矩阵求导

3种标准导数(梯度)公式

1) 自变量是一个标量(Scalar)时：

$$Df(x) = lim_{t\to 0}\frac{f(x+t)-f(x)}{t}$$

2) 自变量是一个向量(Vector)时：

$$D_w f(\boldsymbol{x}) = lim_{t\to 0}\frac{f(\boldsymbol{x}+t\boldsymbol{w})-f(\boldsymbol{x})}{t}$$

(w的维数和x一致) 这个导数的含义是，在n维空间中f(x)所定义的(超)平面上的某个坐标点x相对于w的斜率。

3) 自变量是一个矩阵(Matrix)时：

$$D_W f(\boldsymbol{X}) = lim_{t\to 0}\frac{f(\boldsymbol{X}+t\boldsymbol{W})-f(\boldsymbol{X})}{t}$$

含义和2)类似。（已经无法想象了）

标量f对矩阵X的导数-Trace derivative

标量f对矩阵X的导数，定义为:

$$\frac{\partial f}{\partial X} = \left[\frac{\partial f}{\partial X_{ij}}\right]$$

矩阵导数与微分建立联系：

> $$df = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial f}{\partial X_{ij}}dX_{ij} = \operatorname{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$$

矩阵运算法则

- 加减法：$d(X\pm Y)=dX\pm dY$；矩阵乘法：$d(XY)=(dX)Y+X(dY)$；转置：$d(X^T)=(dX)^T$；迹：$d\operatorname{tr}(X)=\operatorname{tr}(dX)$。
- 逆：$dX^{-1}=-X^{-1}(dX)X^{-1}$。此式可在$XX^{-1}=I$两侧求微分来证明。
- 行列式：$d|X|=\operatorname{tr}(X^{\#}dX)$，其中$X^{\#}$表示$X$的伴随矩阵，在$X$可逆时又可以写作$d|X|=|X|\operatorname{tr}(X^{-1}dX)$。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。
- 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。
- 逐元素函数：$d\sigma(X)=\sigma'(X)\odot dX$，$\sigma(X)=\left[\sigma(X_{ij})\right]$是逐元素运算的标量函数。

Trace trick

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

Example:

> 【线性回归】：$l=\|X\boldsymbol{w}-\boldsymbol{y}\|^2$， 求$\boldsymbol{w}$的最小二乘估计，即求$\frac{\partial l}{\partial w}$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$d(X^TY)=(dX)^TY+X^T(dY)$：

$$dl = (Xd\boldsymbol{w})^T(X\boldsymbol{w}-\boldsymbol{y}) + (X\boldsymbol{w}-\boldsymbol{y})^T(Xd\boldsymbol{w}) = 2(X\boldsymbol{w}-\boldsymbol{y})^TXd\boldsymbol{w}$$

。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

$$\boldsymbol{w} = (X^TX)^{-1}X^T\boldsymbol{y}$$

> 【多元logistic回归】：$l = -\boldsymbol{y}^T\log\mathrm{softmax}(W\boldsymbol{x})$，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；$\mathrm{softmax}(\boldsymbol{a}) = \frac{\exp(\boldsymbol{a})}{\boldsymbol{1}^T\exp(\boldsymbol{a})}$，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：

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

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

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

。对照导数与微分的联系，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：

$$dl = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}}^Td\boldsymbol{a}\right) = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}}^TdW\boldsymbol{x}\right) = \mathrm{tr}\left(\boldsymbol{x}\frac{\partial l}{\partial \boldsymbol{a}}^TdW\right)$$

，得到

$$\frac{\partial l}{\partial W} = \frac{\partial l}{\partial \boldsymbol{a}}\boldsymbol{x}^T$$

> 【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；$\mathrm{softmax}(\boldsymbol{a}) = \frac{\exp(\boldsymbol{a})}{\boldsymbol{1}^T\exp(\boldsymbol{a})}$同例2，$\sigma(\cdot)$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$l = -\boldsymbol{y}^T\log\mathrm{softmax}(\boldsymbol{a}_2)$。在例2中已求出$\frac{\partial l}{\partial \boldsymbol{a}_2} = \mathrm{softmax}(\boldsymbol{a}_2) - \boldsymbol{y}$。使用复合法则，注意此处$\boldsymbol{h}_1, W_2$都是变量：

$$dl = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_2}^Td\boldsymbol{a}_2\right) = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_2}^TdW_2\boldsymbol{h}_1\right) + \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_2}^TW_2d\boldsymbol{h}_1\right)$$

，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$\frac{\partial l}{\partial \boldsymbol{h}_1} = W_2^T\frac{\partial l}{\partial \boldsymbol{a}_2}$。接下来求$\frac{\partial l}{\partial \boldsymbol{a}_1}$，继续使用复合法则，并利用矩阵乘法和逐元素乘法交换的迹技巧：

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

，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：

$$\mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_1}^Td\boldsymbol{a}_1\right) = \mathrm{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_1}^TdW_1\boldsymbol{x}\right) = \mathrm{tr}\left(\boldsymbol{x}\frac{\partial l}{\partial \boldsymbol{a}_1}^TdW_1\right)$$

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

矩阵$F(p\times q)$对矩阵$X(m\times n)$的导数