dots,\boldsymbol{x}_N\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)$，求方差$\Sigma$的最大似然估计。写成数学式是：$l=\log|\Sigma|+\frac{1}{N}\sum_{i=1}^N(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T\Sigma^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})$，求$\frac{\partial l}{\partial \Sigma}$的零点。其中$\boldsymbol{x}_i$是$m\times 1$列向量，$\bar{\boldsymbol{x}}=\frac{1}{N}\sum_{i=1}^N\boldsymbol{x}_i$是样本均值，$\Sigma$是$m\times m$对称正定矩阵，$l$是标量，log表示自然对数。

解：首先求微分，使用矩阵乘法、行列式、逆等运算法则，第一项是$d\log|\Sigma|=|\Sigma|^{-1}d|\Sigma|=\text{tr}(\Sigma^{-1}d\Sigma)$，第二项是$\frac{1}{N}\sum_{i=1}^N(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^Td\Sigma^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})=-\frac{1}{N}\sum_{i=1}^N(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})$。再给第二项套上迹做交换：$\text{tr}\left(\frac{1}{N}\sum_{i=1}^N(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\right)$

$$=\frac{1}{N}\sum_{i=1}^N\text{tr}((\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}))$$

$=\frac{1}{N}\sum_{i=1}^N\text{tr}\left(\Sigma^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\right)=\text{tr}(\Sigma^{-1}S\Sigma^{-1}d\Sigma)$，其中先交换迹与求和，然后将$\Sigma^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})$交换到左边，最后再交换迹与求和，并定义$S=\frac{1}{N}\sum_{i=1}^N(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T$为样本方差矩阵。得到$dl=\text{tr}\left(\left(\Sigma^{-1}-\Sigma^{-1}S\Sigma^{-1}\right)d\Sigma\right)$。对照导数与微分的联系，有

例6【多元logistic回归】：$l = -\boldsymbol{y}^T \log \text{softmax}(W\boldsymbol{x})$，求 $\frac{\partial l}{\partial W}$。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；log表示自然对数，$\text{softmax}(\boldsymbol{a}) = \frac{\exp(\boldsymbol{a})}{\boldsymbol{1}^T \exp(\boldsymbol{a})}$，其中 $\exp(\boldsymbol{a})$ 表示逐元素求指数，$\boldsymbol{1}$ 代表全1向量。

解1：首先将softmax函数代入并写成

$l = -\boldsymbol{y}^T \left(\log(\exp(W\boldsymbol{x})) - \boldsymbol{1}\log(\boldsymbol{1}^T \exp(W\boldsymbol{x}))\right) = -\boldsymbol{y}^T W\boldsymbol{x} + \log(\boldsymbol{1}^T \exp(W\boldsymbol{x}))$，这里要注意逐元素log满足等式 $\log(\boldsymbol{u}/c) = \log(\boldsymbol{u}) - \boldsymbol{1}\log(c)$，以及 $\boldsymbol{y}$ 满足 $\boldsymbol{y}^T\boldsymbol{1} = 1$。求微分，使用矩阵乘法、逐元素函数等法则：

$dl = -\boldsymbol{y}^T dW\boldsymbol{x} + \frac{\boldsymbol{1}^T \left(\exp(W\boldsymbol{x}) \odot (dW\boldsymbol{x})\right)}{\boldsymbol{1}^T \exp(W\boldsymbol{x})}$。再套上迹并做交换，注意可化简 $\boldsymbol{1}^T \left(\exp(W\boldsymbol{x}) \odot (dW\boldsymbol{x})\right) = \exp(W\boldsymbol{x})^T dW\boldsymbol{x}$，这是根据等式 $\boldsymbol{1}^T(\boldsymbol{u} \odot \boldsymbol{v}) = \boldsymbol{u}^T\boldsymbol{v}$，故

$dl = \text{tr}\left(-\boldsymbol{y}^T dW\boldsymbol{x} + \frac{\exp(W\boldsymbol{x})^T dW\boldsymbol{x}}{\boldsymbol{1}^T \exp(W\boldsymbol{x})}\right) = \text{tr}(-\boldsymbol{y}^T dW\boldsymbol{x} + \text{softmax}(W\boldsymbol{x})^T dW\boldsymbol{x}) = \text{tr}(\boldsymbol{x}(\text{softmax}(W\boldsymbol{x}) - \boldsymbol{y})^T dW)$

。对照导数与微分的联系，得到 $\frac{\partial l}{\partial W} = (\text{softmax}(W\boldsymbol{x}) - \boldsymbol{y})\boldsymbol{x}^T$。

解2：定义 $\boldsymbol{a} = W\boldsymbol{x}$，则 $l = -\boldsymbol{y}^T \log \text{softmax}(\boldsymbol{a})$，先同上求出 $\frac{\partial l}{\partial \boldsymbol{a}} = \text{softmax}(\boldsymbol{a}) - \boldsymbol{y}$，再利用复合法则：

$dl = \text{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}}^T d\boldsymbol{a}\right) = \text{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}}^T dW\boldsymbol{x}\right) = \text{tr}\left(\boldsymbol{x}\frac{\partial l}{\partial \boldsymbol{a}}^T dW\right)$，得到 $\frac{\partial l}{\partial W} = \frac{\partial l}{\partial \boldsymbol{a}}\boldsymbol{x}^T$。

最后一例留给经典的神经网络。神经网络的求导术是学术史上的重要成果，还有个专门的名字叫做BP算法，我相信如今很多人在初次推导BP算法时也会颇费一番脑筋，事实上使用矩阵求导术来推导并不复杂。为简化起见，我们推导二层神经网络的BP算法。

例7【二层神经网络】：$l = -\boldsymbol{y}^T \log \text{softmax}(W_2\sigma(W_1\boldsymbol{x}))$，求 $\frac{\partial l}{\partial W_1}$ 和 $\frac{\partial l}{\partial W_2}$。其中 $\boldsymbol{y}$ 是除一个元素为1外其它元素为0的的 $m \times 1$ 列向量，W_2 是 $m \times p$ 矩阵，W_1 是 $p \times n$ 矩

σ 是逐元素sigmoid函数 $\sigma(a) = \frac{1}{1+\exp(-a)}$。

解：定义 $\boldsymbol{a}_1 = W_1\boldsymbol{x}$，$\boldsymbol{h}_1 = \sigma(\boldsymbol{a}_1)$，$\boldsymbol{a}_2 = W_2\boldsymbol{h}_1$，则 $l = -\boldsymbol{y}^T \log\operatorname{softmax}(\boldsymbol{a}_2)$。在前例中已求出 $\frac{\partial l}{\partial \boldsymbol{a}_2} = \operatorname{softmax}(\boldsymbol{a}_2) - \boldsymbol{y}$。使用复合法则，

$$dl = \operatorname{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_2}^T d\boldsymbol{a}_2\right) = \operatorname{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_2}^T dW_2\boldsymbol{h}_1\right) + \underbrace{\operatorname{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_2}^T W_2 d\boldsymbol{h}_1\right)}_{dl_2}$$

，使用矩阵乘法交换的迹技巧从第一项得到 $\frac{\partial l}{\partial W_2} = \frac{\partial l}{\partial \boldsymbol{a}_2}\boldsymbol{h}_1^T$，从第二项得到 $\frac{\partial l}{\partial \boldsymbol{h}_1} = W_2^T\frac{\partial l}{\partial \boldsymbol{a}_2}$。接下来对第二项继续使用复合法则来求 $\frac{\partial l}{\partial \boldsymbol{a}_1}$，并利用矩阵乘法和逐元素乘法交换的迹技巧：

$$dl_2 = \operatorname{tr}\left(\frac{\partial l}{\partial \boldsymbol{h}_1}^T d\boldsymbol{h}_1\right) = \operatorname{tr}\left(\frac{\partial l}{\partial \boldsymbol{h}_1}^T (\sigma'(\boldsymbol{a}_1)\odot d\boldsymbol{a}_1)\right) = \operatorname{tr}\left(\left(\frac{\partial l}{\partial \boldsymbol{h}_1}\odot\sigma'(\boldsymbol{a}_1)\right)^T d\boldsymbol{a}_1\right)$$

，得到 $\frac{\partial l}{\partial \boldsymbol{a}_1} = \frac{\partial l}{\partial \boldsymbol{h}_1}\odot\sigma'(\boldsymbol{a}_1)$。为求 $\frac{\partial l}{\partial W_1}$，再用一次复合法则：

$$dl_2 = \operatorname{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_1}^T d\boldsymbol{a}_1\right) = \operatorname{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_1}^T dW_1\boldsymbol{x}\right) = \operatorname{tr}\left(\boldsymbol{x}\frac{\partial l}{\partial \boldsymbol{a}_1}^T dW_1\right)$$

，得到 $\frac{\partial l}{\partial W_1} = \frac{\partial l}{\partial \boldsymbol{a}_1}\boldsymbol{x}^T$。

推广：样本 $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$，

$l = -\sum_{i=1}^N \boldsymbol{y}_i^T \log\operatorname{softmax}(W_2\sigma(W_1\boldsymbol{x}_i + \boldsymbol{b}_1) + \boldsymbol{b}_2)$，其中 $\boldsymbol{b}_1$ 是 $p\times 1$ 列向量，$\boldsymbol{b}_2$ 是 $m\times 1$ 列向量，其余定义同上。

解1：定义 $\boldsymbol{a}_{1,i} = W_1\boldsymbol{x}_i + \boldsymbol{b}_1$，$\boldsymbol{h}_{1,i} = \sigma(\boldsymbol{a}_{1,i})$，$\boldsymbol{a}_{2,i} = W_2\boldsymbol{h}_{1,i} + \boldsymbol{b}_2$，则 $l = -\sum_{i=1}^N \boldsymbol{y}_i^T \log\operatorname{softmax}(\boldsymbol{a}_{2,i})$。先同上可求出 $\frac{\partial l}{\partial \boldsymbol{a}_{2,i}} = \operatorname{softmax}(\boldsymbol{a}_{2,i}) - \boldsymbol{y}_i$。使用复合法则，

$$dl = \operatorname{tr}\left(\sum_{i=1}^N \frac{\partial l}{\partial \boldsymbol{a}_{2,i}}^T d\boldsymbol{a}_{2,i}\right) = \operatorname{tr}\left(\sum_{i=1}^N \frac{\partial l}{\partial \boldsymbol{a}_{2,i}}^T dW_2\boldsymbol{h}_{1,i}\right) + \operatorname{tr}\left(\sum_{i=1}^N \frac{\partial l}{\partial \boldsymbol{a}_{2,i}}^T W_2 d\boldsymbol{h}_{1,i}\right) + \operatorname{tr}\left(\sum_{i=1}^N \frac{\partial l}{\partial \boldsymbol{a}_{2,i}}^T d\boldsymbol{b}_2\right)$$

三项得到 $\frac{\partial l}{\partial \boldsymbol{b}_2} = \sum_{i=1}^{N} \frac{\partial l}{\partial \boldsymbol{a}_{2,i}}$ 。接下来对第二项继续使用复合法则，得到 $\frac{\partial l}{\partial \boldsymbol{a}_{1,i}} = \frac{\partial l}{\partial \boldsymbol{h}_{1,i}} \odot \sigma'(\boldsymbol{a}_{1,i})$ 。为求 $\frac{\partial l}{\partial W_1}, \frac{\partial l}{\partial \boldsymbol{b}_1}$ ，再用一次复合法则：

$$dl_2 = \operatorname{tr}\left(\sum_{i=1}^{N} \frac{\partial l}{\partial \boldsymbol{a}_{1,i}}^T d\boldsymbol{a}_{1,i}\right) = \operatorname{tr}\left(\sum_{i=1}^{N} \frac{\partial l}{\partial \boldsymbol{a}_{1,i}}^T dW_1 \boldsymbol{x}_i\right) + \operatorname{tr}\left(\sum_{i=1}^{N} \frac{\partial l}{\partial \boldsymbol{a}_{1,i}}^T d\boldsymbol{b}_1\right)$$

，得到 $\frac{\partial l}{\partial W_1} = \sum_{i=1}^{N} \frac{\partial l}{\partial \boldsymbol{a}_{1,i}} \boldsymbol{x}_i^T$ ， $\frac{\partial l}{\partial \boldsymbol{b}_1} = \sum_{i=1}^{N} \frac{\partial l}{\partial \boldsymbol{a}_{1,i}}$ 。

解2：可以用矩阵来表示N个样本，以简化形式。定义 $X = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$ ， $A_1 = [\boldsymbol{a}_{1,1}, \cdots, \boldsymbol{a}_{1,N}] = W_1 X + \boldsymbol{b}_1 \boldsymbol{1}^T$ ， $H_1 = [\boldsymbol{h}_{1,1}, \cdots, \boldsymbol{h}_{1,N}] = \sigma(A_1)$ ， $A_2 = [\boldsymbol{a}_{2,1}, \cdots, \boldsymbol{a}_{2,N}] = W_2 H_1 + \boldsymbol{b}_2 \boldsymbol{1}^T$ ，注意这里使用全1向量来扩展维度。先同上求出 $\frac{\partial l}{\partial A_2} = [\operatorname{softmax}(\boldsymbol{a}_{2,1}) - \boldsymbol{y}_1, \cdots, \operatorname{softmax}(\boldsymbol{a}_{2,N}) - \boldsymbol{y}_N]$ 。使用复合法则，

$$dl = \operatorname{tr}\left(\frac{\partial l}{\partial A_2}^T dA_2\right) = \operatorname{tr}\left(\frac{\partial l}{\partial A_2}^T dW_2 H_1\right) + \underbrace{\operatorname{tr}\left(\frac{\partial l}{\partial A_2}^T W_2 dH_1\right)}_{dl_2} + \operatorname{tr}\left(\frac{\partial l}{\partial A_2}^T d\boldsymbol{b}_2 \boldsymbol{1}^T\right)$$

，从第一项得到 $\frac{\partial l}{\partial W_2} = \frac{\partial l}{\partial A_2} H_1^T$ ，从第二项得到 $\frac{\partial l}{\partial H_1} = W_2^T \frac{\partial l}{\partial A_2}$ ，从第三项得到到 $\frac{\partial l}{\partial \boldsymbol{b}_2} = \frac{\partial l}{\partial A_2} \boldsymbol{1}$ 。接下来对第二项继续使用复合法则，得到 $\frac{\partial l}{\partial A_1} = \frac{\partial l}{\partial H_1} \odot \sigma'(A_1)$ 。为求 $\frac{\partial l}{\partial W_1}, \frac{\partial l}{\partial \boldsymbol{b}_1}$ ，再用一次复合法则：

$$dl_2 = \operatorname{tr}\left(\frac{\partial l}{\partial A_1}^T dA_1\right) = \operatorname{tr}\left(\frac{\partial l}{\partial A_1}^T dW_1 X\right) + \operatorname{tr}\left(\frac{\partial l}{\partial A_1}^T d\boldsymbol{b}_1 \boldsymbol{1}^T\right)$$

，得到 $\frac{\partial l}{\partial W_1} = \frac{\partial l}{\partial A_1} X^T$ ， $\frac{\partial l}{\partial \boldsymbol{b}_1} = \frac{\partial l}{\partial A_1} \boldsymbol{1}$ 。

下篇见zhuanlan.zhihu.com/p/24...。

编辑于 2020-03-06

文章被以下专栏收录

深度学习于图网络
不定时更新图网络学习笔记

数学

推荐阅读

矩阵求导术（下）

本文承接上篇 https://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...

长躯鬼侠

矩阵求导浅析（一）

本文主要关注标量函数对矩阵的导，并提供一种简明直观易操作矩阵求导方法。 推荐矩阵求导的专栏文章： 矩阵求导术（上）阵求导术（下） 机器学习中的向量求导 1.内积 向量 ...

倚楼

461 条评论 切换为时间排序

写下你的评论...

resurrectcore 2017-01-07
网上有个pdf叫做 The Matrix Cookbook.
74

知乎用户 回复 resurrectcore 2017-01-07
我看过，相信写这篇文章的人也看过。感觉这里的方法比硬背公式简单，而且对于搞机

matrix cookbook是当字典用的。

2

查看全部 9 条回复

王赟 Maigo 2017-01-07

赞啊！终于找到系统的方法了！

59

方不觉 2017-01-07

nbnb，一直对矩阵求导感觉无从下手，这几条法则比背公式好记多了！

30

知乎用户 2017-01-09

参考文献大概是：Matrix Calculus:Derivation and Simple Application HU Pili

15

知乎用户 2019-07-27

matrixcalculus.org 曾经打算记过 后来不怎么用就放弃了 要用的时候直接查公式（逃

10

黑与白 回复 知乎用户 2019-10-15

大佬，请问这个网站是什么网站呀

赞

DreamYun 回复 知乎用户 2020-04-28

这个牛逼了，不得了！！！

赞

展开其他 3 条回复

知乎用户 2017-01-23

刚开始搞CV时候，发现典型的CVPR paper总有这么几个矩阵求导公式，与那些只知道刷baseline的文章相比，顿时逼格高大上啊！唉，作者你要是早几年写这个专栏一定能解救一批苦逼的PhD啊！

6

算

6

程许 回复 maja 2020-08-08

我知道你的意思是 ，假如a,b,x都是n1维矩阵，那么从a^T*x=b^T*x并不能必然推出a=b，因为存在恰好是位于零空间的可能。我想这里面的gap可能可以类似于多项式插值定理来弥补：如果f(x)=g(x)在n+2个不同点上都成立，而且f,g的次数都不超过n次，那么f和g恒等。不过作者没有在本文证明或举出这种类型的定理，所以确实可以认为文章的推导只是intuitive，并不严格。

赞

长躯鬼侠 (作者) 回复 程许 2020-08-08

给定a^T x= b^T x对任意x成立，是可以推出a=b的。

3

知乎用户 2017-01-07

写的真好，醍醐灌顶

6

lang的飞起 2017-11-07

刚开始学，学习了。那个线性回归那里，怎么一步就推出2倍了，没看明白

5

长躯鬼侠 (作者) 回复 lang的飞起 2017-11-07

两个向量的内积，a^Tb=b^Ta，加起来就是2倍了

2

虚有 2020-09-30

两年前看:这是什么玩意儿?

两年后：卧槽，这不就是矩阵理论里没讲的章节吗？ np。

给大佬点赞，讲的太好太详细了，看了这么多本矩阵理论的教材，不是缺这个就是缺那个。还是您讲的好。

4

知乎用户 2017-02-27

相思作坊半世離殤 回复 知乎用户 2017-03-20

感谢分享资料。看了这篇pdf和楼主写的文章，明白了许多

1

bb just 回复 相思作坊半世離殤 2019-06-04

链接打不开了，请问你还有这个质量么。

赞

查看全部 7 条回复

Zzzzzz 2017-01-07

行列式微分的那个可以把行列式放进tr里变成伴随矩阵。行列式对其中元素的偏导显然就是其代数余子式，应该不用行列式不为0。这公式叫jacobi's formula

4

长躯鬼侠 (作者) 回复 Zzzzzz 2017-01-07

嗯，你说得对。

赞

知乎用户 回复 Zzzzzz 2017-01-10

他的意思是当行列式不为零时，可以把伴随矩阵写成行列式乘以逆吧

赞

maja 2017-01-08

没有回应，可能你没看懂我的意思：这样吧我们从梯度下降流来考虑：

1）全微分的出发点是没有问题的，但用迹是没有解决问题的，问题的本质是如何定义符号 “df / dX” ，f必然是n^2元函数，根据标量对n元函数有 df = sum (partial f / X(i,j) * dX(i,j)) 如果定义了 dX 为矩阵形式就决定了 df / dX 。如果用矩阵内积运算 则 就确定了 df / dX 的矩阵形式 衡量的就是df 对于dX变化率.

回到基础问题一 f = a^{T} * X * b 现在我们要根据 df = df/dX • dX 来求解. 所以要确定df的形式 矩阵乘法始终是一个记号！ 和行列式通过伴随矩阵建立联系. 通俗的运算就是加 乘 = 定义记号-累乘-S(i, a)=\sum_{i=1}^{n} a_i 则 f=S(i,a)S(k,x_ik * b_k) 然后两边对X诸元素全微分，df=S(<i,k>, dX) {S(p, a) * S(q, b)} . 提取公因式，df = a^T * b ? dX • 1

这个是我们可以用矩阵描述梯度下降流的朴素原理，现在考虑第二个问题：

这就是为什呢题主用迹会出错，因为就不对.

3

知乎用户 回复 maja 2020-09-05

楼主这一套方法，来源应该就是张贤达的《矩阵分析与应用》里面的，方法从计算层面上应该是没问题的，但是一个潜在的问题就是，这一套方法和用到jacobi或者符合函数的时候，总是会差一个转置，使用df / dx 行向量的约定，jacobi和符合求导能够更简洁，也和流形上的微积分的符号一致。而像这里的全部都用列向量的情况，和流形上微积分的公式就不一致，差转置

2

平常心 回复 知乎用户 2020-12-19

对

赞

展开其他 2 条回复

紫杉 2017-01-07

期待下集。。我都忘了很多这部分内容了

3

阳光陈靖文 回复 紫杉 01-22

去主页把链接贴过来了长躯鬼侠：矩阵求导术（下）不用谢

2

覃含章 2017-01-07

写的很清楚，很实用!

3

白夜行 2020-07-04

2

柴鹰 2017-12-28

虽然写了很多，但是看得不是很懂。看来真的是要靠背公式活下去了。

2

黑与白 回复 柴鹰 2019-10-15

我能感觉到答主写的很有用，但是我也确实没有理解，而现在又着急使用，请问您是背哪些公式？可以分享一下吗

赞

老大 回复 黑与白 03-29

我猜你俩是没看懂全微分公式，导致后面的推导没法convince你自己，其实楼主确实全微分公式那里省略了一点，不过可以自己写写看

赞

知乎用户 2017-01-21

之前看吴恩达的cs229就对这个矩阵求导和迹的联系十分懵，多谢作者

2

甄景贤 2017-01-08

这个超有用，功德无量 :)

2

CoolMan 2020-05-18

牛逼牛逼，有些比张贤达的书写的还清楚

1

1 2 3 4 ... 15 下一页