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

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

即，在向量中 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

  - 在矩阵中微分与矩阵倒数的关系可以写成

$$df=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{df}{dx_{i,j}}dx_{i,j}\ \text{变成乘积模式为}df=tr\Big(\frac{\mathbf{df}}{\mathbf{dX}}^T*\mathbf{dX}\Big)$$

即，在矩阵中 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

  - 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】**：$l=\|X\boldsymbol{w}-\boldsymbol{y}\|^2$，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：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$dl=\frac{\partial l}{\partial \boldsymbol{w}}^Td\boldsymbol{w}$，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$\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)$，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 $\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T d\Sigma^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}) = -\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})$。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 $dl = \text{tr}\left(\left(\Sigma^{-1}-\Sigma^{-1}S\Sigma^{-1}\right)d\Sigma\right)$。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；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 $l = -\boldsymbol{y}^T\left(\log(\exp(W\boldsymbol{x})) - \boldsymbol{1}\log(\boldsymbol{1}^T\exp(W\boldsymbol{x}))\right) = -\boldsymbol{y}^TW\boldsymbol{x} + \log(\boldsymbol{1}^T\exp(W\boldsymbol{x}))$，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：$dl = -\boldsymbol{y}^TdW\boldsymbol{x} + \frac{\boldsymbol{1}^T\left(\exp(W\boldsymbol{x})\odot(dW\boldsymbol{x})\right)}{\boldsymbol{1}^T\exp(W\boldsymbol{x})}$。再套上迹并做交换，注意可化简 $\boldsymbol{1}^T\left(\exp(W\boldsymbol{x})\odot(dW\boldsymbol{x})\right) = \exp(W\boldsymbol{x})^TdW\boldsymbol{x}$，这是根据等式 $\boldsymbol{1}^T(\boldsymbol{u}\odot\boldsymbol{v}) = \boldsymbol{u}^T\boldsymbol{v}$，故 $dl = \text{tr}\left(-\boldsymbol{y}^TdW\boldsymbol{x} + \frac{\exp(W\boldsymbol{x})^TdW\boldsymbol{x}}{\boldsymbol{1}^T\exp(W\boldsymbol{x})}\right) = \text{tr}(-\boldsymbol{y}^TdW\boldsymbol{x} + \text{softmax}(W\boldsymbol{x})^TdW\boldsymbol{x}) = \text{tr}(\boldsymbol{x}(\text{softmax}(W\boldsymbol{x})-\boldsymbol{y})^TdW)$。对照导数与微分的联系，得到结果

解2：定义$\boldsymbol{a} = W\boldsymbol{x}$，则$l = -\boldsymbol{y}^T \log \mathrm{softmax}(\boldsymbol{a})$，先同上求出$\frac{\partial l}{\partial \boldsymbol{a}} = \mathrm{softmax}(\boldsymbol{a}) - \boldsymbol{y}$，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$\frac{\partial l}{\partial W} = \frac{\partial l}{\partial \boldsymbol{a}}\boldsymbol{x}^T$。

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】：$l = -\boldsymbol{y}^T \log \mathrm{softmax}(W_2\sigma(W_1\boldsymbol{x}))$，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$p \times n$矩阵，$\boldsymbol{x}$是$n \times 1$列向量，$l$是标量；log表示自然对数，$\mathrm{softmax}(\boldsymbol{a}) = \frac{\exp(\boldsymbol{a})}{\boldsymbol{1}^T \exp(\boldsymbol{a})}$同上，$\boldsymbol{\sigma}$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$l = -\boldsymbol{y}^T \log \mathrm{softmax}(\boldsymbol{a}_2)$。在前例中已求出$\frac{\partial l}{\partial \boldsymbol{a}_2} = \mathrm{softmax}(\boldsymbol{a}_2) - \boldsymbol{y}$。使用复合法则，$dl = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_2}^T d\boldsymbol{a}_2\right) = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_2}^T dW_2\boldsymbol{h}_1\right) + \underbrace{\mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_2}^T W_2 d\boldsymbol{h}_1\right)}_{dl_2}$，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：$dl_2 = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{h}_1}^T d\boldsymbol{h}_1\right) = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{h}_1}^T (\sigma'(\boldsymbol{a}_1) \odot d\boldsymbol{a}_1)\right) = \mathrm{tr}\left(\left(\frac{\partial l}{\partial \boldsymbol{h}_1} \odot \sigma'(\boldsymbol{a}_1)\right)^T d\boldsymbol{a}_1\right)$，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：$dl_2 = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_1}^T d\boldsymbol{a}_1\right) = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_1}^T dW_1\boldsymbol{x}\right) = \mathrm{tr}\left(\boldsymbol{x}\frac{\partial l}{\partial \boldsymbol{a}_1}^T dW_1\right)$，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，

$$dl = \operatorname{tr}\left(\sum_{i=1}^{N} \frac{\partial l}{\partial \boldsymbol{a}_{2,i}}^T d\boldsymbol{a}_{2,i}\right) = \operatorname{tr}\left(\sum_{i=1}^{N} \frac{\partial l}{\partial \boldsymbol{a}_{2,i}}^T dW_2 \boldsymbol{h}_{1,i}\right) + \underbrace{\operatorname{tr}\left(\sum_{i=1}^{N} \frac{\partial l}{\partial \boldsymbol{a}_{2,i}}^T W_2 d\boldsymbol{h}_{1,i}\right)}_{dl_2} + \operatorname{tr}\left(\sum_{i=1}^{N} \frac{\partial l}{\partial \boldsymbol{a}_{2,i}}^T d\boldsymbol{b}_2\right)$$

，从第一项得到得到$\frac{\partial l}{\partial W_2} = \sum_{i=1}^{N} \frac{\partial l}{\partial \boldsymbol{a}_{2,i}} \boldsymbol{h}_{1,i}^T$，从第二项得到$\frac{\partial l}{\partial \boldsymbol{h}_{1,i}} = W_2^T \frac{\partial l}{\partial \boldsymbol{a}_{2,i}}$，从第三项得到到$\frac{\partial l}{\partial \boldsymbol{b}_2} = \sum_{i=1}^{N} \frac{\partial l}{\partial \boldsymbol{a}_{2,i}}$。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$\frac{\partial l}{\partial W_1} = \frac{\partial l}{\partial A_1} X^T$，$\frac{\partial l}{\partial \boldsymbol{b}_1} = \frac{\partial l}{\partial A_1} \boldsymbol{1}$。

- ## 矩阵对矩阵求导
    - 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
    - 向量化$\mathbf{vec}(X) = [X_{11}, \dots, X_{m1}, X_{12}, \dots, X_{m2}, \dots, X_{1n}, \dots, X_{mn}]^T$(mn×1)，并定义矩阵F对矩阵X的导数$\frac{\partial F}{\partial X} = \frac{\partial \mathbf{vec}(F)}{\partial \mathbf{vec}(X)}$(mn×pq)。
    - 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：$\mathbf{vec}(dF) = \mathbf{vec}(AdX) = (I_n \otimes A)\mathbf{vec}(dX)$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$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = A^T$。