矩阵求导总结（一）

2020-01-12

本文由于篇幅原因拆成两篇，下一篇见这里。

标量对向量或矩阵求导

基本方法

y是一个标量，$\mathbf{x}$是向量，A是矩阵。标量对向量或矩阵求导，即对逐个元素的求导

- $\frac{\partial y}{\partial \mathbf{x}}$结果是一个与$\mathbf{x}$维度相同的向量
- $\frac{\partial y}{\partial A}$结果是一个与$A$维度相同的矩阵

实际应用中，一个类似这样的公式$l = (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})$，求$\frac{\partial l}{\partial \beta}$，两种思路

- 将矩阵写开，变成标量形式，加各种$\sum_{i=1}^{n}$，用l对每个β_i求导后，按照求导后应该有的维度，把结果拼起来
    - 当l形式比较简单时适用，复杂形式请用微分法
- **微分法**：右边套一个迹，等式两端同时取微分，目标是写成这种形式$\mathrm{d}l = \mathrm{tr}(\mathbf{b}^T\mathrm{d}\boldsymbol{\beta})$，则可得$\frac{\partial l}{\partial \beta} = \mathbf{b}$
    - 如果是标量对矩阵求导也一样，写成这种形式$\mathrm{d}l = \mathrm{tr}(A^T\mathrm{d}X)$，则$\frac{\partial l}{\partial X} = A$
    - 套迹取微分后的推导，主要用到**微分运算法则**和**迹的性质**，二者都会列在下面。其他说明：
        - 右边可以套一个迹，是因为等式左右两边都是标量；取迹的目的是方便右侧变形，而迹保持不变，举例如下
            - 比如最后推出这种形式：$\mathrm{d}l = \mathrm{tr}(\mathbf{b}\mathrm{d}\boldsymbol{\beta}^T)$，则$\frac{\partial l}{\partial \beta} = \mathbf{b}$。这是用到了迹内转置、交换位置的性质
            - 经常等式右侧的$\mathrm{d}\boldsymbol{\beta}$都不在最后，要用迹内交换位置的性质，交换位置的原则是保持矩阵相乘有意义，这也是减少计算错误的有效手段。经常是$\mathrm{d}\boldsymbol{\beta}$后面的一整块直接移到最前面。
        - 微分$\mathrm{d}X$与X维度相同，这个性质可以帮助判断是否保持了矩阵相乘有意义

微分运算法则

- 常数微分：$\mathrm{d}X=\mathrm{O}$，如果X由常数组成，O与X维度相同
- 微分加减法：$\mathrm{d}(X+Y)=\mathrm{d}X+\mathrm{d}Y$， $\mathrm{d}(X-Y)=\mathrm{d}X-\mathrm{d}Y$
- 微分乘法：$\mathrm{d}(XY)=(\mathrm{d}X)Y+X(\mathrm{d}Y)$
- 微分转置：$\mathrm{d}(X^T)=(\mathrm{d}X)^T$
- 微分的迹：$\mathrm{d}\,\mathrm{tr}(X)=\mathrm{tr}(\mathrm{d}X)$
- 微分哈达马乘积：$\mathrm{d}(X\odot Y)=X\odot\mathrm{d}Y+\mathrm{d}X\odot Y$
- 逐元素函数微分：$\mathrm{d}\sigma(X)=\sigma'(X)\odot\mathrm{d}X$，其中$\sigma$是对$X$中每个元素进行函数变换，结果与$X$维度相同；求导结果的矩阵每个元素为$\sigma'(x_{ij})\mathrm{d}x_{ij}$
- 逆矩阵微分：$\mathrm{d}X^{-1}=-X^{-1}\mathrm{d}XX^{-1}$。此式可通过$XX^{-1}=\mathrm{I}$左右两侧求微分推得。
- 行列式微分：$\mathrm{d}|X|=|X|\,\mathrm{tr}(X^{-1}\mathrm{d}X)$，这里默认$X$可逆，因为如果不可逆$|X|$就是0了。更一般的表示是$\mathrm{d}|X|=\mathrm{tr}(X^{\#}\mathrm{d}X)$，其中$X^{\#}$是$X$的伴随矩阵。
  - 直观理解：$|X|=\sum_{j=1}^n x_{ij}X_{ji}^{\#}$，这对任意$i$都成立，所以$|X|$对$x_{ij}$的导数就应该是$X_{ji}^{\#}$，因此$\frac{\partial|X|}{\partial X}=X^{\#\mathrm{T}}$，所以微分形式就是$\mathrm{d}|X|=\mathrm{tr}(X^{\#}\mathrm{d}X)$
  - 注：如果这样写，$n|X|=\sum_{i=1}^n\sum_{j=1}^n x_{ij}X_{ji}^{\#}$，那岂不是$\frac{\partial|X|}{\partial X}=\frac{1}{n}X^{\#\mathrm{T}}$？这个式子不对，因为$X_{mn}^{\#}$里也会含有一些$x_{ij}$的项，所以没有那么简单（而上面$X_{ji}^{\#}$中确实不含$x_{ij}$的项）。

迹的性质

- 标量的迹等于自身：$\mathrm{tr}(a)=a$
- 转置：$\mathrm{tr}(A^T)=\mathrm{tr}(A)$
- 线性：$\mathrm{tr}(A\pm B)=\mathrm{tr}(A)\pm\mathrm{tr}(B)$
- 交换：$\mathrm{tr}(A^TB)=\mathrm{tr}(B^TA)$，其中$A$与$B$维度相同，迹结果等于$\sum_{i,j}A_{ij}B_{ij}$
  - 类似地有：$\mathrm{tr}\big(A^T(B\odot C)\big)=\mathrm{tr}\big((A\odot B)^TC\big)$，其中$A,B,C$维度相同，迹结果为$\sum_{i,j}A_{ij}B_{ij}C_{ij}$

微分法的背后原理

为什么标量对向量求导，写成$\mathrm{d}l=\mathrm{tr}(\mathbf{b}^T\mathrm{d}\boldsymbol{\beta})$，则$\frac{\partial l}{\partial\beta}=\mathbf{b}$；标量对矩阵求导，写成$\mathrm{d}l=\mathrm{tr}(A^T\mathrm{d}X)$，则$\frac{\partial l}{\partial X}=A$？

- 标量对向量求导：等式右侧其实是$\sum_i b_i\mathrm{d}\beta_i$，那么$\frac{\partial l}{\partial\beta_i}=b_i$，自然可得$\frac{\partial l}{\partial\beta}=\mathbf{b}$
- 标量对矩阵求导同理，等式右侧是$\sum_{ij}a_{ij}\mathrm{d}X_{ij}$，那么$\frac{\partial l}{\partial X_{ij}}=a_{ij}$，自然可得$\frac{\partial l}{\partial X}=A$

这借鉴了多元情形下的全微分公式，全微分是梯度向量与微分向量的内积

$$\mathrm{d}f=\sum_i\frac{\partial f}{\partial x_i}\mathrm{d}x_i=\left[\frac{\partial f}{\partial\mathbf{x}}\right]^T\mathrm{d}\mathbf{x}$$

了解这个原理后，我们可以发现写成其他形式也是可以的，比如**内积**

- 标量对向量求导，写成$\mathrm{d}l=\langle\mathbf{b},\mathrm{d}\boldsymbol{\beta}\rangle$，则$\frac{\partial l}{\partial\beta}=\mathbf{b}$
- 标量对矩阵求导，写成$\mathrm{d}l=\langle A,\mathrm{d}X\rangle$，则$\frac{\partial l}{\partial X}=A$
    - 注：矩阵的内积是，对应位置相乘，再将所有数相加

内积形式的应用可以参见下一节：**哈达马乘积的处理。**

哈达马乘积的处理

遇到$\odot\mathrm{d}\mathbf{y}$这种情况，还是要努力转化成我们熟知的形式。这里举一个例子，提供三种方法

题目：$l=\mathbf{x}^T\exp(\mathbf{y})$，求$\frac{\partial l}{\partial\mathbf{y}}$。

1、内积方法

$$\begin{aligned}\mathrm{d}l&=\mathbf{x}^T[\exp(\mathbf{y})\odot\mathrm{d}\mathbf{y}]\\&=\langle\mathbf{x},\exp(\mathbf{y})\odot\mathrm{d}\mathbf{y}\rangle\\&=\langle\mathbf{x}\odot\exp(\mathbf{y}),\mathrm{d}\mathbf{y}\rangle\end{aligned}$$

所以$\frac{\partial l}{\partial\mathbf{y}}=\mathbf{x}\odot\exp(\mathbf{y})$。

上面最后一个等式是一个性质，也很好理解，只要写成$\sum_i x_i\exp(y_i)y_i$即可；当三者都是矩阵时，这条性质也成立。

2、迹的性质

对哈达马乘积，迹也有和上面内积类似的性质：A,B,C同维度时，$\operatorname{tr}((A\odot B)^TC)=\operatorname{tr}(A^T(B\odot C))$。

如果用这条性质来做的话，就可以直接写出

$$\mathrm{d}l=\operatorname{tr}\left([\mathbf{x}\odot\exp(\mathbf{y})]^T\mathrm{d}\mathbf{y}\right)$$

3、矩阵相乘

当出现的是向量的哈达马乘积时，还有第三种做法。令$Z=\operatorname{diag}(\mathbf{y})$，则

$$\mathrm{d}l=\mathbf{x}^T[\exp(\mathbf{y})\odot\mathrm{d}\mathbf{y}]=\mathbf{x}^TZ\mathrm{d}\mathbf{y}$$

这就是我们熟知的形式了。

例题

1、标量对向量求导。已知$l = \mathbf{x}^T A\mathbf{x}$，求$\frac{\partial l}{\partial \mathbf{x}}$。

- 解法1：右侧写成标量形式

  $$l = \sum_{ij} x_i a_{ij} x_j$$

  对向量中元素逐个求导如下

  $$\begin{aligned}\frac{\partial l}{\partial x_k} &= \sum_{j\neq k} a_{kj}x_j + \sum_{i\neq k} x_i a_{ik} + 2a_{kk}x_k \\ &= \sum_j a_{kj}x_j + \sum_i x_i a_{ik} \\ &= A_{k,:}\mathbf{x} + \mathbf{x}^T A_{:,k} \\ &= A_{k,:}\mathbf{x} + A^T_{k,:}\mathbf{x}\end{aligned}$$

  拼合可得

  $$\frac{\partial l}{\partial \mathbf{x}} = (A + A^T)\mathbf{x}$$

- 解法2：微分法

  $$\begin{aligned}\mathrm{d}l &= \mathrm{d}\left[\mathrm{tr}(\mathbf{x}^T A\mathbf{x})\right] = \mathrm{tr}\left[\mathrm{d}(\mathbf{x}^T A\mathbf{x})\right] \\ &= \mathrm{tr}\left[\mathrm{d}(\mathbf{x}^T A)\mathbf{x} + \mathbf{x}^T A\mathrm{d}\mathbf{x}\right] \\ &= \mathrm{tr}\left[\mathrm{d}\mathbf{x}^T A\mathbf{x} + \mathbf{x}^T A\mathrm{d}\mathbf{x}\right] \\ &= \mathrm{tr}\left[\mathbf{x}^T A^T\mathrm{d}\mathbf{x} + \mathbf{x}^T A\mathrm{d}\mathbf{x}\right] = \mathrm{tr}\left[\mathbf{x}^T(A^T + A)\mathrm{d}\mathbf{x}\right]\end{aligned}$$

  因此$\frac{\partial l}{\partial \mathbf{x}} = (A + A^T)\mathbf{x}$。

2、标量对矩阵求导。已知$l = \mathbf{a}^T X\mathbf{b}$，求$\frac{\partial l}{\partial X}$。

- 解法1：右侧写成标量形式

  $$l = \sum_{ij} a_i x_{ij} b_j$$

  对向量中元素逐个求导可得$\frac{\partial l}{\partial X_{ij}} = a_i b_j$。所以$\frac{\partial l}{\partial X} = \mathbf{a}\mathbf{b}^T$

- 解法2：微分法

$$
\begin{aligned}
\mathrm{d}l &= \mathrm{d}\left[\operatorname{tr}(\mathbf{a}^T X\mathbf{b})\right] = \operatorname{tr}\left[\mathrm{d}(\mathbf{a}^T X\mathbf{b})\right] \\
&= \operatorname{tr}\left[\mathbf{a}^T \mathrm{d}(X\mathbf{b})\right] = \operatorname{tr}\left[\mathbf{a}^T \mathrm{d}X\mathbf{b}\right] \\
&= \operatorname{tr}\left[\mathbf{b}\mathbf{a}^T \mathrm{d}X\right]
\end{aligned}
$$

因此$\frac{\partial l}{\partial \mathbf{x}} = \mathbf{a}\mathbf{b}^T$。

3、多元正态分布Σ的极大似然估计，需要计算对数似然对Σ的导数

$$
l = \log|\Sigma| + \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})^T \Sigma^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})
$$

使用微分法

$$
\begin{aligned}
\mathrm{d}l\ &= \frac{1}{|\Sigma|}\mathrm{d}|\Sigma| + \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})^T \mathrm{d}(\Sigma^{-1})(\mathbf{x}_i - \overline{\mathbf{x}}) \\
&= \operatorname{tr}(\Sigma^{-1}\mathrm{d}\Sigma) - \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})^T \Sigma^{-1}\mathrm{d}\Sigma\Sigma^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) \\
&= \operatorname{tr}(\Sigma^{-1}\mathrm{d}\Sigma) - \frac{1}{N}\sum_{i=1}^{N}\Sigma^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \Sigma^{-1}\mathrm{d}\Sigma \\
&= \operatorname{tr}\left(\left[\Sigma^{-1} - \frac{1}{N}\sum_{i=1}^{N}\Sigma^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \Sigma^{-1}\right]\mathrm{d}\Sigma\right) \\
&= \operatorname{tr}\left(\left[\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1}\right]\mathrm{d}\Sigma\right)
\end{aligned}
$$

所以$\frac{\partial l}{\partial \Sigma} = (\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1})^T$。

向量矩阵间求导

机器学习中常见的是标量对向量或矩阵求导，但如果涉及求二阶导，或者使用链式法则，则需要向量对向量求导，或者矩阵对矩阵求导。

向量对向量求导

向量$\mathbf{y}$长度为m，向量$\mathbf{x}$长度为n，$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$结果有两种写法

- 分子布局：得到一个$m \times n$的矩阵，一般叫雅克比矩阵
- 分母布局：得到一个$n \times m$的矩阵，一般叫梯度矩阵

这两者本质相同，只是写法不同，互为转置。上文标量对向量、矩阵求导中使用的是分母布局，因此下文统一也都用分母布局方式，此时

$$
\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}
$$

向量对向量求导，只要写成这种形式

$$
\mathrm{d}\mathbf{y} = \left[\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right]^T \mathrm{d}\mathbf{x}
$$

这与之前标量求导相比，只是少了一个迹。

矩阵对矩阵求导

矩阵$Y_{p\times q}$对矩阵$X_{m\times n}$求导，需要产生出$pq \times mn$个值，为了不产生太高维的数组，我们可以将X, Y矩阵都拉成向量，把各列堆起来即可，如下所示

$$
\mathrm{vec}(X) = \ [X_{11}, \ldots, X_{m1}, X_{12}, \ldots, X_{m2}, \ldots, X_{1n}, \ldots, X_{mn}]^T (\mathrm{mn} \times 1)
$$

则矩阵对矩阵求导，可转化为向量对向量求导，$\frac{\partial Y}{\partial X} = \frac{\partial \mathrm{vec}(Y)}{\partial \mathrm{vec}(X)}(mn \times pq)$，导数与微分的关系如下

$$
\mathrm{vec}(\mathrm{d}Y) = \left[\frac{\partial Y}{\partial X}\right]^T \mathrm{vec}(\mathrm{d}X)
$$

所以矩阵对矩阵求导的步骤为，先两侧取微分，然后两侧取vec，再将$\mathrm{vec}(\mathrm{d}X)$放到最右边即可。这个过程需要用到向量化的性质，以及Kronecker积和交换矩阵相关的恒等式

向量化

- 线性：$\mathrm{vec}(A+B) = \mathrm{vec}(A) + \mathrm{vec}(B)$
- 矩阵乘法：$\mathrm{vec}(AXB) = (B^T \otimes A)\mathrm{vec}(X)$
  - $\mathrm{vec}(AX) = \mathrm{vec}(AXI) = (I \otimes A)\mathrm{vec}(X)$
  - $\otimes$表示Kronecker积，$A_{m\times n} \otimes B_{p\times q} = [A_{ij}B]_{mp\times nq}$
- 转置：$\mathrm{vec}(A^T) = K_{mn}\mathrm{vec}(A_{m\times n})$
  - 其中K_{mn}是交换矩阵，维度为$mn \times mn$，将按列有限的向量化变成按行优先的向量化
- 逐元素乘法：$\mathrm{vec}(A \odot X) = \mathrm{diag}(A)\mathrm{vec}(X)$
  - 其中$\mathrm{diag}(A)$维度为$mn \times mn$，是A中元素按列优先排成的对角阵

Kronecker积和交换矩阵相关的恒等式

- $(A \otimes B)^T = A^T \otimes B^T$
- $\mathrm{vec}(\mathbf{a}\mathbf{b}^T) = \mathbf{b} \otimes \mathbf{a}$

- $(A\otimes B)(C\otimes D)=(AC)\otimes(BD)$
- $K_{mn}=K_{nm}^T,\quad K_{mn}K_{nm}=I$
- $K_{pm}(A\otimes B)K_{nq}=B\otimes A$，其中$A$的维度为$m\times n$，$B$的维度是$p\times q$

向量对矩阵求导，或者矩阵对向量求导，都是按照矩阵对矩阵求导的方式来做，只不过向量取vec是它本身而已。

例题

1、向量对向量求导。$\mathbf{y}=A\mathbf{x}$，求$\frac{\partial\mathbf{y}}{\partial\mathbf{x}}$

解：$\mathrm{d}\mathbf{y}=A\mathrm{d}\mathbf{x}$，所以$\frac{\partial\mathbf{y}}{\partial\mathbf{x}}=A^T$

2、矩阵对矩阵求导。$Y=AX$，求$\frac{\partial Y}{\partial X}$

解：$\mathrm{d}Y=A\mathrm{d}X$，向量化如下

$$\mathrm{vec}(\mathrm{d}Y)=\mathrm{vec}(A\mathrm{d}X)=(I\otimes A)\mathrm{vec}(\mathrm{d}X)$$

所以$\frac{\partial Y}{\partial X}=I\otimes A^T$

3、二阶导。$f=\log|X|$，X维度为$n\times n$，求$\nabla_X f$和$\nabla_X^2 f$。

解：易知$\nabla_X f=X^{-1\mathrm{T}}$，等式两端同时取微分并向量化可得

$$\begin{aligned}\mathrm{vec}(\mathrm{d}\nabla_X f)&=\mathrm{vec}(\mathrm{d}X^{-1\mathrm{T}})\\&=-\mathrm{vec}([X^{-1}\mathrm{d}XX^{-1}]^T)\\&=-K_{nn}\mathrm{vec}(X^{-1}\mathrm{d}XX^{-1})\\&=-K_{nn}(X^{-1\mathrm{T}}\otimes X^{-1})\mathrm{vec}(\mathrm{d}X)\end{aligned}$$

因此$\nabla_X^2 f=-K_{nn}(X^{-1\mathrm{T}}\otimes X^{-1})$，这是个对称矩阵。当$X$是对称矩阵时，$\nabla_X^2 f=X^{-1}\otimes X^{-1}$。

4、逐元素函数。$F=A\exp(XB)$，各矩阵维度分别为$A_{l\times m},X_{m\times n},B_{n\times p}$，求$\frac{\partial F}{\partial X}$。

解：等式两端同时取微分并向量化可得

$$\begin{aligned}\mathrm{vec}(\mathrm{d}F)&=\mathrm{vec}(\mathrm{d}A\exp(XB))\\&=\mathrm{vec}(A\left[\exp(XB)\odot\mathrm{d}XB\right])\\&=(I_p\otimes A)\mathrm{vec}(\left[\exp(XB)\odot\mathrm{d}XB\right])\\&=(I_p\otimes A)\mathrm{diag}(\exp(XB))\mathrm{vec}(\mathrm{d}XB)\\&=(I_p\otimes A)\mathrm{diag}(\exp(XB))(B^T\otimes I_m)\mathrm{vec}(\mathrm{d}X)\end{aligned}$$

因此$\frac{\partial F}{\partial X}=(B\otimes I_m)\mathrm{diag}(\exp(XB))(I_p\otimes A^T)$

链式法则和更多例题请见下一篇。

速查手册 # 数学 # 线性代数

‹ 线性代数总结

矩阵求导总结（二） ›