矩阵求导的理解（重要！）

codingli...
公众号：L的算法成长之路

《矩阵求导术》重点笔记

首先是标量对矩阵的求导

一元微积分（标量对标量）中的导数与微分的关系： $df = f'(x)dx$

多元微积分（标量对向量）中的梯度与微分的关系： $df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i = \frac{\partial f^T}{\partial x} dx$

（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 $\frac{\partial f}{\partial x}$ (nx1)与微分向量 dx (nx1)的内积）

矩阵导数与微分建立联系： $df = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial f}{\partial X_{ij}} dX_{ij} = tr(\frac{\partial f^T}{\partial X} dX)$

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， $tr(A^T B) = \sum_{i,j} A_{ij} B_{ij}$ ，即 $tr(A^T B)$ 是矩阵A,B的内积。

这里表示，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。

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： $dtr(X) = tr(dX)$

2.逆： $dX^{-1} = -X^{-1}dXX^{-1}$

3.行列式： $d|X| = tr(X^* dX)$ ， X^* 表示X的伴随矩阵

如果X可逆，上式可写成 $d|X| = |X|tr(X^{-1}dX)$

4.逐元素乘法： $d(X \odot Y) = dX \odot Y + X \odot dY$ ，$\odot$ 代表尺寸相同的矩阵逐元素相乘

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。

一些迹技巧：

1.标量套上迹： $a = tr(a)$

2.转置： $tr(A^T) = tr(A)$

3.线性： $tr(A \pm B) = tr(A) \pm tr(B)$

4.矩阵乘法交换： $tr(AB) = tr(BA)$ ，其中 A 与 B^T 尺寸相同，两侧都等于 $\sum_{i,j} A_{ij}B_{ij}$

5.矩阵乘法/逐元素乘法交换： $tr(A^T(B \odot C)) = tr((A \odot B)^T C)$ ，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，即能得到导数。**

关于复合：

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 $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y}\frac{\partial y}{\partial x}$ ，但这里我们不能沿用链式法则，因为矩阵对矩阵的导数 $\frac{\partial Y}{\partial X}$ 截止目前仍是未定义的。我们直接从微分入手建立复合法则：先写出 $df = tr(\frac{\partial f}{\partial Y}^T dY)$ ，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 $\frac{\partial f}{\partial X}$ 。

来看几个例子：

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，这里的 $\boldsymbol{a}, \boldsymbol{b}$ 是常量，$d\boldsymbol{a} = \boldsymbol{0}, d\boldsymbol{b} = \boldsymbol{0}$，得到：$df = \boldsymbol{a}^T dX \boldsymbol{b}$，再套上迹并做矩阵乘法交换：$df = \text{tr}(\boldsymbol{a}^T dX \boldsymbol{b}) = \text{tr}(\boldsymbol{b}\boldsymbol{a}^T dX)$，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 $\frac{\partial f}{\partial X} = (\boldsymbol{b}\boldsymbol{a}^T)^T = \boldsymbol{a}\boldsymbol{b}^T$。

注意：这里不能用 $\frac{\partial f}{\partial X} = \boldsymbol{a}^T \frac{\partial X}{\partial X} \boldsymbol{b} = ?$，导数与乘常数矩阵的交换是不合法则的运算（而微分是合法的）。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 $df = \text{tr}(\boldsymbol{a}^T(\exp(X\boldsymbol{b}) \odot (dX\boldsymbol{b}))) = \text{tr}((\boldsymbol{a} \odot \exp(X\boldsymbol{b}))^T dX\boldsymbol{b}) = \text{tr}(\boldsymbol{b}(\boldsymbol{a} \odot \exp(X\boldsymbol{b}))^T dX)$，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 $\frac{\partial f}{\partial X} = (\boldsymbol{b}(\boldsymbol{a} \odot \exp(X\boldsymbol{b}))^T)^T = (\boldsymbol{a} \odot \exp(X\boldsymbol{b}))\boldsymbol{b}^T$。

例3：$f = \text{tr}(Y^T M Y), Y = \sigma(WX)$，求 $\frac{\partial f}{\partial X}$。其中 W 是 $l \times m$ 列向量，X 是 $m \times n$ 矩阵，Y 是 $l \times n$ 矩阵，M 是 $l \times l$ 对称矩阵，σ 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：$df = \text{tr}((dY)^T M Y) + \text{tr}(Y^T M dY) = 2\text{tr}(Y^T M dY)$，对照导数与微分的联系，得到 $\frac{\partial f}{\partial Y} = 2MY$。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：$df = \text{tr}\left(\frac{\partial f}{\partial Y}^T (\sigma'(WX) \odot (WdX))\right) = \text{tr}\left(\left(\frac{\partial f}{\partial Y} \odot \sigma'(WX)\right)^T W dX\right)$，对照导数与微分的联系，得到 $\frac{\partial f}{\partial X} = W^T\left(\frac{\partial f}{\partial Y} \odot \sigma'(WX)\right) = W^T((2M\sigma(WX)) \odot \sigma'(WX))$。

例4【线性回归】：$l = \|X\boldsymbol{w} - \boldsymbol{y}\|^2$，求 $\boldsymbol{w}$ 的最小二乘估计，即求 $\frac{\partial l}{\partial \boldsymbol{w}}$ 的零点。其中 $\boldsymbol{y}$ 是 $m \times 1$ 列向量，X 是 $m \times n$ 矩阵，$\boldsymbol{w}$ 是 $n \times 1$ 列向量，l 是标量。

解：严格来说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：$dl = (Xd\boldsymbol{w})^T(X\boldsymbol{w} - \boldsymbol{y}) + (X\boldsymbol{w} - \boldsymbol{y})^T(Xd\boldsymbol{w}) = 2(X\boldsymbol{w} - \boldsymbol{y})^TXd\boldsymbol{w}$。对照导数与微分的联系 $dl = \frac{\partial l}{\partial \boldsymbol{w}}^T d\boldsymbol{w}$，得到 $\frac{\partial l}{\partial \boldsymbol{w}} = (2(X\boldsymbol{w} - \boldsymbol{y})^TX)^T = 2X^T(X\boldsymbol{w} - \boldsymbol{y})$。$\frac{\partial l}{\partial \boldsymbol{w}}$ 的零点即 $\boldsymbol{w}$ 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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim N(\boldsymbol{\mu}, \Sigma)$，求方差 Σ 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，Σ 是 $m \times m$ 对称正定矩阵，l 是标量。

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 $\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T d\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = -\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$。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 $dl = \text{tr}\left((\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1})d\Sigma\right)$。对照导数与微分的联系，有 $\frac{\partial l}{\partial \Sigma} = (\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1})^T$，其零点即 Σ 的最大似然估计为 $\Sigma = S$。

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；$\text{softmax}(\boldsymbol{a}) = \frac{\exp(\boldsymbol{a})}{\boldsymbol{1}^T\exp(\boldsymbol{a})}$，其中 $\exp(\boldsymbol{a})$ 表示逐元素求指数，$\boldsymbol{1}$ 代表全1向量。

解：首先将softmax函数代入并写成 $l = -\boldsymbol{y}^T\left(\log(\exp(W\boldsymbol{x})) - \boldsymbol{1}\log(\boldsymbol{1}^T\exp(W\boldsymbol{x}))\right) = -\boldsymbol{y}^TW\boldsymbol{x} + \log(\boldsymbol{1}^T\exp(W\boldsymbol{x}))$，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：$dl = -\boldsymbol{y}^TdW\boldsymbol{x} + \frac{\boldsymbol{1}^T(\exp(W\boldsymbol{x}) \odot (dW\boldsymbol{x}))}{\boldsymbol{1}^T\exp(W\boldsymbol{x})}$。再套上迹并做交换，注意可化简 $\boldsymbol{1}^T(\exp(W\boldsymbol{x}) \odot (dW\boldsymbol{x})) = \exp(W\boldsymbol{x})^TdW\boldsymbol{x}$，这是根据等式 $\boldsymbol{1}^T(\boldsymbol{u} \odot \boldsymbol{v}) = \boldsymbol{u}^T\boldsymbol{v}$，故

$dl = \mathrm{tr}\left(-\boldsymbol{y}^T dW\boldsymbol{x} + \frac{\exp(W\boldsymbol{x})^T dW\boldsymbol{x}}{\boldsymbol{1}^T \exp(W\boldsymbol{x})}\right) = \mathrm{tr}(\boldsymbol{x}(\mathrm{softmax}(W\boldsymbol{x}) - \boldsymbol{y})^T dW)$ 。对照导数与微分
系，得到 $\frac{\partial l}{\partial W} = (\mathrm{softmax}(W\boldsymbol{x}) - \boldsymbol{y})\boldsymbol{x}^T$ 。

另解：定义 $\boldsymbol{a} = W\boldsymbol{x}$ ，则 $l = -\boldsymbol{y}^T \log \mathrm{softmax}(\boldsymbol{a})$ ，先如上求出 $\frac{\partial l}{\partial \boldsymbol{a}} = \mathrm{softmax}(\boldsymbol{a}) - \boldsymbol{y}$ ，再利用复合法则：$dl = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}}^T d\boldsymbol{a}\right) = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}}^T dW\boldsymbol{x}\right) = \mathrm{tr}\left(\boldsymbol{x}\frac{\partial l}{\partial \boldsymbol{a}}^T dW\right)$ ，得到 $\frac{\partial l}{\partial W} = \frac{\partial l}{\partial \boldsymbol{a}}\boldsymbol{x}^T$ 。

然后是矩阵对矩阵的求导

先定义向量 f (p×1)对向量 x (m×1)的导数：

有 $df = \frac{\partial f}{\partial x}^T dx$ ；

再定义矩阵的（按列优化）向量化

$vec(X) = [X_{11}, \ldots, X_{m1}, X_12, \ldots X_{m2}, \ldots, X_{1n}, \ldots, X_{mn}]^T$ (mn×1)，并定义矩阵F对矩阵X的导数 $\frac{\partial F}{\partial X} = \frac{\partial vec(F)}{\partial vec(X)}$ (mn×pq)。导数与微分有联系：

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

向量化的技巧：

1.线性： $vec(A+B) = vec(A) + vec(B)$

2.矩阵乘法： $vec(AXB) = (B^T \otimes A)vec(X)$ ，其中 $\otimes$ 代表Kronecker积，A(m×n)与B(p×
Kronecker积是 $A \otimes B = [A_{ij}B]$ (mp×nq)。

3.转置： $\text{vec}(A^T) = K_{mn}\text{vec}(A)$，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，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 $\frac{\partial F}{\partial X} = \frac{\partial Y}{\partial X}\frac{\partial F}{\partial Y}$ 。

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 $F = D^T B^T XAC$ 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 $Y = B^T XA$，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，
   $\text{vec}(AXB^T) = K_{pm}\text{vec}(BX^T A^T) = K_{pm}(A \otimes B)\text{vec}(X^T) = K_{pm}(A \otimes B)K_{nq}\text{vec}(X)$ 。

例子：

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 $\mathrm{vec}(d\nabla_X f) = -K_{nn}\mathrm{vec}(X^{-1}dXX^{-1}) = -K_{nn}(X^{-1T} \otimes X^{-1})\mathrm{vec}(dX)$ ，对照导数与微分的联系，得到 $\nabla^2_X f = -K_{nn}(X^{-1T} \otimes X^{-1})$ ，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：
$\mathrm{vec}(dF) = (I_p \otimes A)\mathrm{vec}(\exp(XB) \odot (dXB))$ ，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】： $l = -y\boldsymbol{x}^T\boldsymbol{w} + \log(1 + \exp(\boldsymbol{x}^T\boldsymbol{w}))$ ，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 $\nabla^2_{\boldsymbol{w}} l = \boldsymbol{x}\sigma'(\boldsymbol{x}^T\boldsymbol{w})\boldsymbol{x}^T$ 。

推广：样本 $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ ， $l = \sum_{i=1}^{N}\left(-y_i\boldsymbol{x}_i^T\boldsymbol{w} + \log(1 + \exp(\boldsymbol{x}_i^T\boldsymbol{w}))\right)$ ，求 $\nabla_{\boldsymbol{w}} l$ 和 $\nabla^2_{\boldsymbol{w}} l$ 。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 $l = -\boldsymbol{y}^T X\boldsymbol{w} + \mathbf{1}^T \log(\mathbf{1} + \exp(X\boldsymbol{w}))$，进而可以求得 $\nabla_{\boldsymbol{w}} l = X^T(\sigma(X\boldsymbol{w}) - \boldsymbol{y})$， $\nabla^2_w l = X^T \text{diag}(\sigma'(X\boldsymbol{w}))X$ 。

例5【多元logistic回归】： $l = -\boldsymbol{y}^T \log \text{softmax}(W\boldsymbol{x}) = -\boldsymbol{y}^T W\boldsymbol{x} + \log(\mathbf{1}^T \exp(W\boldsymbol{x}))$，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 $\boldsymbol{a} = W\boldsymbol{x}$， $d\text{softmax}(\boldsymbol{a}) = \frac{\exp(\boldsymbol{a}) \odot d\boldsymbol{a}}{\mathbf{1}^T \exp(\boldsymbol{a})} - \frac{\exp(\boldsymbol{a})(\mathbf{1}^T(\exp(\boldsymbol{a}) \odot d\boldsymbol{a}))}{(\mathbf{1}^T \exp(\boldsymbol{a}))^2}$，这里需要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 $\mathbf{1}^T(\exp(\boldsymbol{a}) \odot d\boldsymbol{a}) = \exp(\boldsymbol{a})^T d\boldsymbol{a}$，故有 $d\text{softmax}(\boldsymbol{a}) = D\text{softmax}(\boldsymbol{a})d\boldsymbol{a}$，其中 $D\text{softmax}(\boldsymbol{a}) = \frac{\text{diag}(\exp(\boldsymbol{a}))}{\mathbf{1}^T \exp(\boldsymbol{a})} - \frac{\exp(\boldsymbol{a}) \exp(\boldsymbol{a})^T}{(\mathbf{1}^T \exp(\boldsymbol{a}))^2}$，代入有 $d\nabla_W l = D\text{softmax}(\boldsymbol{a})d\boldsymbol{a}\boldsymbol{x}^T = D\text{softmax}(W\boldsymbol{x})dW\boldsymbol{x}\boldsymbol{x}^T$，做向量化并使用矩阵乘法的技巧，得到 $\nabla^2_W l = (\boldsymbol{x}\boldsymbol{x}^T) \otimes D\text{softmax}(W\boldsymbol{x})$ 。

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 $df = \text{tr}(\nabla^T_X f dX)$，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 $df = \nabla^T_{\boldsymbol{x}} f d\boldsymbol{x}$；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：

Matrix calculus - Wikipedia

通过一个例子快速上手矩阵求导 - NoGeek - CSDN博客

矩阵求导术（上）

矩阵求导术（下）

发布于 2018-11-19

文章被以下专栏收录

算法修炼之路
机器学习和深度学习的相关算法等

推荐阅读

矩阵求导与矩阵微分

矩阵求导与矩阵微分符号定义 使用大写的粗体字母表示矩阵 \mathbf{A}、\mathbf{F} 使用小写的粗体字母表示向量 \mathbf{x}、\mathbf{f}，这里默认为列向量 使用小写的正体字母表示标量 x、f...

说谎的傻子　发表于科学与技术

向量恒等式

这次的内容相当平凡。最近学磁流体力学，发现对向量恒等式还是不熟悉，以前只是记一下 \mathbf{A} \times(\mathbf{B} \times \mathbf{C})=(\mathbf{A} \cdot \mathbf{C}) \mathbf{B}-(\math...

魂魄妖妖梦　发表于A Tri...

微分的四种理解

一 微分是无穷小？物理 分看做是一个很小的量， 时总是很方便的，但是 严谨的感觉。实际上，它 谨，第二次数学危机就是 的。 严谨性与明晰性是

像牛一样的猫

5 条评论

切换为时间排序

写下你的评论...

知乎用户　04-19
这句话是啥意思？因为矩阵对矩阵的导数截至目前是未定义的。为啥？
赞

codingling (作者) 回复 知乎用户　04-20
因为不知道y对x的导数，没法用链式法则
赞

阳光陈靖文　01-22

刚看完长躯鬼侠的矩阵求导术，就读到你的笔记哈哈哈，总结的不错

赞

雁过无声 2020-06-06

666

赞

han 2019-08-27

清晰，易懂

赞