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:

$$Df(x) = lim_{t
ightarrow 0} \, rac{f(x+t) - f(x)}{t}$$

2) 自变量是一个向量(Vector)时:

$$D_w f(oldsymbol{x}) = lim_{t o 0} \, rac{f(oldsymbol{x} + toldsymbol{w}) - f(oldsymbol{x})}{t}$$

(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:

$$D_W f(m{X}) = lim_{t
ightarrow 0} \, rac{f(m{X} + tm{W}) - f(m{X})}{t}$$

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:

$$rac{\partial f}{\partial X} = \left[rac{\partial f}{\partial X_{ij}}
ight]$$

矩阵导数与微分建立联系:

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

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: $d\mathrm{tr}(X)=\mathrm{tr}(dX)$ 。
- 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 = tr(a)
- 转置: $\operatorname{tr}(A^T) = \operatorname{tr}(A)$.
- 线性: $tr(A \pm B) = tr(A) \pm tr(B)$ 。
- 矩阵乘法交换: tr(AB) = tr(BA), 其中 $A = B^T$ 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: $\operatorname{tr}(A^T(B \odot C)) = \operatorname{tr}((A \odot B)^T C)$, 其中A, B, C尺寸相同。两侧都等于 $\sum_{i,j} A_{ij} B_{ij} C_{ij}$ 。

Example:

【线性回归】: $l = ||Xw - y||^2$, 求w的最小二乘估计,即求 $\frac{\partial l}{\partial w}$ 的零点。其中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: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 $d(X^T Y) = (dX)^T Y + X^T (dY)$:

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

。对照导数与微分的联系 $dl=rac{\partial l^T}{\partial w}dw$,得到 $rac{\partial l}{\partial w}=\left(2(Xw-y)^TX\right)^T=2X^T(Xw-y)$ 。 $rac{\partial l}{\partial w}$ 的零点即w的最小二乘估计为

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

【多元logistic回归】: $l = -y^T \log \operatorname{softmax}(Wx)$, 求 $\frac{\partial l}{\partial W}$ 。其中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,x是 $n \times 1$ 列向量,l是标量; $\operatorname{softmax}(a) = \frac{\exp(a)}{1^T \exp(a)}$,其中 $\exp(a)$ 表示逐元素求指数,1代表全1向量。

解: 首先将softmax函数代入并写成

$$l = -\boldsymbol{y}^T \left(\log(\exp(W\boldsymbol{x})) - \boldsymbol{1} \log(\boldsymbol{1}^T \exp(W\boldsymbol{x})) \right) = -\boldsymbol{y}^T W \boldsymbol{x} + \log(\boldsymbol{1}^T \exp(W\boldsymbol{x}))$$

,这里要注意逐元素 \log 满足等式 $\log(\boldsymbol{u}/c) = \log(\boldsymbol{u}) - \mathbf{1}\log(c)$,以及 \boldsymbol{y} 满足 $\boldsymbol{y}^T\mathbf{1} = 1$ 。求微分,使用矩阵乘法、逐元素函数等法则:

$$dl = -oldsymbol{y}^T dWoldsymbol{x} + rac{oldsymbol{1}^T (\exp(Woldsymbol{x}) \odot (dWoldsymbol{x}))}{oldsymbol{1}^T \exp(Woldsymbol{x})}$$

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

$$dl = \operatorname{tr}\!\left(-oldsymbol{y}^T dW oldsymbol{x} + rac{\exp(W oldsymbol{x})^T dW oldsymbol{x}}{oldsymbol{1}^T \exp(W oldsymbol{x})}
ight) = \operatorname{tr}(oldsymbol{x}(\operatorname{softmax}(W oldsymbol{x}) - oldsymbol{y})^T dW)$$

。对照导数与微分的联系,得到 $\frac{\partial l}{\partial W} = (\operatorname{softmax}(W\boldsymbol{x}) - \boldsymbol{y})\boldsymbol{x}^T$ 。

另解: 定义a=Wx,则 $l=-y^T\log \operatorname{softmax}(a)$,先如上求出 $\frac{\partial l}{\partial a}=\operatorname{softmax}(a)-y$,再利用复合法则:

$$dl = ext{tr}igg(rac{\partial l}{\partial oldsymbol{a}}^T doldsymbol{a}igg) = ext{tr}igg(oldsymbol{x}rac{\partial l}{\partial oldsymbol{a}}^T dWoldsymbol{x}igg) = ext{tr}igg(oldsymbol{x}rac{\partial l}{\partial oldsymbol{a}}^T dWigg)$$

, 得到

$$rac{\partial l}{\partial W} = rac{\partial l}{\partial oldsymbol{a}} \, oldsymbol{x}^T$$

【二层神经网络】: $l = -\mathbf{y}^T \log \operatorname{softmax}(W_2\sigma(W_1\mathbf{x}))$, 求 $\frac{\partial l}{\partial W_1}$ 和 $\frac{\partial l}{\partial W_2}$ 。其中 \mathbf{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, \mathbf{x} 是 $n \times 1$ 列向量,l是标量; $\operatorname{softmax}(\mathbf{a}) = \frac{\exp(\mathbf{a})}{\mathbf{1}^T \exp(\mathbf{a})}$ 同例 2, $\sigma(\cdot)$ 是逐元素 $\operatorname{sigmoid}$ 函数 $\sigma(\mathbf{a}) = \frac{1}{1+\exp(-\mathbf{a})}$ 。

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 $l = -\boldsymbol{y}^T \log \operatorname{softmax}(\boldsymbol{a}_2)$ 。 在 例 2 中 已 求 出 $\frac{\partial l}{\partial \boldsymbol{a}_2} = \operatorname{softmax}(\boldsymbol{a}_2) - \boldsymbol{y}$ 。 使用复合法则,注意此处 \boldsymbol{h}_1, W_2 都是变量:

$$dl = ext{tr}igg(rac{\partial l}{\partial oldsymbol{a}_2}^T doldsymbol{a}_2igg) = ext{tr}igg(rac{\partial l}{\partial oldsymbol{a}_2}^T dW_2oldsymbol{h}_1igg) + ext{tr}igg(rac{\partial l}{\partial oldsymbol{a}_2}^T W_2 doldsymbol{h}_1igg)$$

,使用矩阵乘法交换的迹技巧从第一项得到 $\frac{\partial l}{\partial W_2} = \frac{\partial l}{\partial a_2} h_1^T$,从第二项得到 $\frac{\partial l}{\partial h_1} = W_2^T \frac{\partial l}{\partial a_2}$ 。接下来求 $\frac{\partial l}{\partial a_1}$,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:

$$\mathrm{tr}igg(rac{\partial l}{\partial oldsymbol{h}_1}^T doldsymbol{h}_1igg) = \mathrm{tr}igg(rac{\partial l}{\partial oldsymbol{h}_1}^T (\sigma'(oldsymbol{a}_1)\odot doldsymbol{a}_1)igg) = \mathrm{tr}igg(igg(rac{\partial l}{\partial oldsymbol{h}_1}\odot \sigma'(oldsymbol{a}_1)igg)^T doldsymbol{a}_1igg)$$

,得到 $\frac{\partial l}{\partial a_1} = \frac{\partial l}{\partial h_1} \odot \sigma'(a_1)$ 。为求 $\frac{\partial l}{\partial W_1}$,再用一次复合法则:

$$\mathrm{tr}igg(rac{\partial l}{\partial oldsymbol{a}_1}^T doldsymbol{a}_1igg) = \mathrm{tr}igg(rac{\partial l}{\partial oldsymbol{a}_1}^T dW_1oldsymbol{x}igg) = \mathrm{tr}igg(oldsymbol{x}rac{\partial l}{\partial oldsymbol{a}_1}^T dW_1igg)$$

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

矩阵 $F(p \times q)$ 对矩阵 $X(m \times n)$ 的导数