知识补充

0.1 矩阵求导

标量 f 对矩阵 X 的导数，定义为 $\frac{\partial f}{\partial X}=\left[\frac{\partial f}{\partial X_{ij}}\right]$，即 f 对 X 逐元素求导排成与 X 尺寸相同的矩阵。然而，这个定义在计算中并不好用，实用上的原因是在对较复杂的函数难以逐元素求导。

一元微积分中的导数（标量对标量的导数）与微分有联系：$df=f'(x)dx$；多元微积分中的梯度（标量对向量的导数）也与微分有联系：$df=\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}dx_i=\frac{\partial f}{\partial \boldsymbol{x}}^T d\boldsymbol{x}$，这里第一个等号是全微分公式，第二个等号表达了梯度与微分的联系：全微分 df 是 $n\times 1$ 梯度向量 $\frac{\partial f}{\partial \boldsymbol{x}}$ 与 $n\times 1$ 微分向量 $d\boldsymbol{x}$ 的内积；受此启发，我们将矩阵导数与微分建立联系：

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

其中 tr 代表迹 (trace) 是方阵对角线元素之和，满足性质：对尺寸相同的矩阵 A,B，$\operatorname{tr}(A^TB)=\sum_{i,j}A_{ij}B_{ij}$，即 $\operatorname{tr}(A^TB)$ 是矩阵 A,B 的内积。与梯度相似，这里第一个等号是全微分公式，第二个等号表达了矩阵导数与微分的联系：全微分 df 是 $m\times n$ 导数 $\frac{\partial f}{\partial X}$ 与 $m\times n$ 微分矩阵 dX 的内积。

矩阵微分的运算法则：

1. 加减法：$d(X\pm Y)=dX\pm dY$；
2. 矩阵乘法：$d(XY)=dXY+XdY$；
3. 转置：$d(X^T)=(dX)^T$；
4. 迹：$d\operatorname{tr}(X)=\operatorname{tr}(dX)$；
5. 逆：$dX^{-1}=-X^{-1}dXX^{-1}$。此式可在 $XX^{-1}=I$ 两侧求微分来证明。
6. 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 $d|X|=|X|\operatorname{tr}(X^{-1}dX)$。
7. 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。
8. 逐元素函数：$d\sigma(X)=\sigma'(X)\odot dX$ $\sigma(X)=[\sigma(X_{ij})]$ 是逐元素运算的标量函数。

迹运算法则：

1. 标量套上迹：$a=\operatorname{tr}(a)$
2. 转置：$\operatorname{tr}(A^T)=\operatorname{tr}(A)$
3. 线性：$\operatorname{tr}(A\pm B)=\operatorname{tr}(A)\pm\operatorname{tr}(B)$
4. 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 $\sum_{i,j}A_{ij}B_{ji}$。
5. 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 $\sum_{i,j}A_{ij}B_{ij}C_{ij}$。

矩阵求导法则：**若标量函数 f 是矩阵 X 经加减乘法、行列式、逆、逐元素函数等运算构成，则使用相应的运算法则对 f 求微分，再使用迹运算法则给 df 套上迹并将其它项交换至 dX 左侧，即能得到导数。**

复合法则：假设已求得 $\frac{\partial f}{\partial Y}$，而 Y 是 X 的函数，如何求 $\frac{\partial f}{\partial X}$ 呢? 在微积分中有标量求导的链式法则 $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y}\frac{\partial y}{\partial x}$，但这里不能沿用链式法则，因为矩阵对矩阵的导数 $\frac{\partial Y}{\partial X}$ 截至目前仍是未定义的。直接从微分入手建立复合法则：先写出 $df = \mathrm{tr}\left(\frac{\partial f}{\partial Y}^T dY\right)$，再将 dY 用 dX 表示出来代入，并使用迹运算法则将其他项交换至 dX 左侧，即可得到 $\frac{\partial f}{\partial X}$。

例 1：$f = \boldsymbol{a}^T X \boldsymbol{b}$ $\frac{\partial f}{\partial X}$。其中 $\boldsymbol{a}$ 是 $m\times 1$ 列向量，X 是 $m \times n$ 矩阵，$\boldsymbol{b}$ 是 $n \times 1$ 列向量，f 是标量。
解：先使用矩阵乘法法则求微分：$df = \boldsymbol{a}^T dX \boldsymbol{b}$，再套上迹并做矩阵乘法交换：

$$\begin{aligned} df &= \mathrm{tr}(\boldsymbol{a}^T dX \boldsymbol{b}) \qquad (df\ is\ a\ scalar) \\ &= \mathrm{tr}(\boldsymbol{b}\boldsymbol{a}^T dX) \qquad (\mathrm{tr}(AB)\ =\ \mathrm{tr}(BA)) \end{aligned}$$

对照导数与微分的联系 $df = \mathrm{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$，得到 $\frac{\partial f}{\partial X} = (\boldsymbol{b}\boldsymbol{a}^T)^T = \boldsymbol{a}\boldsymbol{b}^T$。

例 2【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：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：

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

对照导数与微分的联系 $dl = \frac{\partial l}{\partial \boldsymbol{w}}^T d\boldsymbol{w}$，得到

$$\frac{\partial l}{\partial \boldsymbol{w}} = (2(X\boldsymbol{w} - \boldsymbol{y})^T X)^T = 2X^T (X\boldsymbol{w} - \boldsymbol{y})$$

$\frac{\partial l}{\partial \boldsymbol{w}}$ 的零点即 $\boldsymbol{w}$ 的最小二乘估计为 $\boldsymbol{w} = (X^T X)^{-1} X^T \boldsymbol{y}$。

0.2 参数估计

总体的均值（mean）常用 μ 表示。方差用 σ^2 表示。样本的均值用 $\bar{X}$ 表示，方差用 S^2 表示。(S^2 通常用 $n-1$ 为底。这样是想让结果跟接近总体的方差，又称为无偏估计。)

$$\begin{aligned} E(x) = \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ D(x) &= \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2 \\ E(x^2) &= E(x)^2 + D(x) \end{aligned}$$

0.2.1 矩估计

样本 k 阶矩：$a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$

总体 k 阶矩：$\alpha_k = EX^k, \mu_k = E(X - EX)^k$

在 k 阶矩存在的情况下，根据大数定律有：

$$a_k \xrightarrow{p} \alpha_k, \qquad m_k \xrightarrow{p} \mu_k$$

于是让总体的原点矩与样本的原点矩相等，解出参数，所得结果即为参数的矩估计值。

矩估计方法应用的原则是：能用低阶矩处理的就不用高阶矩。矩估计法的优点是简单易行, 有些情况下不需要事先知道总体是什么分布. 缺点是，当总体类型已知时，没有充分利用分布提供的信息. 一般场合下, 矩估计量不具有唯一性.

0.2.2 最大似然估计

设样本 $X=(X_1,\cdots,X_n)$ 有概率函数

$$f(x;\theta)=f(x;\theta_1,\cdots,\theta_k)$$

这里参数 $\theta=(\theta_1,\cdots,\theta_k)\in\Theta, x=(x_1,\cdots,x_n)$ 为样本 X 的观察值. 当固定 x 时把 $f(x;\theta)$ 看成为 θ 的函数，称为似然函数, 常记为 $L(x;\theta)$ 或 $L(\theta)$.

当把参数 θ 看成变动时，也就得到"在不同的 θ 值下能观察到 x 的可能性大小, 即 $L(x;\theta)$"。使得 $L(x;\theta)$ 最大的 θ 值称为 θ 最大似然估计值。

设 $X=(X_1,\cdots,X_n)$ 为从具有概率函数 $f_\theta(x)$ 的总体中抽取的样本，θ 为未知参数或者参数向量. $x=(x_1,\cdots,x_n)$ 为样本的观察值。若在给定 x 时, 值 $\hat{\theta}=\hat{\theta}(x)$ 满足下式

$$L(\hat{\theta})=\max_{\theta\in\Theta}\{L(x;\theta)\}$$

则称 $\hat{\theta}$ 为参数 θ 的最大似然估计值, 而 $\hat{\theta}(X)$ 称为参数 θ 的最大似然估计量。

求最大似然估计值相当于求似然函数的最大值。在简单样本的情况下,

$$L(\hat{\theta})=\prod_{i=1}^{n}f_\theta(x_i)$$

而把似然函数的对数 $l(\theta)=\log L(\theta)$ 称为对数似然函数。

当似然函数对变量 θ 单调时, 可以容易得到其最大值点. 反之，当似然函数为非单调函数且对变量 θ 可微分时，可以求其驻点: 令

$$\frac{dl(\theta)}{d\theta}=0 \qquad (\text{或者}\frac{dL(\theta)}{d\theta}=0)$$

当 θ 为多维时, 比如 $\theta=(\theta_1,\cdots,\theta_k)$ 时令

$$\frac{\partial l(\theta)}{\partial\theta_i}=0 \qquad (\text{或者}\frac{\partial L(\theta)}{\partial\theta_i}=0) \quad i=1,\cdots,k$$

然后判断此驻点是否是最大值点。