矩阵计算 笔记

蒲 飞

一、矩阵的导数

记$X=\begin{pmatrix}x_1\\x_2\\\cdots\\x_n\end{pmatrix}\in R^{n\times 1}$，则

$$\frac{\partial x}{\partial x}=\frac{\partial\begin{pmatrix}x_1\\\cdots\\x_n\end{pmatrix}}{\partial\begin{pmatrix}x_1\\\cdots\\x_n\end{pmatrix}}=\begin{pmatrix}\frac{\partial x_1}{\partial x_1}\\\frac{\partial x_2}{\partial x_2}\\\cdots\\\frac{\partial x_n}{\partial x_n}\end{pmatrix}=\begin{pmatrix}1\\1\\\cdots\\1\end{pmatrix}\in R^{\mathrm{n}\times 1}$$

$$\frac{\partial x^T}{\partial x}=\frac{\partial(x_1,x_2,\ldots,x_n)}{\partial\begin{pmatrix}x_1\\\cdots\\x_n\end{pmatrix}}=\begin{pmatrix}\frac{\partial x_1}{\partial x_1}&\frac{\partial x_2}{\partial x_1}&\cdots&\frac{\partial x_n}{\partial x_1}\\\frac{\partial x_1}{\partial x_2}&\frac{\partial x_2}{\partial x_2}&\cdots&\frac{\partial x_n}{\partial x_2}\\\cdots&\cdots&\cdots&\cdots\\\frac{\partial x_1}{\partial x_n}&\frac{\partial x_2}{\partial x_n}&\cdots&\frac{\partial x_n}{\partial x_n}\end{pmatrix}=I_{\mathrm{n}\times n}$$

$$\frac{\partial x}{\partial x^T}=\frac{\partial\begin{pmatrix}x_1\\\cdots\\x_n\end{pmatrix}}{\partial(x_1,x_2,\ldots,x_n)}=\begin{pmatrix}\frac{\partial x_1}{\partial x_1}&\frac{\partial x_1}{\partial x_2}&\cdots&\frac{\partial x_1}{\partial x_n}\\\frac{\partial x_2}{\partial x_1}&\frac{\partial x_2}{\partial x_2}&\cdots&\frac{\partial x_2}{\partial x_n}\\\cdots&\cdots&\cdots&\cdots\\\frac{\partial x_n}{\partial x_1}&\frac{\partial x_n}{\partial x_2}&\cdots&\frac{\partial x_n}{\partial x_n}\end{pmatrix}=I_{\mathrm{n}\times n}$$

1、若矩阵A和向量y均与向量x无关，这时有

$$\frac{\partial x^T Ay}{\partial x}=\frac{\partial x^T}{\partial x}Ay=Ay$$

注意到

$$y^T Ax=<A^T y,x>=<x,A^T y>=x^T A^T y$$

所以

$$\frac{\partial y^T Ax}{\partial x}=\frac{\partial x^T A^T y}{\partial x}=A^T y$$

推论：

$$\frac{\partial x^T Ax}{\partial x}=Ax+A^T x$$

特别当A为对称矩阵时，有

$$\frac{\partial x^T Ax}{\partial x}=2Ax$$

2、设$Y\in R^{m\times n}$，x 是一个标量，则有 $\frac{dY^{-1}}{dx}=-Y^{-1}\frac{dY}{dx}Y^{-1}$

证明： 由 $\frac{d(YY^{-1})}{dx}=\frac{dI}{dx}=O_{m\times n}$，有 $Y\frac{d(Y^{-1})}{dx}+\frac{dY}{dx}Y^{-1}=O_{m\times n}$，因此 $\frac{d(Y^{-1})}{dx}=-Y^{-1}\frac{dY}{dx}Y^{-1}$。

二、迹（trace）计算

单变量高斯分布的概率密度函数如下（均值为 μ，方差为 σ）：

$$N(x|\mu,\sigma)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

多变量高斯分布（假设 n 维）的概率密度函数如下（均值为 μ，方差矩阵为 Σ)：

$$N(x|u,\Sigma)=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$$

矩阵迹的性质

1、$tr(\alpha A+\beta B)=\alpha tr(A)+\beta tr(B)$

2、$tr(A)=tr(A^T)$

3、$tr(AB)=\mathrm{tr}(BA)$

证明： 设 $A=(\mathrm{a}_{ij})_{n\times n}$ ，$B=(b_{ij})_{n\times n}$，$C=AB=(\mathrm{c}_{ij})_{n\times n}$，$D=BA=(\mathrm{d}_{ij})_{n\times n}$ ，则有：

$$c_{ij}=\sum_{k=1}^{n}a_{ik}b_{kj}\text{，}\quad d_{\mathrm{ij}}=\sum_{k=1}^{n}b_{ik}a_{kj}$$

因此

$$tr(AB)=\sum_{t=1}^{n}\sum_{k=1}^{n}a_{tk}b_{kt}\text{，}\quad tr(BA)=\sum_{t=1}^{n}\sum_{k=1}^{n}b_{tk}a_{kt}$$

进一步推导有

$$tr(AB)=\sum_{t=1}^{n}\sum_{k=1}^{n}a_{tk}b_{kt}=\sum_{k=1}^{n}\sum_{t=1}^{n}a_{tk}b_{kt}=\sum_{k=1}^{n}\sum_{t=1}^{n}a_{tk}b_{kt}=\sum_{k=1}^{n}\sum_{t=1}^{n}b_{kt}a_{tk}=\sum_{t=1}^{n}\sum_{k=1}^{n}b_{tk}a_{kt}=$$
$$\sum_{t=1}^{n}\sum_{k=1}^{n}b_{tk}a_{kt}=tr(BA)$$

4、$tr(ABC)=\mathrm{tr}(CBA)=\mathrm{tr}(BCA)$

证明： 根据性质 3

$$tr(ABC)=\mathrm{tr}((AB)C)=\mathrm{tr}(C(AB))=\mathrm{tr}(CAB)$$
$$tr(ABC)=\mathrm{tr}(A(BC))=tr((BC)A)=\mathrm{tr}(BCA)$$

5、对任何向量 $x\in R^{\mathrm{n}},y\in R^n$ 和矩阵 $A\in R^{n\times n}$，显然 x^TAy 是一个标量，因此有

$$x^T Ay = tr(x^T Ay) = tr(Ayx^T)$$

6、多元变量分布中期望 E 与协方差Σ的性质：

$$E[xx^T] = \Sigma + uu^T$$

证明： $\Sigma = E[(\text{x-}u)(x-u)^T] = E[\text{xx}^T - xu^T - ux^T + uu^T]$

$$= E[\text{xx}^T] - uu^T - uu^T + uu^T = E[xx^T] - uu^T$$

7、$E(\text{x}^T\text{Ax}) = \text{tr}(\Sigma) + \mu^T A\mu$

证明： 因为 x^TAx 是一个标量，可得 $\text{x}^T\text{Ax} = \text{tr}(x^T Ax)$，从而有：

$$E(\text{x}^T\text{Ax}) = E[\text{tr}(x^T Ax)] = E[tr(Axx^T)]$$

$$= tr(E(Axx^T)) = tr(AE(xx^T)) = tr(A(\Sigma + \mu\mu^T))$$

$$= tr(A\Sigma) + tr(A\mu^T\mu) = \text{tr}(A\Sigma) + tr(A\mu^T\mu)$$

$$= tr(A\Sigma) + tr(\mu^T A\mu) = tr(A\Sigma) + \mu^T A\mu$$

8、多元变量高斯分布的 KL 散度

定义： $D_{KL}(P_1 \| P_2) = E_{P_1}[\log \frac{P_1}{P_2}]$

$$D_{KL}(P_1 \| P_2) = E_{P_1}[\log P_1 - \log P_2]$$

$$= \frac{1}{2} E_{P_1}[-\log|\Sigma_1| - (\text{x}-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) + \log|\Sigma_2| + (\text{x}-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)]$$

$$= \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2} E_{P_1}\{-tr[(\text{x}-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)] + tr[(\text{x}-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)]\}$$

$$= \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2} E_{P_1}\{-tr[\Sigma_1^{-1}(x-\mu_1)(\text{x}-\mu_1)^T] + tr[\Sigma_2^{-1}(x-\mu_2)(\text{x}-\mu_2)^T]\}$$

$$= \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2} E_{P_1}\{-tr[\Sigma_1^{-1}(x-\mu_1)(\text{x}-\mu_1)^T]\} + \frac{1}{2} E_{P_1}\{tr[\Sigma_2^{-1}(x-\mu_2)(\text{x}-\mu_2)^T]\}$$

$$=\frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}-\frac{1}{2}tr\{E_{P_1}[\Sigma_1^{-1}(x-\mu_1)(\mathrm{x}-\mu_1)^T]\}+\frac{1}{2}tr\{E_{P_1}[\Sigma_2^{-1}(x-\mu_2)(\mathrm{x}-\mu_2)^T]\}$$

$$=\frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}-\frac{1}{2}tr\{\Sigma_1^{-1}E_{P_1}[(x-\mu_1)(\mathrm{x}-\mu_1)^T]\}+\frac{1}{2}tr\{\Sigma_2^{-1}E_{P_1}[(x-\mu_2)(\mathrm{x}-\mu_2)^T]\}$$

（注意到：方差矩阵Σ_1^{-1}是常量）

$$=\frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}-\frac{1}{2}tr\{\Sigma_1^{-1}E_{P_1}[(x-\mu_1)(\mathrm{x}-\mu_1)^T]\}+\frac{1}{2}tr\{E_{P_1}[\Sigma_2^{-1}(xx^T-\mu_2x^T-\mathrm{x}\mu_2^T+\mu_2\mu_2^T)]\}$$

（注意到：$E_{P_1}[(x-\mu_1)(\mathrm{x}-\mu_1)^T]=\Sigma_1$）

$$=\frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}-\frac{1}{2}tr\{\Sigma_1^{-1}\Sigma_1\}+\frac{1}{2}tr\{\Sigma_2^{-1}E_{P_1}(xx^T-\mu_2x^T-\mathrm{x}\mu_2^T+\mu_2\mu_2^T)\}$$

（注意到：$E_{p_1}(\mathrm{xx}^T)=\Sigma_1+\mu_1\mu_1^T$，$E_{p_1}(\mathrm{x})=\mu_1$和$E_{p_1}(\mathrm{x}^T)=\mu_1^T$）

$$=\frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}-\frac{1}{2}n+\frac{1}{2}tr\{\Sigma_2^{-1}(\Sigma_1+\mu_1\mu_1^T-\mu_2\mu_1^T-\mu_1\mu_2^T+\mu_2\mu_2^T)\}$$

$$=\frac{1}{2}\{\log\frac{|\Sigma_2|}{|\Sigma_1|}-n+tr(\Sigma_2^{-1}\Sigma_1)+tr\{\Sigma_2^{-1}(\mu_1\mu_1^T-\mu_2\mu_1^T-\mu_1\mu_2^T+\mu_2\mu_2^T)\}\}$$

$$=\frac{1}{2}\{\log\frac{|\Sigma_2|}{|\Sigma_1|}-n+tr(\Sigma_2^{-1}\Sigma_1)+tr\{\Sigma_2^{-1}\mu_1\mu_1^T-\Sigma_2^{-1}\mu_2\mu_1^T-\Sigma_2^{-1}\mu_1\mu_2^T+\Sigma_2^{-1}\mu_2\mu_2^T\}\}$$

（注意到：$tr(\Sigma_2^{-1}\mu_1\mu_1^T)=tr(\mu_1^T\Sigma_2^{-1}\mu_1)$，$tr(\Sigma_2^{-1}\mu_2\mu_2^T)=tr(\mu_2^T\Sigma_2^{-1}\mu_2)$

和 $tr(\Sigma_2^{-1}\mu_2\mu_1^T)=tr(\mu_1^T\Sigma_2^{-1}\mu_2)$）

$$=\frac{1}{2}\{\log\frac{|\Sigma_2|}{|\Sigma_1|}-n+tr(\Sigma_2^{-1}\Sigma_1)+tr\{\mu_1^T\Sigma_2^{-1}\mu_1-2\mu_1^T\Sigma_2^{-1}\mu_2+\mu_2^T\Sigma_2^{-1}\mu_2\}\}$$

（注意到：$tr(\mu_1^T\Sigma_2^{-1}\mu_1)=\mu_1^T\Sigma_2^{-1}\mu_1$，$tr(\mu_2^T\Sigma_2^{-1}\mu_2)=\mu_2^T\Sigma_2^{-1}\mu_2$

和 $tr(\mu_1^T\Sigma_2^{-1}\mu_2)=\mu_1^T\Sigma_2^{-1}\mu_2$）

$$=\frac{1}{2}\{\log\frac{|\Sigma_2|}{|\Sigma_1|}-n+tr(\Sigma_2^{-1}\Sigma_1)+(\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\}$$

三、实值函数关于矩阵变量的梯度计算

实值函数 $f(A)$ 相对于 $m\times n$ 矩阵 A 的梯度为一个 $m\times n$ 矩阵，定义为：

$$\frac{\partial f(A)}{\partial A}=\begin{bmatrix}\frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}}\\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}}\\ \cdots & \cdots & \cdots & \cdots\\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}}\end{bmatrix}=\nabla_A f(A)$$

其中，A_{ij} 是矩阵 A 第 i 行第 j 列元素。

1、迹的梯度：$\frac{\partial \mathrm{tr}(A)}{\partial A}=I$

证明： 由 $\mathrm{tr}(A)=\sum_{k=1}^{n}A_{kk}$， 则有：

$$[\frac{\partial \mathrm{tr}(A)}{\partial A}]_{ij}=\frac{\partial \mathrm{tr}(A)}{\partial A_{ij}}=\frac{\partial}{\partial A_{ij}}[\sum_{k=1}^{n}A_{kk}]=\{\begin{matrix}1, & i=j\\ 0, & i\neq j\end{matrix}$$

所以，$\frac{\partial \mathrm{tr}(A)}{\partial A}=I$

2、$\frac{\partial tr(AB)}{\partial A}=\frac{\partial tr(BA)}{\partial A}=B^T$ ，其中 $A\in R^{m\times n}$， $B\in R^{n\times m}$

证明： 首先，矩阵乘积 $[AB]_{ij}=\sum_{l=1}^{n}A_{il}B_{lj}$。因此，$tr(AB)=\sum_{p=1}^{m}\sum_{l=1}^{n}A_{pl}B_{lp}$，于是 $\frac{\partial tr(AB)}{\partial A}$ 是一个 $m\times n$ 矩阵，其元素为 $[\frac{\partial tr(AB)}{\partial A}]_{ij}=\frac{\partial tr(AB)}{\partial A_{ij}}=\frac{\partial}{\partial A_{ij}}(\sum_{p=1}^{m}\sum_{l=1}^{n}A_{pl}B_{lp})=B_{ji}$。从而，$\frac{\partial tr(AB)}{\partial A}=\nabla_A tr(AB)=B^T$。又 $\frac{\partial tr(AB)}{\partial A}=\frac{\partial tr(BA)}{\partial A}$，所以 $\frac{\partial tr(BA)}{\partial A}=B^T$。

3、设 $x, y \in R^{n\times 1}$，则有 $\frac{\partial tr(xy^T)}{\partial x} = \frac{\partial tr(yx^T)}{\partial x} = y$

证明： 易知，$tr(xy^T) = tr(yx^T) = x^T y$，所以 $\frac{\partial tr(xy^T)}{\partial x} = \frac{\partial tr(yx^T)}{\partial x} = \frac{\partial (x^T y)}{\partial x}$=y。

4、单个矩阵迹的梯度矩阵

设 A 是 $n\times n$ 矩阵，则有 $\frac{\partial tr(A^{-1})}{\partial A} = -(-A^{-1})^T$

证明：

5、设 $A \in R^{m\times n}$、$x \in R^{m\times 1}$、$y \in R^{n\times 1}$，则有 $\frac{\partial (x^T Ay)}{\partial A} = xy^T$

证明： 因为 $[\frac{\partial (x^T Ay)}{\partial A}]_{ij} = \frac{\partial}{\partial A_{ij}}(x^T Ay) = \frac{\partial}{\partial A_{ij}}(\sum_{q=1}^{n}\sum_{p=1}^{m} x_p A_{pq} y_q) = x_i y_j$，所以有 $\frac{\partial (x^T Ay)}{\partial A} = xy^T$。

6、设 $A \in R^{m\times n}$，则有 $\frac{\partial tr(A^T A)}{\partial A} = 2A$。

证明： $[\frac{\partial tr(A^T A)}{\partial A}]_{ij} = \frac{\partial}{\partial A_{ij}}(\sum_{t=1}^{n}\sum_{k=1}^{m} A_{kt} A_{kt}) = \frac{\partial}{\partial A_{ij}}(A_{ij} A_{ij}) = 2A_{ij}$，$tr(A^T A)$ 只计算 $A^T A$ 所有第 t 行第 t 列的元素（t=1,2,...,n）。

7、设 $A\in R^{n\times n}$，则有 $\dfrac{\partial(\det(A))}{\partial A}=\det(A)A^{-T}$，这里 $A^{-T}=(A^{-1})^{T}$

证明： 记 $A_{(ij)}=(-1)^{i+j}\det(A_{-(i)(j)})$，其中 $A_{(ij)}$ 是 A_{ij} 的代数余子式。$\det(A_{-(i)(j)})$ 表示行列式 $\det(A)$ 中去掉第 i 行和第 j 列的元素后组成的 $(n-1)\times(n-1)$ 行列式。易知 $\det(A)=\sum_{i=1}^{n}A_{ij}(-1)^{i+j}\det(A_{-(i)(j)})=\sum_{i=1}^{n}A_{ij}A_{(ij)}$。再记 $adj(A)=(A_{(ji)})=(A_{(ij)})^{T}$，因为有 $\sum_{k}A_{ik}(adj(A)_{kj})=\begin{cases}\det(A), & i=j\\ 0, & i\neq j\end{cases}$，即每一行的元素 A_{ik} 乘以该元素对应的代数余子式 $A_{(ik)}$ 之和等于 $\det(A)$，$adj(A)_{ki}$ 对应的元素就是 $A_{(ik)}$。所以 $Aadj(A)=adj(A)A=\det(A)I$，从而 $A^{-1}=\dfrac{1}{\det(A)}adj(A)$。另一方面，$\det(A)$ 中与 A_{ij} 有关的项只有 $A_{ij}A_{(ij)}$，又 $A_{ij}A_{(ij)}=A_{ij}[(adj(A))_{ji}]^{T}$，所以，$\dfrac{\partial(\det A)}{\partial A}=(adj(A))^{T}$。故 $\dfrac{\partial(\det A)}{\partial A}=\det(A)A^{-T}$。

8、设 $A\in R^{m\times n}$，$B\in R^{n\times m}$，则有 $\dfrac{\partial tr(AB)}{\partial A}=\dfrac{\partial tr(BA)}{\partial A}=B^{T}$

证明： 由 $tr(AB)=\sum_{K=1}^{m}\sum_{l=1}^{n}A_{kl}B_{lk}$，有 $[\dfrac{\partial tr(AB)}{\partial A}]_{ij}=\dfrac{\partial}{\partial A_{ij}}[\sum_{k=1}^{m}\sum_{l=1}^{n}A_{kl}B_{lk}]=B_{ji}$，所以 $\dfrac{\partial tr(AB)}{\partial A}=B^{T}$。又 $tr(AB)=tr(BA)$，所以 $\dfrac{\partial tr(BA)}{\partial A}=B^{T}$。

9、设 $A\in R^{n\times n}$ 非奇异、$x\in R^{n\times 1}$、$y\in R^{n\times 1}$，则有 $\dfrac{\partial(x^{T}A^{-1}y)}{\partial A}=-A^{-T}xy^{T}A^{-T}$，其中 $A^{-T}=(A^{-1})^{T}$。

证明：