# 小锅的机器学习笔记–数学基础

## 数据集

每一个样本都是一个 $p$ 维向量，记为:

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \tag{1}$$

我们不妨假设收集到的数据集一共有 $N$ 个样本点，那么数据集可用 $X_{N\times p}$ 来表示，记为：

$$X_{N\times p} = (x_1, x_2, \ldots, x_N)^T = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{Np} \end{pmatrix}_{N\times p} \tag{2}$$

## 频率派

我们认为 $\theta$ 是一个未知的常量，而 $X_{N\times p}$ 是一个随机变量。那么每一个观测都是由 $p\left(x_i|\theta\right)$ 所产生的，假设样本之间相互独立，那么对于整个数据集的观测为：

$$p\left(X_{N\times p}|\theta\right) = \prod_{i=1}^{N} p\left(x_i|\theta\right) \tag{3}$$

为了求最合适的 $\theta$，我们采用 $MLE$(极大对数释然估计) 的方法求解：

$$\theta_{MLE} = \underset{\theta}{argmax} \log\, p\left(X_{N\times p}|\,\theta\right) = \underset{\theta}{argmax} \sum_{i=1}^{N} \log\, p\left(x_i|\theta\right) \tag{4}$$

## 贝叶斯派

贝叶斯认为：$\theta$ 不是一个常量，而是满足一个预设的先验分布 $\theta$ $p(\theta)$，那么依赖观察集的后验可以写成是:

$$p(\theta|X) = \frac{p(\theta, X)}{p(X)} = \frac{p(X|\theta)\times p(\theta)}{p(X)} \tag{5}$$

如果这个 $p$ 可能是离散型的随机变量，或者是连续型的随机变量。

$$p(X) = \begin{cases} \int p(X|\theta)\times p(\theta)d\theta & \text{if } \theta\text{是连续型的随机变量} \\ \sum_{i=0}^{n} p(X|\theta_i)\times p(\theta_i) & \text{if } \theta\text{是离散型的随机变量} \end{cases}$$

可以注意到分母 $p(X)$ 和 $\theta$ 没有关系，为了求最合适的 $\theta$, 最大化这个 $p(\theta|X)$：

$$\theta_{MAP} = \underset{\theta}{argmax}\, p(\theta|X) = \underset{\theta}{argmax}\, p(X|\theta)\times p(\theta) \tag{6}$$

## 高斯分布

### Data 假设

$$x_i = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1p} \end{pmatrix} \tag{7}$$

$$X_{N\times p} = (x_1, x_2, \ldots, x_N)^T = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{Np} \end{pmatrix}_{N\times p} \tag{8}$$

### 一维的高斯分布

设 $p=1$ 且我们设 $X$ 满足高斯分布 $N(\mu,\sigma)$ $X$ 的概率密度为为:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{9}$$

最后利用极大释然估计 $X$ 的分布 $N(\mu,\sigma)$

$$\begin{aligned} \log p(X) &= \log \prod_{i=0}^{N} p\left(x_i|\theta\right) \\ &= \sum_{i=1}^{N} \log\ p\left(x_i|\theta\right) \\ &= -\sum_{i=1}^{N} \log \sqrt{2\pi} + \log\sigma + \frac{\left(x_i-\mu\right)^2}{2\sigma^2} \end{aligned}$$

那么 $\mu_{MLE}$ 为:

$$\begin{aligned} \mu_{MLE} &= \underset{\mu}{argmax} \log p(X) \\ &= \underset{\mu}{argmax}\ -\sum_{i=1}^{N} \left(x_i-\mu\right)^2 \\ &= \underset{\mu}{argmin} \sum_{i=1}^{N} \left(x_i-\mu\right)^2 \end{aligned}$$

因为:

$$\frac{\partial \sum_{i=1}^{N} \left(x_i-\mu\right)^2}{\partial\mu} = 2\sum_{i=1}^{N} \left(u-x_i\right) = 0 \tag{10}$$

所以有：

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{11}$$

对于另外一个参数 $\sigma$:

$$\begin{aligned}
\sigma_{MLE} &= \underset{\sigma}{argmax} \log p(X) \\
&= \underset{\sigma}{argmax}\ -\sum_{i=1}^{N}(\log\sigma + \frac{(x_i-\mu)^2}{2\sigma^2}) \\
&= \underset{\sigma}{argmin}\ \sum_{i=1}^{N}(\log\sigma + \frac{(x_i-\mu)^2}{2\sigma^2})
\end{aligned}$$

因为:

$$\frac{\partial \sum_{i=1}^{N}(\log\sigma + \frac{(x_i-\mu)^2}{2\sigma^2})}{\partial \sigma^2} = \sum_{i=1}^{N}\frac{1}{2\sigma^2} - \frac{(x_i-\mu)^2}{2\sigma^4} = 0 \tag{12}$$

所以有:

$$\sigma_{MLE}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \tag{13}$$

求解 $\theta$ 的分布时，我们是先计算出 $\mu_{MLE}$ 然后利用 $\mu_{MLE}$ 求解得到 $\sigma_{MLE}^2$。

$$E[\mu_{MLE}] = E[\frac{1}{N}\sum_{i=1}^{N} x_i] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \mu \tag{14}$$

$$\begin{aligned}
E[\sigma_{MLE}^2] &= E[\frac{1}{N}\ \sum_{i=1}^{N}\ x_i^2] - \mu_{MLE}^2 \\
&= \frac{1}{N}E[\sum_{i=1}^{N}(x_i^2-\mu^2)] - E[\mu_{MLE}^2-\mu^2] \\
&= \frac{1}{N}\sum_{i=1}^{N}(E[x_i^2]-E[x_i]^2) - (E[\mu_{MLE}^2]-\mu^2)
\end{aligned}$$

因为:

$$\begin{aligned}
\mu &= E[\mu_{MLE}] \\
&= \frac{1}{N}\sum_{i=1}^{N}(E[x_i^2]-E[x_i]^2) - (E[\mu_{MLE}^2]-E[\mu_{MLE}]^2) \\
&= \frac{1}{N}\sum_{i=1}^{N} Var[x_i] - Var[u_{MLE}]
\end{aligned}$$

因为:

$$Var[u_{MLE}] = \frac{1}{N^2}\sum_{i=1}^{N} Var[x_i] = \frac{1}{N}\sigma^2$$

$$Var[x_i] = \sigma^2$$

所以:

$$E[\sigma^2_{MLE}] = \sigma^2 - \frac{1}{N}\sigma^2 = \frac{N-1}{N}\sigma^2$$

所以可以发现对 $\mu_{MLE}$ 是无偏的，但是对 $\sigma_{MLE}$ 的估计是有偏的，上述的点估计方法会把 $\sigma$ 估计小，通常取：

$$\sigma = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2 \tag{15}$$

来进行修正。

### 多维的高斯分布

首先 $x$ 是一个 $p$ 维的随机变量：

$$x \quad \sim\sim \quad N(\mu, \sigma) \quad \hookrightarrow \quad f(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\text{exp}[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))] \tag{16}$$

其中 $\mu$ 为：

$$\mu = (u_1, u_2, \ldots, u_p)^T \tag{17}$$

$\Sigma$ 协方差矩阵，一般而言是半正定矩阵，这里我们只考虑正定矩阵:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \sigma_{31} & \sigma_{32} & \ldots & \sigma_{3p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix} \tag{18}$$

因为 $\Sigma$==$\Sigma^T$，所以这个 $\Sigma$ 一定可以奇异值分解（也就是相似对角化）。

$$\Sigma = U\Lambda U^T \tag{19}$$

其中：

$$U = (u_1, u_2, u_3, ......, u_p)_{p\times p} \quad u_i\text{是 } p \text{ 维列向量} \quad \text{并且: } UU^T = U^TU = E$$

$$\Lambda = diag(\lambda_1, \lambda_2, ...., \lambda_p)$$

所以有:

$$\Sigma = U\lambda U^T = (u_1, u_2, ..., u_p)\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_p \end{pmatrix}\begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_p^T \end{pmatrix}$$

$$= (u_1\lambda_1, u_2\lambda_2, ..., u_p\lambda_p)\begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_p^T \end{pmatrix}$$

$$= \sum_{i=1}^{p} u_i\lambda_i u_i^T$$

所以：

$$\Sigma^{-1} = (U\lambda U^T)^{-1} = U\lambda^{-1}U^T = \sum_{i=1}^{p} u_i\frac{1}{\lambda_i}u_i^T \tag{20}$$

现在我们设：

$$\triangle = (x-\mu)^T\Sigma^{-1}(x-\mu) = \sum_{i=1}^{p}(x-\mu)^T u_i\frac{1}{\lambda_i}u_i^T(x-\mu) \tag{21}$$

显然有:

$$(x-\mu)^T u_i = u_i^T(x-\mu)$$

令：

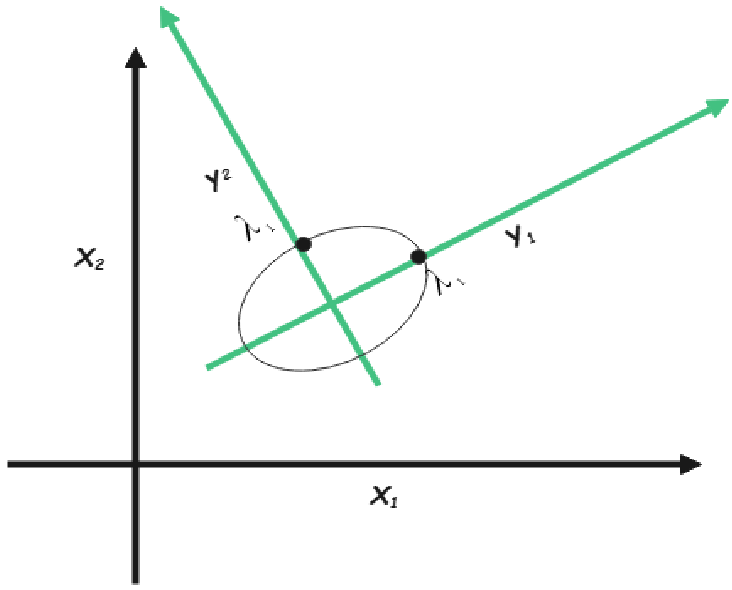
$$y_i = (x-\mu)^T u_i \tag{22}$$

$$Y = (y_1, y_2, ....., y_p)^T \tag{23}$$

所以:

$$\triangle = \sum_{i=1}^{p}\frac{y_i^2}{\lambda_i} \tag{24}$$

那么 $x$ 和 $Y$ 之间有什么关系呢？我们现在设 $x$ 为 $p$ 维空间的一组向量，那么这个 $Y$ 就是对 $x$ 做了一轮可逆线性变换得到的。假设 $x$ 是二维向量，并且规定 $\triangle = 1$，那么 $x$ 和 $y$ 之间的关系就如下图一样, 也就是当 $x$ 是一个二维向量时，固定 $f(x)$ 的值, 那么




也等价于固定 $\triangle$ 的值：

$$f(x) = Val \tag{25}$$

所有满足 $f(x) = Val$ 的二维向量 $x$ 都在一个椭圆上，并且这个椭圆的圆心在向量 $x$ 的坐标系中的坐标是均值向量 $\mu$。当 $p \geq 3$，$x$ 是更高维度的向量时，所有满足 $f(x) = Val$ 的二维向量 $x$ 都在一个超椭球面上。

### 高斯分布-边缘概率和条件概率

设 $x$ 是两个随机变量 $x_m, x_n$ 的联合随机分布：

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} x_m \\ x_n \end{pmatrix} \qquad \text{且满足:} \quad m + n = p \tag{26}$$

我们设:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mu_m \\ \mu_n \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix} = \begin{pmatrix} \Sigma_{mm} & \Sigma_{mn} \\ \Sigma_{nm} & \Sigma_{nn} \end{pmatrix} \tag{27}$$

**然后我们引出一个推论：**

已知随机变量 $X \sim\sim N(\mu,\Sigma)$，并且 $x \in R^p$，随机变量 $Y = AX + B$，其中：矩阵 $A_{q\times p}$，向量 $B_q$，那么 $Y$ 也服从高斯分布，且:

$$Y \sim\sim N(A\mu, A\Sigma A^T) \tag{28}$$

**首先求解 $p(x_m)$：** 我们设:

$$x_m = \begin{pmatrix} I_m & 0_n \end{pmatrix} \begin{pmatrix} x_m \\ x_n \end{pmatrix} \tag{29}$$

那么根据引出的推论:

$$A = \begin{pmatrix} I_m & 0_n \end{pmatrix} \qquad B = 0 \tag{30}$$

所以:

$$E[x_m] = \begin{pmatrix} I_m & 0_n \end{pmatrix} \begin{pmatrix} \mu_m \\ \mu_n \end{pmatrix} = \mu_m \tag{31}$$

$$Var[x_m] = \begin{pmatrix} I_m & 0_n \end{pmatrix} \begin{pmatrix} \Sigma_{mm} & \Sigma_{mn} \\ \Sigma_{nm} & \Sigma_{nn} \end{pmatrix} \begin{pmatrix} I_m \\ 0_n \end{pmatrix} = \Sigma_{mm} \tag{32}$$

所以我们知道 (同理可求 $x_n$)：

$$x_m \sim\sim N(\mu_m, \Sigma_{mm}) \tag{33}$$

**然后求解 $p(x_n|x_m)$：** 首先我们记:

$$x_{n.m} = x_n - \Sigma_{nm}\Sigma_{mm}^{-1}x_m = \begin{pmatrix} -\Sigma_{nm}\Sigma_{mm}^{-1} & I_n \end{pmatrix} \begin{pmatrix} x_m \\ x_n \end{pmatrix} \tag{34}$$

所以显然：

$$E[x_{n.m}] = \begin{pmatrix} -\Sigma_{nm}\Sigma_{mm}^{-1} & I_n \end{pmatrix} \begin{pmatrix} \mu_m \\ \mu_n \end{pmatrix} = \mu_n - \Sigma_{nm}\Sigma_{mm}^{-1}\mu_m \tag{35}$$

$$Var[x_{n.m}] = \begin{pmatrix} -\Sigma_{nm}\Sigma_{mm}^{-1} & I_n \end{pmatrix} \begin{pmatrix} \Sigma_{mm} & \Sigma_{mn} \\ \Sigma_{nm} & \Sigma_{nn} \end{pmatrix} \begin{pmatrix} -\Sigma_{mm}^{-1}\Sigma_{nm}^{T} \\ I_n \end{pmatrix} \tag{36}$$

$$= \Sigma_{nn} - \Sigma_{nm}\Sigma_{mm}^{-1}\Sigma_{mn} \tag{37}$$

所以我们知道:

$$x_{n.m} \sim\sim N(\mu_n - \Sigma_{nm}\Sigma_{mm}^{-1}\mu_m, \Sigma_{nn} - \Sigma_{nm}\Sigma_{mm}^{-1}\Sigma_{mn}) \tag{38}$$

可写成：

$$x_n = x_{n.m} + \Sigma_{nm}\Sigma_{mm}^{-1}x_m \tag{39}$$

我们设随机变量 $Z = x_n|x_m$, 其实这个时候 $x_m$ 是一个常量，那么本质上就是：

$$Z = x_{n.m} + \Sigma_{nm}\Sigma_{mm}^{-1}x_m \quad x_m \text{ 当做常量} \tag{40}$$

借助引出的推论可知:

$$E[Z] = E[x_{n.m}] + \Sigma_{nm}\Sigma_{mm}^{-1}x_m = \mu_n + \Sigma_{nm}\Sigma_{mm}^{-1}(x_m - u_m) \tag{41}$$

$$Var[Z] = Var[x_{n.m}] = \Sigma_{nn} - \Sigma_{nm}\Sigma_{mm}^{-1}\Sigma_{mn} \tag{42}$$

所以我们知道了 (同理可求 $x_m|x_n$)：

$$x_n|x_m \quad \sim\sim \quad N(\mu_n + \Sigma_{nm}\Sigma_{mm}^{-1}(x_m - u_m), \Sigma_{nn} - \Sigma_{nm}\Sigma_{mm}^{-1}\Sigma_{mn}) \tag{43}$$

### 高斯线性模型的求解

已知 $x \quad \sim\sim \quad N(\mu, \Lambda^{-1})$，且 $y|x \quad \sim\sim \quad N(Ax + b, L^{-1})$，求 $y$ 和 $x|y$ 的分布。
首先计算 $y$ 的分布，因为:
$y|x \quad \sim\sim \quad N(Ax + b, L^{-1})$，这句话的意思是当 $x$ 作为一个常量的时候，$y$ 服从一个均值为 $Ax + b$，方差为 $L^{-1}$ 的高斯分布，且 $y$ 与 $x$ 之间存在某种线性关系, 所以我们设:。

$$y = Ax + b + \epsilon \tag{44}$$

$$\text{且} \quad \epsilon \quad \text{满足:} \quad \epsilon \quad \sim\sim \quad N(0, L^{-1})$$
$$\text{并且} \quad \epsilon \quad \text{与} \quad x \quad \text{相互独立}$$

根据引出的推论我们可以知道:

$$\begin{aligned} E[y] &= E[Ax + b + \epsilon] = E[Ax + b] + E[\epsilon] \\ &= A\mu + b \\ Var[y] &= Var[Ax + b + \epsilon] = Var[Ax + b] + Var[\epsilon] \\ &= A\Lambda^{-1}A^T + L^{-1} \end{aligned}$$

所以：

$$y \quad \sim\sim \quad N(A\mu + b, A\Lambda^{-1}A^T + L^{-1}) \tag{45}$$

下面求解 $x|y$ 的分布, 首先记:

$$Z = \begin{pmatrix} y \\ x \end{pmatrix} \tag{46}$$

很显然 $Z$ 就是 $x, y$ 的联合分布，$Z$ 的均值和方差为:

$$Z \quad \sim\sim \quad N(\begin{bmatrix} A\mu + b \\ \mu \end{bmatrix}, \begin{bmatrix} A\Lambda^{-1}A^T + L^{-1} & Cov(y, x) \\ Cov(x, y) & \Lambda^{-1} \end{bmatrix}) \tag{47}$$

下面求解 $Cov(x,y)$:

$$
\begin{aligned}
Cov(x,y) &= E[(x-E[x])(y-E[y])^T] = E[(x-E[x])(Ax+b+\epsilon-E[Ax+b+\epsilon])^T] \\
&= E[(x-E[x])(Ax-A\mu+\epsilon)^T] = E[(x-\mu)(x-\mu)^T A^T] + E[(x-\mu)\epsilon^T]
\end{aligned}
$$

我们知道 $x$ 与 $\epsilon$ 相互独立，所以 $x-\mu$ 与 $\epsilon$ 相互独立，所以有:

$$
E[(x-\mu)\epsilon^T] = E[(x-\mu)]E[\epsilon^T] = 0 \tag{48}
$$

对于 $E[(x-\mu)(x-\mu)^T A^T]$, 有:

$$
E[(x-\mu)(x-\mu)^T A^T] = E[(x-\mu)(x-\mu)^T]A^T = Var(x)A^T = \Lambda^{-1}A^T \tag{49}
$$

所以有:

$$
Cov(x,y) = \Lambda^{-1}A^T \tag{50}
$$

并且：

$$
Cov(y,x) = Cov(x,y)^T = A\Lambda^{-1} \tag{51}
$$

由公式 (43) 我们可以知道:

$$
E[x|y] = \mu + \Lambda^{-1}A^T(A\Lambda^{-1}A^T + L^{-1})^{-1}(y - A\mu - b) \tag{52}
$$

$$
Var[x|y] = \Lambda^{-1} - \Lambda^{-1}A^T(L^{-1} + A\Lambda^{-1}A^T)^{-1}A\Lambda^{-1} \tag{53}
$$

所以:

$$
x|y \quad \sim\sim \quad N(E[x|y], Var[x|y]) \tag{54}
$$

# 矩阵求导

## 分子布局与分母布局:

第一种情况，设:

$$
y \in R \qquad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots x_m \end{pmatrix} \qquad x \in R_{m\times 1}
$$

也就是 $y$ 是一个标量，$x$ 是一个 m 维的列向量，求 $\dfrac{\mathrm{d}y}{\mathrm{d}x}$:

$$
= \begin{cases} \left[\dfrac{\mathrm{d}y}{\mathrm{d}x_1}, \dots, \dfrac{\mathrm{d}y}{\mathrm{d}x_m}\right] & \text{分子布局} \\[2em] \left[\dfrac{\mathrm{d}y}{\mathrm{d}x_1}, \dots, \dfrac{\mathrm{d}y}{\mathrm{d}x_m}\right]^T & \text{分母布局} \end{cases} \tag{55}
$$

第二种情况，设:

$$x \in R \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots y_m \end{pmatrix} \qquad y \in R_{m\times 1}$$

也就是 $x$ 是一个标量，$y$ 是一个 m 维的列向量，求 $\dfrac{\mathrm{d}y}{\mathrm{d}x}$:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \begin{cases} \left[\dfrac{\mathrm{d}y_1}{\mathrm{d}x}, \dots, \dfrac{\mathrm{d}y_m}{\mathrm{d}x}\right]^T & \text{分子布局} \\ \\ \left[\dfrac{\mathrm{d}y_1}{\mathrm{d}x}, \dots, \dfrac{\mathrm{d}y_m}{\mathrm{d}x}\right] & \text{分母布局} \end{cases} \tag{56}$$

就如上所述的那样，其实分子分母布局的区别就是：如果求导之后得到的向量的行数与分子的行数相等就是分子布局，如果和分母的行数相等就是分母布局。然后扩展到向量对向量求导中，设:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} \qquad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}_{m\times 1} \tag{57}$$

求 $\dfrac{\mathrm{d}y}{\mathrm{d}x}$:

$$\text{分子布局:} \qquad \frac{\mathrm{d}y}{\mathrm{d}x} = \begin{bmatrix} \dfrac{\mathrm{d}y_1}{\mathrm{d}x} \\ \\ \dfrac{\mathrm{d}y_2}{\mathrm{d}x} \\ \vdots \\ \dfrac{\mathrm{d}y_n}{\mathrm{d}x} \end{bmatrix} = \begin{bmatrix} \dfrac{\mathrm{d}y_1}{\mathrm{d}x_1} & \dfrac{\mathrm{d}y_1}{\mathrm{d}x_2} & \cdots & \dfrac{\mathrm{d}y_1}{\mathrm{d}x_m} \\ \\ \dfrac{\mathrm{d}y_2}{\mathrm{d}x_1} & \dfrac{\mathrm{d}y_2}{\mathrm{d}x_2} & \cdots & \dfrac{\mathrm{d}y_2}{\mathrm{d}x_m} \\ \\ \vdots & \vdots & \ddots & \vdots \\ \\ \dfrac{\mathrm{d}y_n}{\mathrm{d}x_1} & \dfrac{\mathrm{d}y_n}{\mathrm{d}x_2} & \cdots & \dfrac{\mathrm{d}y_n}{\mathrm{d}x_m} \end{bmatrix}_{n\times m} \tag{58}$$

向量对向量求导得到的结果是一个矩阵，分子布局就是求导得到的矩阵行数和分子一

样，列拉伸到与分母一样。

$$\text{分母布局:}\qquad \frac{\mathrm{d}y}{\mathrm{d}x}=\begin{bmatrix}\frac{\mathrm{d}y}{\mathrm{d}x_1}\\ \frac{\mathrm{d}y}{\mathrm{d}x_2}\\ \vdots\\ \frac{\mathrm{d}y}{\mathrm{d}x_m}\end{bmatrix}=\begin{bmatrix}\frac{\mathrm{d}y_1}{\mathrm{d}x_1} & \frac{\mathrm{d}y_2}{\mathrm{d}x_1} & \cdots & \frac{\mathrm{d}y_n}{\mathrm{d}x_1}\\ \frac{\mathrm{d}y_1}{\mathrm{d}x_2} & \frac{\mathrm{d}y_2}{\mathrm{d}x_2} & \cdots & \frac{\mathrm{d}y_n}{\mathrm{d}x_2}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\mathrm{d}y_1}{\mathrm{d}x_m} & \frac{\mathrm{d}y_2}{\mathrm{d}x_m} & \cdots & \frac{\mathrm{d}y_m}{\mathrm{d}x_m}\end{bmatrix}_{m\times n} \tag{59}$$

分子布局就是求导得到的矩阵行数和分母一样，行拉伸到与分母一样。

**下面的讨论全部基于分母布局:**

**二阶导数:**

$$\text{设:}\quad x=\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_m\end{bmatrix}_{m\times 1}\qquad ,\quad f(x)\in R\qquad \text{求:}\quad \frac{\mathrm{d}^2f(x)}{\mathrm{d}x^2}$$

解，记:

$$g=\frac{\mathrm{d}f(x)}{\mathrm{d}x}=\begin{bmatrix}\frac{\mathrm{d}f(x)}{\mathrm{d}x_1}\\ \frac{\mathrm{d}f(x)}{\mathrm{d}x_2}\\ \vdots\\ \frac{\mathrm{d}f(x)}{\mathrm{d}x_m}\end{bmatrix}$$

所以:

$$\frac{\mathrm{d}^2f(x)}{\mathrm{d}x^2}=\frac{\mathrm{d}g}{\mathrm{d}x}=\begin{bmatrix}\frac{\mathrm{d}g}{\mathrm{d}x_1}\\ \frac{\mathrm{d}g}{\mathrm{d}x_2}\\ \vdots\\ \frac{\mathrm{d}g}{\mathrm{d}x_m}\end{bmatrix}=\begin{bmatrix}\frac{\partial^2f(x)}{\partial x_1\partial x_1} & \frac{\partial^2f(x)}{\partial x_2\partial x_1} & \cdots & \frac{\partial^2f(x)}{\partial x_m\partial x_1}\\ \frac{\partial^2f(x)}{\partial x_1\partial x_2} & \frac{\partial^2f(x)}{\partial x_2\partial x_2} & \cdots & \frac{\partial^2f(x)}{\partial x_m\partial x_2}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial^2f(x)}{\partial x_1\partial x_m} & \frac{\partial^2f(x)}{\partial x_2\partial x_m} & \cdots & \frac{\partial^2f(x)}{\partial x_m\partial x_m}\end{bmatrix} \tag{60}$$

上述二次求导过程和高数里面学的求导过程类似，只不过利用到了布局知识。

**复核求导**

加法:

$$x=\begin{bmatrix}x_1\\x_2\\\vdots\\x_m\end{bmatrix}\in R^{m\times1}\quad y=f(x)=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}\in R^{n\times1}\quad z=g(x)=\begin{bmatrix}z_1\\z_2\\\vdots\\z_n\end{bmatrix}\in R^{n\times1}\tag{61}$$

求 $\dfrac{\mathrm{d}(y+z)}{\mathrm{d}x}$:

$$\frac{\mathrm{d}(y+z)}{\mathrm{d}x}=\frac{\mathrm{d}y}{\mathrm{d}x}+\frac{\mathrm{d}z}{\mathrm{d}x}=\begin{bmatrix}\dfrac{\mathrm{d}y}{\mathrm{d}x_1}\\\dfrac{\mathrm{d}y}{\mathrm{d}x_2}\\\vdots\\\dfrac{\mathrm{d}y}{\mathrm{d}x_m}\end{bmatrix}_{m\times1}+\begin{bmatrix}\dfrac{\mathrm{d}z}{\mathrm{d}x_1}\\\dfrac{\mathrm{d}z}{\mathrm{d}x_2}\\\vdots\\\dfrac{\mathrm{d}z}{\mathrm{d}x_m}\end{bmatrix}_{m\times1}$$

$$=\begin{bmatrix}\dfrac{\mathrm{d}y_1}{\mathrm{d}x_1}&\dfrac{\mathrm{d}y_2}{\mathrm{d}x_1}&\cdots&\dfrac{\mathrm{d}y_n}{\mathrm{d}x_1}\\\dfrac{\mathrm{d}y_1}{\mathrm{d}x_2}&\dfrac{\mathrm{d}y_2}{\mathrm{d}x_2}&\cdots&\dfrac{\mathrm{d}y_n}{\mathrm{d}x_2}\\\vdots&\vdots&\ddots&\vdots\\\dfrac{\mathrm{d}y_1}{\mathrm{d}x_m}&\dfrac{\mathrm{d}y_2}{\mathrm{d}x_m}&\cdots&\dfrac{\mathrm{d}y_n}{\mathrm{d}x_m}\end{bmatrix}_{m\times n}+\begin{bmatrix}\dfrac{\mathrm{d}z_1}{\mathrm{d}x_1}&\dfrac{\mathrm{d}z_2}{\mathrm{d}x_1}&\cdots&\dfrac{\mathrm{d}z_n}{\mathrm{d}x_1}\\\dfrac{\mathrm{d}z_1}{\mathrm{d}x_2}&\dfrac{\mathrm{d}z_2}{\mathrm{d}x_2}&\cdots&\dfrac{\mathrm{d}z_n}{\mathrm{d}x_2}\\\vdots&\vdots&\ddots&\vdots\\\dfrac{\mathrm{d}z_1}{\mathrm{d}x_m}&\dfrac{\mathrm{d}z_2}{\mathrm{d}x_m}&\cdots&\dfrac{\mathrm{d}z_n}{\mathrm{d}x_m}\end{bmatrix}_{m\times n}$$

乘法: 设:

$$x=\begin{bmatrix}x_1\\x_2\\\vdots\\x_m\end{bmatrix}_{m\times1}\qquad y=f(x)=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}_{n\times1}\qquad z=g(x)=\begin{bmatrix}z_1\\z_2\\\vdots\\z_n\end{bmatrix}_{n\times1}$$

求 $\dfrac{\mathrm{d}y^Tz}{\mathrm{d}x}$:

首先我们观察到 $y^Tz$ 是标量，那么显然 $\dfrac{\mathrm{d}y^Tz}{\mathrm{d}x} \in R^{n\times 1}$：

$$\text{所以:}\quad \left.\frac{\mathrm{d}y^Tz}{\mathrm{d}x}\right|_{m\times 1} = \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{m\times n} \cdot \left.\frac{\mathrm{d}y^Tz}{\mathrm{d}y}\right|_{n\times 1} + \left.\frac{\mathrm{d}z}{\mathrm{d}x}\right|_{m\times n} \cdot \left.\frac{\mathrm{d}y^Tz}{\mathrm{d}z}\right|_{n\times 1} \tag{62}$$

下面计算 $\left.\dfrac{\mathrm{d}y^Tz}{\mathrm{d}y}\right|_{n\times 1}$，$\left.\dfrac{\mathrm{d}y^Tz}{\mathrm{d}z}\right|_{n\times 1}$：

$$\left.\frac{\mathrm{d}y^Tz}{\mathrm{d}y}\right|_{n\times 1} = \begin{bmatrix} \dfrac{\partial \sum_{i=0}^{n} y_i z_i}{\partial z_1} \\ \dfrac{\partial \sum_{i=0}^{n} y_i z_i}{\partial z_2} \\ \vdots \\ \dfrac{\partial \sum_{i=0}^{n} y_i z_i}{\partial z_n} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = y \qquad \left.\frac{\mathrm{d}y^Tz}{\mathrm{d}z}\right|_{n\times 1} = \begin{bmatrix} \dfrac{\partial \sum_{i=0}^{n} y_i z_i}{\partial y_1} \\ \dfrac{\partial \sum_{i=0}^{n} y_i z_i}{\partial y_2} \\ \vdots \\ \dfrac{\partial \sum_{i=0}^{n} y_i z_i}{\partial y_n} \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = z \tag{63}$$

所有有:

$$\left.\frac{\mathrm{d}y^Tz}{\mathrm{d}x}\right|_{m\times 1} = \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{m\times n} \cdot z \quad + \quad \left.\frac{\mathrm{d}z}{\mathrm{d}x}\right|_{m\times n} \cdot y \tag{64}$$

### 链式法则

**设：**

$$x \in R \qquad y = f(x) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \qquad z = g(y) = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$$

求证：

$$\left.\frac{\mathrm{d}z}{\mathrm{d}x}\right|_{1\times n} = \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{1\times m} \cdot \left.\frac{\mathrm{d}z}{\mathrm{d}y}\right|_{m\times n} \tag{65}$$

解, 因为:

$$\left.\frac{\mathrm{d}z}{\mathrm{d}x}\right|_{1\times n} = \begin{bmatrix} \left.\dfrac{\mathrm{d}z_1}{\mathrm{d}x}\right|_{1\times 1} & \left.\dfrac{\mathrm{d}z_2}{\mathrm{d}x}\right|_{1\times 1} & \dots & \left.\dfrac{\mathrm{d}z_n}{\mathrm{d}x}\right|_{1\times 1} \end{bmatrix} \tag{66}$$

又因为:

$$\left.\frac{\mathrm{d}z_i}{\mathrm{d}x}\right|_{1\times 1} = \left.\frac{\partial y}{\partial x}\right|_{1\times m} \cdot \left.\frac{\partial z_i}{\partial y}\right|_{m\times 1} \qquad i = 1, 2, 3, ...., n \tag{67}$$

所以:

$$\left.\frac{\mathrm{d}z}{\mathrm{d}x}\right|_{1\times n}=\begin{bmatrix}\left.\frac{\partial y}{\partial x}\right|_{1\times m}\cdot\left.\frac{\partial z_1}{\partial y}\right|_{m\times 1} & \left.\frac{\partial y}{\partial x}\right|_{1\times m}\cdot\left.\frac{\partial z_2}{\partial y}\right|_{m\times 1} & \cdots & \left.\frac{\partial y}{\partial x}\right|_{1\times m}\cdot\left.\frac{\partial z_n}{\partial y}\right|_{m\times 1}\end{bmatrix}$$

$$=\left.\frac{\partial y}{\partial x}\right|_{1\times m}\cdot\begin{bmatrix}\left.\frac{\partial z_1}{\partial y}\right|_{m\times 1} & \left.\frac{\partial z_2}{\partial y}\right|_{m\times 1} & \cdots & \left.\frac{\partial z_n}{\partial y}\right|_{m\times 1}\end{bmatrix}$$

$$=\left.\frac{\partial y}{\partial x}\right|_{1\times m}\cdot\begin{bmatrix}\frac{\partial z_1}{\partial y_1} & \frac{\partial z_2}{\partial y_1} & \cdots & \frac{\partial z_n}{\partial y_1}\\ \frac{\partial z_1}{\partial y_2} & \frac{\partial z_2}{\partial y_2} & \cdots & \frac{\partial z_n}{\partial y_2}\\ \vdots & \vdots\ldots & \vdots & \\ \frac{\partial z_1}{\partial y_m} & \frac{\partial z_2}{\partial y_m} & \cdots & \frac{\partial z_n}{\partial y_m}\end{bmatrix}_{m\times n}$$

$$=\left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{1\times m}\cdot\left.\frac{\mathrm{d}z}{\mathrm{d}y}\right|_{m\times n}$$

**设：**

$$x=\begin{bmatrix}x_1\\x_2\\\vdots\\x_m\end{bmatrix}_{m\times 1}\qquad y=f(x)=\begin{bmatrix}y_1\\y_2\\\vdots\\y_k\end{bmatrix}_{k\times 1}\qquad z=g(y)=\begin{bmatrix}z_1\\z_2\\\vdots\\z_n\end{bmatrix}_{n\times 1}$$

求证：

$$\left.\frac{\mathrm{d}z}{\mathrm{d}x}\right|_{m\times n}=\left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{m\times k}\cdot\left.\frac{\mathrm{d}z}{\mathrm{d}y}\right|_{k\times n}\tag{68}$$

解, 因为:

$$\left.\frac{\mathrm{d}z}{\mathrm{d}x}\right|_{m\times n}=\begin{bmatrix}\frac{\partial z}{\partial x_1}\\ \frac{\partial z}{\partial x_2}\\ \vdots\\ \frac{\partial z}{\partial x_m}\end{bmatrix}_{m\times n}\tag{69}$$

因为 (65):

所以有:

$$\frac{\partial z}{\partial x_i}=\frac{\partial y}{\partial x_i}\cdot\frac{\partial z}{\partial y} \tag{70}$$

所以:

$$\left.\frac{\mathrm{d}z}{\mathrm{d}x}\right|_{m\times n}=\begin{bmatrix}\frac{\partial y}{\partial x_1}\cdot\frac{\partial z}{\partial y}\\ \frac{\partial y}{\partial x_2}\cdot\frac{\partial z}{\partial y}\\ \vdots\\ \frac{\partial y}{\partial x_m}\cdot\frac{\partial z}{\partial y}\end{bmatrix}_{m\times n}=\begin{bmatrix}\frac{\partial y}{\partial x_1}\\ \frac{\partial y}{\partial x_2}\\ \vdots\\ \frac{\partial y}{\partial x_m}\end{bmatrix}_{m\times k}\cdot\left.\frac{\partial z}{\partial y}\right|_{k\times n}=\left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{m\times k}\cdot\left.\frac{\mathrm{d}z}{\mathrm{d}y}\right|_{k\times n}$$

**同理可得:**

$$\left.\frac{\mathrm{d}z}{\mathrm{d}x^T}\right|_{n\times m}=\left.\frac{\mathrm{d}z}{\mathrm{d}y^T}\right|_{n\times k}\cdot\left.\frac{\mathrm{d}y}{\mathrm{d}x^T}\right|_{k\times m} \tag{71}$$

**设:**

$$X=\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_m\end{bmatrix}_{m\times 1}\qquad Y=f(X)=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}_{n\times 1}\qquad Z=G(Y)\in R$$

求证:

$$\left.\frac{\partial Z}{\partial X}\right|_{m\times 1}=\left.\frac{\partial Y^T}{\partial X}\right|_{m\times n}\cdot\left.\frac{\partial Z}{\partial Y}\right|_{n\times 1} \tag{72}$$

因为:

$$\left.\frac{\partial Z}{\partial X}\right|_{m\times 1}=\begin{bmatrix}\left.\frac{\partial Z}{\partial x_1}\right|_{1\times 1}\\ \left.\frac{\partial Z}{\partial x_2}\right|_{1\times 1}\\ \vdots\\ \left.\frac{\partial Z}{\partial x_m}\right|_{1\times 1}\end{bmatrix}_{m\times 1} \tag{73}$$

又因为:

$$\left.\frac{\partial Z}{\partial x_j}\right|_{1\times1}=\sum_{i=1}^{n}\left.\frac{\partial Z}{\partial y_i}\right|_{1\times1}\cdot\left.\frac{\partial y_i}{\partial x_j}\right|_{1\times1}$$

$$=\begin{bmatrix}\left.\frac{\partial y_1}{\partial x_j}\right|_{1\times1} & \left.\frac{\partial y_2}{\partial x_j}\right|_{1\times1} & \cdots & \left.\frac{\partial y_n}{\partial x_j}\right|_{1\times1}\end{bmatrix}_{1\times n}\cdot\begin{bmatrix}\left.\frac{\partial z}{\partial y_1}\right|_{1\times1}\\ \left.\frac{\partial z}{\partial y_2}\right|_{1\times1}\\ \vdots\\ \left.\frac{\partial z}{\partial y_n}\right|_{1\times1}\end{bmatrix}_{n\times1}$$

$$=\left.\frac{\partial Y^T}{\partial x_j}\right|_{1\times n}\cdot\left.\frac{\partial Z}{\partial Y}\right|_{n\times1}$$

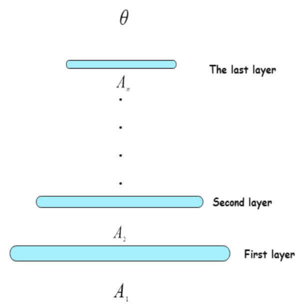
所以有：

$$\left.\frac{\partial Z}{\partial X}\right|_{m\times1}=\begin{bmatrix}\left.\frac{\partial Y^T}{\partial x_1}\right|_{1\times n}\cdot\left.\frac{\partial Z}{\partial Y}\right|_{n\times1}\\ \left.\frac{\partial Y^T}{\partial x_2}\right|_{1\times n}\cdot\left.\frac{\partial Z}{\partial Y}\right|_{n\times1}\\ \vdots\\ \left.\frac{\partial Y^T}{\partial x_m}\right|_{1\times n}\cdot\left.\frac{\partial Z}{\partial Y}\right|_{n\times1}\end{bmatrix}_{m\times1}=\begin{bmatrix}\left.\frac{\partial Y^T}{\partial x_1}\right|_{1\times n}\\ \left.\frac{\partial Y^T}{\partial x_2}\right|_{1\times n}\\ \vdots\\ \left.\frac{\partial Y^T}{\partial x_m}\right|_{1\times n}\end{bmatrix}_{m\times n}\cdot\left.\frac{\partial Z}{\partial Y}\right|_{n\times1}=\left.\frac{\partial Y^T}{\partial X}\right|_{m\times n}\cdot\left.\frac{\partial Z}{\partial Y}\right|_{n\times1}$$

**神经网络的反向传播:**

设在神经网络中有 $n$ 层，第 $i$ 层的输入为 $A_i$，且 $A_i$ 为列向量，最后的输出结果为 $\theta$, 且 $\theta\in R$:

求 $\frac{\mathrm{d}\theta}{\mathrm{d}A_1}$:




由 (72) 可以知道:

$$\frac{\mathrm{d}\theta}{\mathrm{d}A_1} = \frac{\mathrm{d}A_n^T}{\mathrm{d}A_1} \cdot \frac{\mathrm{d}\theta}{\mathrm{d}A_n} = (\frac{\mathrm{d}A_n}{\mathrm{d}A_1^T})^T \cdot \frac{\mathrm{d}\theta}{\mathrm{d}A_n} \tag{74}$$

然后继续把 $\frac{\mathrm{d}A_n}{\mathrm{d}A_1^T}$ 按照 (71) 展开:

$$\begin{aligned}
\frac{\mathrm{d}A_n}{\mathrm{d}A_1^T} &= \frac{\mathrm{d}A_n}{\mathrm{d}A_{n-1}^T} \cdot \frac{\mathrm{d}A_{n-1}}{\mathrm{d}A_1^T} \\
&= \frac{\mathrm{d}A_n}{\mathrm{d}A_{n-1}^T} \cdot \frac{\mathrm{d}A_{n-1}}{\mathrm{d}A_{n-2}^T} \cdot \frac{\mathrm{d}A_{n-2}}{\mathrm{d}A_1^T} \\
&= \frac{\mathrm{d}A_n}{\mathrm{d}A_{n-1}^T} \cdot \frac{\mathrm{d}A_{n-1}}{\mathrm{d}A_{n-2}^T} \cdot \frac{\mathrm{d}A_{n-2}}{\mathrm{d}A_{n-3}^T} \cdot \frac{\mathrm{d}A_{n-3}}{\mathrm{d}A_1^T} \\
&= \frac{\mathrm{d}A_n}{\mathrm{d}A_{n-1}^T} \cdot \frac{\mathrm{d}A_{n-1}}{\mathrm{d}A_{n-2}^T} \cdot \frac{\mathrm{d}A_{n-2}}{\mathrm{d}A_{n-3}^T} \cdot \frac{\mathrm{d}A_{n-3}}{\mathrm{d}A_{n-4}^T} \cdot \frac{\mathrm{d}A_{n-4}}{\mathrm{d}A_1^T} \\
&= \frac{\mathrm{d}A_n}{\mathrm{d}A_{n-1}^T} \cdot \frac{\mathrm{d}A_{n-1}}{\mathrm{d}A_{n-2}^T} \cdot \frac{\mathrm{d}A_{n-2}}{\mathrm{d}A_{n-3}^T} \cdot \frac{\mathrm{d}A_{n-3}}{\mathrm{d}A_{n-4}^T} \cdot \frac{\mathrm{d}A_{n-4}}{\mathrm{d}A_{n-5}^T} \cdots\cdots \frac{\mathrm{d}A_3}{\mathrm{d}A_2^T} \cdot \frac{\mathrm{d}A_2}{\mathrm{d}A_1^T}
\end{aligned}$$

所以有:

$$\frac{\mathrm{d}\theta}{\mathrm{d}A_1} = (\frac{\mathrm{d}A_n}{\mathrm{d}A_{n-1}^T} \cdot \frac{\mathrm{d}A_{n-1}}{\mathrm{d}A_{n-2}^T} \cdot \frac{\mathrm{d}A_{n-2}}{\mathrm{d}A_{n-3}^T} \cdot \frac{\mathrm{d}A_{n-3}}{\mathrm{d}A_{n-4}^T} \cdot \frac{\mathrm{d}A_{n-4}}{\mathrm{d}A_{n-5}^T} \cdots\cdots \frac{\mathrm{d}A_3}{\mathrm{d}A_2^T} \cdot \frac{\mathrm{d}A_2}{\mathrm{d}A_1^T})^T \cdot \frac{\mathrm{d}\theta}{\mathrm{d}A_n} \tag{75}$$

**几种常用的推论:**

**推论一：**设:

$$A = [a_1, a_2, \ldots, a_m]_{1\times m} \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots x_m \end{bmatrix}_{m\times 1} \tag{76}$$

求 $\dfrac{\mathrm{d}(AX)}{\mathrm{d}X}$，$\dfrac{\mathrm{d}(AX)}{\mathrm{d}X^T}$:

$$\frac{\mathrm{d}(AX)}{\mathrm{d}X}=\begin{bmatrix}\dfrac{\partial(AX)}{\partial x_1}\\ \dfrac{\partial(AX)}{\partial x_2}\\ \vdots\\ \dfrac{\partial(AX)}{\partial x_m}\end{bmatrix}_{m\times 1}=\begin{bmatrix}\dfrac{\partial\sum_{i=1}^m x_i a_i}{\partial x_1}\\ \dfrac{\partial\sum_{i=1}^m x_i a_i}{\partial x_2}\\ \vdots\\ \dfrac{\partial\sum_{i=1}^m x_i a_i}{\partial x_m}\end{bmatrix}_{m\times 1}=\begin{bmatrix}a_1\\ a_2\\ \vdots\\ a_m\end{bmatrix}_{m\times 1}=A \tag{77}$$

$$\frac{\mathrm{d}(AX)}{\mathrm{d}X^T}=\begin{bmatrix}\dfrac{\partial(AX)}{\partial x_1} & \dfrac{\partial(AX)}{\partial x_2} & \dots & \dfrac{\partial(AX)}{\partial x_m}\end{bmatrix}_{1\times m}=\begin{bmatrix}a_1 & a_2 & \dots & a_m\end{bmatrix}=A^T \tag{78}$$

**推论二：**设:

$$A=\begin{bmatrix}a_{11} & a_{12} & \dots & a_{1m}\\ a_{21} & a_{22} & \dots & a_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ a_{n1} & a_{n2} & \dots & a_{nm}\end{bmatrix}_{n\times m}\qquad X=\begin{bmatrix}x_1\\ x_2\\ \vdots x_m\end{bmatrix}_{m\times 1} \tag{79}$$

求 $\dfrac{\mathrm{d}(AX)}{\mathrm{d}X}$，$\dfrac{\mathrm{d}(AX)}{\mathrm{d}X^T}$: 记:

$$Y=AX=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}\qquad \text{且:}\qquad y_i=\sum_{k=1}^m a_{ik}x_k$$

$$\frac{\mathrm{d}(AX)}{\mathrm{d}X} = \begin{bmatrix} \frac{\partial Y}{\partial x_1} \\ \frac{\partial Y}{\partial x_2} \\ \vdots \\ \frac{\partial Y}{\partial x_m} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_m} & \frac{\partial y_2}{\partial x_m} & \cdots & \frac{\partial y_n}{\partial x_m} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial \sum_{k=1}^m a_{1k}x_k}{\partial x_1} & \frac{\partial \sum_{k=1}^m a_{2k}x_k}{\partial x_1} & \cdots & \frac{\partial \sum_{k=1}^m a_{nk}x_k}{\partial x_1} \\ \frac{\partial \sum_{k=1}^m a_{1k}x_k}{\partial x_2} & \frac{\partial \sum_{k=1}^m a_{2k}x_k}{\partial x_2} & \cdots & \frac{\partial \sum_{k=1}^m a_{nk}x_k}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \sum_{k=1}^m a_{1k}x_k}{\partial x_m} & \frac{\partial \sum_{k=1}^m a_{2k}x_k}{\partial x_m} & \cdots & \frac{\partial \sum_{k=1}^m a_{nk}x_k}{\partial x_m} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} & a_{21} & \ldots & a_{n1} \\ a_{12} & a_{22} & \ldots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \ldots & a_{nm} \end{bmatrix} = A^T$$

同理可得:

$$\frac{\mathrm{d}(AX)}{\mathrm{d}X^T} = A \tag{80}$$

**设:**

$$X = [x_1, x_2, x_3, ......, x_n]_{n\times 1}^T$$

求 $\frac{\mathrm{d}\|X\|}{\mathrm{d}X}$，$\frac{\mathrm{d}\|X\|}{\mathrm{d}X^T}$:

$$\frac{\mathrm{d}\|X\|}{\mathrm{d}X} = \begin{bmatrix} \frac{\mathrm{d}\|X\|}{\mathrm{d}x_1} \\ \frac{\mathrm{d}\|X\|}{\mathrm{d}x_2} \\ \vdots \\ \frac{\mathrm{d}\|X\|}{\mathrm{d}x_n} \end{bmatrix} = \begin{bmatrix} \frac{\mathrm{d}\sum_{i=1}^{n} x_i^2}{\mathrm{d}x_1} \\ \frac{\mathrm{d}\sum_{i=1}^{n} x_i^2}{\mathrm{d}x_2} \\ \vdots \\ \frac{\mathrm{d}\sum_{i=1}^{n} x_i^2}{\mathrm{d}x_n} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{bmatrix} = 2X \tag{81}$$

$$\frac{\mathrm{d}\|X\|}{\mathrm{d}X^T} = \begin{bmatrix} \frac{\mathrm{d}\|X\|}{\mathrm{d}x_1} \\ \frac{\mathrm{d}\|X\|}{\mathrm{d}x_2} \\ \vdots \\ \frac{\mathrm{d}\|X\|}{\mathrm{d}x_n} \end{bmatrix}^T = \begin{bmatrix} \frac{\mathrm{d}\sum_{i=1}^{n} x_i^2}{\mathrm{d}x_1} \\ \frac{\mathrm{d}\sum_{i=1}^{n} x_i^2}{\mathrm{d}x_2} \\ \vdots \\ \frac{\mathrm{d}\sum_{i=1}^{n} x_i^2}{\mathrm{d}x_n} \end{bmatrix}^T = \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{bmatrix}^T = 2X^T \tag{82}$$