矩阵求导

· 标量对矩阵求导

一元微分学中倒数与微分的联系

$$df = f'(x) dx$$

。 将其推广到向量

$$df = \sum_{i=1}^n rac{df}{dx_i} dx_i \; ; \; df = ig(rac{\mathbf{df}}{\mathbf{dx}}ig)^T \mathbf{dx}$$

即,在向量中 全微分df是梯度向量 $\dfrac{\partial f}{\partial x}$ (n×1)与微分向量dx(n×1)的内积

。 在矩阵中微分与矩阵倒数的关系可以写成

$$df = \sum_{i=1}^m \sum_{j=1}^n rac{df}{dx_{i,j}} dx_{i,j}$$
 变成乘积模式为 $df = trig(rac{\mathbf{df}}{\mathbf{dX}}^T * \mathbf{dX}ig)$

即,在矩阵中 全微分df是导数 $\dfrac{\partial f}{\partial oldsymbol{x}}$ (m×n)与微分矩阵 $doldsymbol{X}$ (m×n)的内积

- 。 所以根据标量对向量与矩阵的求导形式可变形替换得到结果
- 例4【**线性回归**】: $m{l} = \| m{X} m{w} m{y} \|^2$, 求 $m{w}$ 的最小二乘估计,即求 $\frac{\partial l}{\partial m}$ 的零点。其中 $m{y}$ 是 $m \times 1$ 列向量,X是 $m \times n$ 矩阵,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: $oldsymbol{l} = (oldsymbol{X}oldsymbol{w} - oldsymbol{y})^T (oldsymbol{X}oldsymbol{w} - oldsymbol{y})$,求微分,使用矩阵乘法、转置等法则:

 $dl = (Xdoldsymbol{w})^T(Xoldsymbol{w}-oldsymbol{y}) + (Xoldsymbol{w}-oldsymbol{y})^T(Xdoldsymbol{w}) = 2(Xoldsymbol{w}-oldsymbol{y})^TXdoldsymbol{w}$: 意这里 $Xdoldsymbol{w}$ 和 $Xoldsymbol{w}-oldsymbol{y}$ 是向量,两个向量的内积满足 $oldsymbol{u}^Toldsymbol{v}=oldsymbol{v}^Toldsymbol{u}$ 。对照导数与微分的联系

$$doldsymbol{l} doldsymbol{u} = rac{\partial l}{\partial oldsymbol{w}}^T doldsymbol{w}$$
,得到 $rac{\partial l}{\partial oldsymbol{w}} = 2X^T(Xoldsymbol{w} - oldsymbol{y})$ 。 $rac{\partial l}{\partial oldsymbol{w}} = oldsymbol{0}$ 即 $X^TXoldsymbol{w} = X^Toldsymbol{y}$,

得到 $oldsymbol{w}$ 的最小二乘估计为 $oldsymbol{w}=(X^TX)^{-1}X^Toldsymbol{y}$ 。

例5【方差的最大似然估计】:样本 $oldsymbol{x}_1,\ldots,oldsymbol{x}_N \sim \mathcal{N}(oldsymbol{\mu},\Sigma)$,求方差 $oldsymbol{\Sigma}$ 的最大似然估计。

写成数学式是:
$$l=\log |\Sigma|+rac{1}{N}\sum_{i=1}^N(m{x}_i-m{ar{x}})^T\Sigma^{-1}(m{x}_i-m{ar{x}})$$
,求 $rac{\partial l}{\partial \Sigma}$ 的零点。其

中 $oldsymbol{x_i}$ 是 $oldsymbol{m} imes oldsymbol{1}$ 列向量, $oldsymbol{ar{x}} = rac{1}{N} \sum^{N} oldsymbol{x_i}$ 是样本均值, $oldsymbol{\Sigma}$ 是 $oldsymbol{m} imes oldsymbol{m}$ 对称正定矩阵, $oldsymbol{l}$ 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

$$d\log |\Sigma| = |\Sigma|^{-1} d|\Sigma| = \operatorname{tr}(\Sigma^{-1} d\Sigma)$$
,第二项是

$$rac{1}{N}\sum_{i=1}^N(oldsymbol{x}_i-ar{oldsymbol{x}})^Td\Sigma^{-1}(oldsymbol{x}_i-ar{oldsymbol{x}})=-rac{1}{N}\sum_{i=1}^N(oldsymbol{x}_i-ar{oldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(oldsymbol{x}_i-ar{oldsymbol{x}})$$

。再给第二项套上迹做交换:
$$\operatorname{tr}\left(rac{1}{N}\sum_{i=1}^N(m{x}_i-ar{m{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(m{x}_i-ar{m{x}})
ight)$$

$$=rac{1}{N}\sum_{i=1}^{N} ext{tr}((oldsymbol{x}_i-ar{oldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(oldsymbol{x}_i-ar{oldsymbol{x}}))$$

$$=rac{1}{N}\sum_{i=1}^N ext{tr}\left(\Sigma^{-1}(oldsymbol{x}_i-ar{oldsymbol{x}})(oldsymbol{x}_i-ar{oldsymbol{x}})^T\Sigma^{-1}d\Sigma
ight)= ext{tr}(\Sigma^{-1}S\Sigma^{-1}d\Sigma)$$
, 其中先

交换迹与求和,然后将 $\mathbf{\Sigma}^{-1}(oldsymbol{x}_i-oldsymbol{ar{x}})$ 交换到左边,最后再交换迹与求和,并定义

$$S = rac{1}{N} \sum_{i=1}^N (oldsymbol{x_i} - ar{oldsymbol{x}}) (oldsymbol{x_i} - ar{oldsymbol{x}})^T$$
为样本方差矩阵。得到

$$dl=\mathrm{tr}\left(\left(\Sigma^{-1}-\Sigma^{-1}S\Sigma^{-1}
ight)d\Sigma
ight)$$
。对照导数与微分的联系,有

$$rac{\partial l}{\partial \Sigma} = (\Sigma^{-1} - \Sigma^{-1} S \Sigma^{-1})^T$$
,其零点即 Σ 的最大似然估计为 $\Sigma = S$ 。

例6【多元logistic回归】: $m{l} = -m{y}^T \log \operatorname{softmax}(m{W}m{x})$,求 $\frac{\partial l}{\partial m{W}}$ 。 其中 $m{y}$ 是除一个元素为1外其它元素为0的 $m{m} imes m{1}$ 列向量, $m{W}$ 是 $m{m} imes m{n}$ 矩阵, $m{x}$ 是 $m{n} imes m{1}$ 列向量, $m{l}$ 是标量; \log 表示自然对数, $m{softmax}(m{a}) = \frac{\exp(m{a})}{m{1}^T \exp(m{a})}$,其中 $\exp(m{a})$ 表示逐元素求指数, $m{1}$ 代表全1向量。

解1: 首先将softmax函数代入并写成

$$oxed{l} = -oldsymbol{y}^T \left(\log(\exp(Woldsymbol{x})) - oldsymbol{1} \log(oldsymbol{1}^T \exp(Woldsymbol{x}))
ight) = -oldsymbol{y}^T Woldsymbol{x} + \log(oldsymbol{1}^T \exp(Woldsymbol{x}))$$

,这里要注意逐元素 \log 满足等式 $\log(oldsymbol{u}/c) = \log(oldsymbol{u}) - oldsymbol{1}\log(oldsymbol{c})$,以及 $oldsymbol{y}$ 满足 $oldsymbol{y}^Toldsymbol{1} = oldsymbol{1}$

。求微分,使用矩阵乘法、逐元素函数等法则:

$$dl = -m{y}^T dWm{x} + rac{m{1}^T \left(\exp(Wm{x}) \odot (dWm{x})
ight)}{m{1}^T \exp(Wm{x})}$$
。再套上迹并做交换,注意可化简

$$\mathbf{1}^T \left(\exp(W oldsymbol{x}) \odot (dW oldsymbol{x})
ight) = \exp(W oldsymbol{x})^T dW oldsymbol{x}$$
,这是根据等式

$$\mathbf{1}^T(\boldsymbol{u}\odot\boldsymbol{v})=\boldsymbol{u}^T\boldsymbol{v}$$
, by

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

。对照导数与微分的联系,得到结果

解2: 定义 $oldsymbol{a} = Woldsymbol{x}$,则 $oldsymbol{l} = -oldsymbol{y}^T \log \operatorname{softmax}(oldsymbol{a})$,先同上求出 $\frac{\partial l}{\partial \boldsymbol{a}} = \operatorname{softmax}(\boldsymbol{a}) - \boldsymbol{y}$,再利用复合法则: $d l = ext{tr}\left(rac{\partial l}{\partial m{a}}^T dm{a}
ight) = ext{tr}\left(rac{\partial l}{\partial m{a}}^T dWm{x}
ight) = ext{tr}\left(m{x}rac{\partial l}{\partial m{a}}^T dW
ight)$,得到 $rac{\partial l}{\partial W} = rac{\partial l}{\partial c} m{x}^T$

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】: $l=-m{y}^T\log \operatorname{softmax}(W_2\sigma(W_1m{x}))$,求 $\dfrac{\partial l}{\partial W_2}$ 和 $\dfrac{\partial l}{\partial W_2}$ 。其 中 $m{y}$ 是除一个元素为1外其它元素为0的的 $m{m} imes m{1}$ 列向量, $m{W_2}$ 是 $m{m} imes m{p}$ 矩阵, $m{W_1}$ 是 $m{p} imes m{n}$ 矩 阵, $m{x}$ 是 $m{n} imes m{1}$ 列向量, $m{l}$ 是标量;log表示自然对数, $m{softmax}(m{a}) = rac{\exp(m{a})}{\mathbf{1}^T \exp(m{a})}$ 同上,

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

解: 定义 $oldsymbol{a}_1 = W_1oldsymbol{x}$, $oldsymbol{h}_1 = \sigma(oldsymbol{a}_1)$, $oldsymbol{a}_2 = W_2oldsymbol{h}_1$,则

 $m{l} = -m{y}^T \log \operatorname{softmax}(m{a_2})$ 。在前例中已求出 $rac{\partial l}{\partial m{a_2}} = \operatorname{softmax}(m{a_2}) - m{y}$ 。使用复合

法则,
$$doldsymbol{l} = \operatorname{tr}\left(rac{\partial oldsymbol{l}}{\partial oldsymbol{a}_2}^T doldsymbol{a}_2
ight) = \operatorname{tr}\left(rac{\partial oldsymbol{l}}{\partial oldsymbol{a}_2}^T dW_2oldsymbol{h}_1
ight) + \underbrace{\operatorname{tr}\left(rac{\partial oldsymbol{l}}{\partial oldsymbol{a}_2}^T W_2 doldsymbol{h}_1
ight)}_{dl_2}$$

,使用矩阵乘法交换的迹技巧从第一项得到 $\dfrac{\partial l}{\partial W_2}=\dfrac{\partial l}{\partial m{a}_2}m{h}_1^T$,从第二项得到

$$rac{\partial l}{\partial m{h}_1} = W_2^T rac{\partial l}{\partial m{a}_2}$$
。接下来对第二项继续使用复合法则来求 $rac{\partial l}{\partial m{a}_1}$,并利用矩阵乘法和逐元素

乘法交换的迹技巧:
$$dl_2=\mathrm{tr}\left(rac{\partial l}{\partial m{h}_1}^T dm{h}_1
ight)=\mathrm{tr}\left(rac{\partial l}{\partial m{h}_1}^T (\sigma'(m{a}_1)\odot dm{a}_1)
ight)=\mathrm{tr}\left(\left(rac{\partial l}{\partial m{h}_1}\odot \sigma'(m{a}_1)
ight)^T dm{a}_1
ight)$$
,得到 $\frac{\partial l}{\partial m{a}_1}=rac{\partial l}{\partial m{h}_1}\odot \sigma'(m{a}_1)$ 。 为求 $\frac{\partial l}{\partial W_1}$,再用一次复合法则: $dl_2=\mathrm{tr}\left(rac{\partial l}{\partial m{a}_1}^T dm{a}_1
ight)=\mathrm{tr}\left(rac{\partial l}{\partial m{a}_1}^T dW_1m{x}
ight)=\mathrm{tr}\left(m{x}rac{\partial l}{\partial m{a}_1}^T dW_1
ight)$,得到 $\frac{\partial l}{\partial W_1}=rac{\partial l}{\partial m{a}_1}m{x}^T$ 。

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

$$l = -\sum_{i=1}^N m{y}_i^T \log \operatorname{softmax}(W_2 \sigma(W_1 m{x}_i + m{b}_1) + m{b}_2)$$
,其中 $m{b}_1$ 是 $m{p} imes m{1}$ 列向量,

 b_2 是 $m \times 1$ 列向量,其余定义同上。

解1:定义 $m{a}_{1,i}=W_1m{x}_i+m{b}_1$, $m{h}_{1,i}=\sigma(m{a}_{1,i})$, $m{a}_{2,i}=W_2m{h}_{1,i}+m{b}_2$,则

$$l = -\sum_{i=1}^N m{y}_i^T \log \operatorname{softmax}(m{a}_{2,i})$$
。先同上可求出 $rac{\partial l}{\partial m{a}_{2,i}} = \operatorname{softmax}(m{a}_{2,i}) - m{y}_i$ 。

使用复合法则。

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

,从第一项得到得到
$$rac{\partial l}{\partial W_2}=\sum_{i=1}^Nrac{\partial l}{\partial m{a}_{2,i}}m{h}_{1,i}^T$$
,从第二项得到 $rac{\partial l}{\partial m{h}_{1,i}}=W_2^Trac{\partial l}{\partial m{a}_{2,i}}$,从

第三项得到到
$$\dfrac{\partial l}{\partial oldsymbol{b_2}} = \sum_{i=1}^N \dfrac{\partial l}{\partial oldsymbol{a_{2,i}}}$$
。接下来对第二项继续使用复合法则,得到

$$rac{\partial l}{\partial m{a}_{1,i}} = rac{\partial l}{\partial m{h}_{1,i}} \odot \sigma'(m{a}_{1,i})$$
。为求 $rac{\partial l}{\partial W_1}, rac{\partial l}{\partial m{b}_1}$,再用一次复合法则:

$$dl_2 = ext{tr}\left(\sum_{i=1}^N rac{\partial l}{\partial oldsymbol{a}_{1,i}}^T doldsymbol{a}_{1,i}
ight) = ext{tr}\left(\sum_{i=1}^N rac{\partial l}{\partial oldsymbol{a}_{1,i}}^T dW_1oldsymbol{x}_i
ight) + ext{tr}\left(\sum_{i=1}^N rac{\partial l}{\partial oldsymbol{a}_{1,i}}^T doldsymbol{b}_1
ight)$$

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

解2:可以用矩阵来表示N个样本,以简化形式。定义 $X=[oldsymbol{x}_1,\cdots,oldsymbol{x}_N]$,

$$A_1 = [m{a}_{1,1}, \cdots, m{a}_{1,N}] = W_1 X + m{b}_1 m{1}^T$$
 , $H_1 = [m{h}_{1,1}, \cdots, m{h}_{1,N}] = \sigma(A_1)$

,
$$A_2 = [m{a}_{2,1}, \cdots, m{a}_{2,N}] = W_2 H_1 + m{b}_2 m{1}^T$$
,注意这里使用全1向量来扩展维度。先同

上求出
$$rac{\partial l}{\partial A_2} = [ext{softmax}(m{a}_{2,1}) - m{y}_1, \cdots, ext{softmax}(m{a}_{2,N}) - m{y}_N]$$
。使用复合法

则,

$$dl = ext{tr}\left(rac{\partial l}{\partial A_2}^T dA_2
ight) = ext{tr}\left(rac{\partial l}{\partial A_2}^T dW_2 H_1
ight) + ext{tr}\left(rac{\partial l}{\partial A_2}^T W_2 dH_1
ight) + ext{tr}\left(rac{\partial l}{\partial A_2}^T doldsymbol{b}_2 oldsymbol{1}^T
ight)$$

,从第一项得到
$$rac{\partial l}{\partial W_2}=rac{\partial l}{\partial A_2}H_1^T$$
,从第二项得到 $rac{\partial l}{\partial H_1}=W_2^Trac{\partial l}{\partial A_2}$,从第三项得到到

$$rac{\partial l}{\partial m{b}_2} = rac{\partial l}{\partial A_2} m{1}$$
。接下来对第二项继续使用复合法则,得到 $rac{\partial l}{\partial A_1} = rac{\partial l}{\partial H_1} \odot \sigma'(A_1)$ 。

为求 $\frac{\partial l}{\partial W_1}$, $\frac{\partial l}{\partial h_1}$, 再用一次复合法则:

$$dl_2 = \mathrm{tr}\left(rac{\partial l}{\partial A_1}^T dA_1
ight) = \mathrm{tr}\left(rac{\partial l}{\partial A_1}^T dW_1 X
ight) + \mathrm{tr}\left(rac{\partial l}{\partial A_1}^T doldsymbol{b}_1oldsymbol{1}^T
ight)$$
, (4)

到
$$rac{\partial l}{\partial W_1}=rac{\partial l}{\partial A_1}X^T$$
, $rac{\partial l}{\partial m{b}_1}=rac{\partial l}{\partial A_1}m{1}$ 。

. 矩阵对矩阵求导

- o 向量m对向量n的求导得到一个mxn的矩阵
- 。 对于矩阵,可以进行行/列优先的向量转化
- \circ 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 $rac{\partial F}{\partial X}=rac{\partial \mathrm{vec}(F)}{\partial \mathrm{vec}(X)}$ (mn×pq)。
- 。 例1: F=AX, X是m×n矩阵, 求 $\dfrac{\partial F}{\partial X}$ 。

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 $rac{\partial F}{\partial X}=I_n\otimes A^T$ 。

特例:如果X退化为向量,即 $m{f}=Am{x}$,则根据向量的导数与微分的关系 $dm{f}=rac{\partial m{f}}{\partial m{x}}^Tdm{x}$,得到 $rac{\partial m{f}}{\partial m{x}}=A^T$ 。