some background need for ml

zhangyifeng

2018 年 10 月 31 日

目录

1 Matrix **2**
- 1.1 notation 2
- 1.2 Basic calculate 2

2 common distribution **2**
- 2.1 gamma distribution 2

3 Matrix Differentiation **3**
- 3.1 Matrix Differentiation-from functional analysis points 3

4 Lagrange duality **10**
- 4.1 application 10
- 4.2 primal problem 10
- 4.3 generalized Lagrange function 10
- 4.4 dual problem 11
- 4.5 KKT(Karush-Kuhn-Tucker)condition 11

1 Matrix

1.1 notation

$\boldsymbol{\alpha} \in \mathbb{R}^n$
$\boldsymbol{x} \in \mathbb{R}^n$
$A \in \mathbb{R}^{m \times n}$
$(A)_{ij} = a_{ij}$
A^T: transpose of A
$tr(A) = \sum_{i=1}^{n} a_{ii}$
$det(A) = \sum_{\sigma \in S_n} par(\sigma) a_{1\sigma_1} a_{2\sigma_2} \ldots_{1\sigma_n}$, where S_n is all set of $n - order$ permutation. par(σ) can be -1 or +1.
also, $det(A) = \sum_{i=1}^{n} a_{ki} A_{ki} (k = 1, 2, \ldots, n) = \sum_{j=1}^{n} a_{jl} A_{jl} (l = 1, 2, \ldots, n)$
Frobenius norm of A:

$$\|A\|_F = (tr(A^T A))^{1/2} = (\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2)^{1/2}$$

it can be regarded as L_2 norm when metrix was extended to vectors

1.2 Basic calculate

$tr(AB) = tr(BA)$
$tr(ABC) = tr(BCA) = tr(CAB)$

2 common distribution

2.1 gamma distribution

gamma function

$\Gamma(a) = \int_0^\infty x^{a-1} \exp^{-x} dx$, where $a > 0$

3 Matrix Differentiation

3.1 Matrix Differentiation-from functional analysis points

假设 X 和 Y 为赋范向量空间，$F: X \to Y$ 是一个映射，那么 F 在 $x_0 \in X$ 可导的意思是说存在一个有界线性算子 $L \in \mathcal{L}(X,Y)$，使得对于任意的 $\epsilon > 0$ 都存在 $\delta > 0$，对于满足 $x \in X\backslash\{x_0\}, \|x - x_0\| < \delta$ 的 x 都有 $\frac{\|F(x)-F(x_0)-L(x-x_0)\|}{\|x-x_0\|} < \epsilon$. 我们称 $L(\|x - x_0\|)$ 为 F 在 x_0 点的微分。

以上定义有一个等价的表述，往往计算起来更方便：对于距离 x_0 足够近的点 x，即 $\lim_{x\to x_0} \frac{o(\|x-x_0\|)}{\|x-x_0\|} = 0$，有 $F(x) = F(x_0)+L(x-x_0)+o(\|x-x_0\|)$. (注：此处 $L(x - x_0)$ 应该理解为线性算子 L 在 $x - x_0$ 这个点的值，而不是 L 乘以 $x - x_0$。不过在有限维空间所有线性算子都可以用矩阵表述，L 在 $x - x_0$, 这个值便正好可以表述为矩阵与向量的乘积 (Riesz 表示定理))

例子 1：假设 $F(X) = X^TX$ 是一个 $\mathbb{R}^{m\times n} \to \mathbb{S}^n$ 的映射，其中 $\mathbb{S}^n$ 为 n 维对称阵的空间。

$$\begin{aligned}
&F(X + \Delta X) - F(X)\\
=&(X + \Delta X)^T(X + \Delta X) - X^TX\\
=&X^T\Delta X + \Delta X^TX + o(\|\Delta X\|)
\end{aligned}$$

所以我们有 $L(\Delta X) = 2X^T\Delta X$，这个就是 F 在 X 点的微分。

例子 2：最小二乘问题 $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ f 是一个 $\mathbb{R}^n \to \mathbb{R}$ 的映射。

$$\begin{aligned}
&f(x + \Delta x) - f(x)\\
=&\frac{1}{2}\|A(x + \Delta x) - b\|^2 - \frac{1}{2}\|Ax - b\|^2\\
=&\frac{1}{2}\|Ax - b + A\Delta x\|^2 - \frac{1}{2}\|Ax - b\|^2\\
=&(Ax - b)^TA\Delta x + o(\|\Delta x\|)
\end{aligned}$$

所以我们有 $L(\Delta x) = (Ax-b)^TA\Delta x$，这个就是 f 在 x 点的微分。在这种情况下，L 这个有界线性算子 (梯度) 可以用矩阵来表述 (Riesz 表示定理)：$L(\Delta x) = \langle\nabla f(x), \Delta x\rangle = (Ax - b)^TA\Delta x$，所以梯度 $\nabla f(x) = A^T(Ax - b)$

总结：在有限维的情况下，我们可以先求 F 的微分 $L(\Delta x)$，利用 Riesz 表示定理，得 $L(\Delta x) = \langle f^{'}(x), \Delta x\rangle$，可求得对应的 gradient vector 或者

jacobi 矩阵 $f^{'}(x)$，也就是导数，显然，这里可以看出，导数和微分差一个转置。

标量 f 对矩阵 X 的导数

核心思想

函数的微分 = 函数的导数和自变量的微分的内积 $= \mathrm{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$

矩阵微分运算法则

加减法：$d(X \pm Y) = dX \pm dY$

矩阵乘法：$d(XY) = (dX)Y + XdY$

转置：$d(X^T) = (dX)^T$

迹：$d\mathrm{tr}(X) = \mathrm{tr}(dX)$

逆：$dX^{-1} = -X^{-1}dXX^{-1}$。此式可在 $XX^{-1} = I$ 两侧求微分来证明

行列式：$d|X| = \mathrm{tr}(X^{\#}dX)$ ，其中 $X^{\#}$ 表示 X 的伴随矩阵，在 X 可逆时又可以写作 $d|X| = |X|\mathrm{tr}(X^{-1}dX)$。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：$d\sigma(X) = \sigma'(X) \odot dX$ ，$\sigma(X) = [\sigma(X_{ij})]$ 是逐元素标量函数运算，$\sigma'(X) = [\sigma'(X_{ij})]$ 是逐元素求导数。

举个例子，$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}, d\sin(X) = \begin{bmatrix} \cos x_{11}dx_{11} & \cos x_{12}dx_{12} \\ \cos x_{21}dx_{21} & \cos x_{22}dx_{22} \end{bmatrix} = \cos(X) \odot dX$。

迹技巧 (trace trick)

标量套上迹：$a = \mathrm{tr}(a)$

转置：$tr(A^T) = tr(A)$

线性：$\mathrm{tr}(A \pm B) = \mathrm{tr}(A) \pm \mathrm{tr}(B)$

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

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 $\frac{\partial f}{\partial x}=\frac{\partial f}{\partial y}\frac{\partial y}{\partial x}$，但这里我们不能沿用链式法则，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。于是我们继续追本溯源，链式法则是从何而来？源头仍然是微分。我们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 (这个是矩阵对矩阵的导数，在下一节我们会了解到)，并使用迹技巧将其他项交换至 dX 左侧，即可得到 $\frac{\partial f}{\partial X}$。

标量对矩阵的一般求导步骤

1. 对标量函数 f 两端作微分，利用微分运算法则化简
2. 对两端作迹运算，利用迹运算法则化简，将 dx 移到最右端
3. 利用微分和矩阵的联系 $df=\mathrm{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$, 求 $\frac{\partial f}{\partial X}$

一些例子

例 1：$f=\boldsymbol{a}^T X\boldsymbol{b}$，求 $\frac{\partial f}{\partial X}$。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：1. 作微分: 这里的 $\boldsymbol{a},\boldsymbol{b}$ 是常量，$d\boldsymbol{a}=\boldsymbol{0}, d\boldsymbol{b}=\boldsymbol{0}$，得：$df=\boldsymbol{a}^T dX\boldsymbol{b}$

2. 作迹运算:$df=\mathrm{tr}(\boldsymbol{a}^T dX\boldsymbol{b})=\mathrm{tr}(\boldsymbol{b}\boldsymbol{a}^T dX)$,注意这里我们根据 $\mathrm{tr}(AB)=\mathrm{tr}(BA)$ $\boldsymbol{a}^T dX\ \boldsymbol{b}$

3. 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：$f=\boldsymbol{a}^T\exp(X\boldsymbol{b})$，求 $\frac{\partial f}{\partial X}$。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，exp 表示逐元素求指数，f 是标量。

解：1. 作微分：$df=\boldsymbol{a}^T(\exp(X\boldsymbol{b})\odot(dX\boldsymbol{b}))$

2. 作迹运算:$df=\mathrm{tr}(\boldsymbol{a}^T(\exp(X\boldsymbol{b})\odot(dX\boldsymbol{b})))=\mathrm{tr}((\boldsymbol{a}\odot\exp(X\boldsymbol{b}))^T dX\boldsymbol{b})=\mathrm{tr}(\boldsymbol{b}(\boldsymbol{a}\odot\exp(X\boldsymbol{b}))^T dX)$

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

例 3【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（此时可以省略第二步: 作迹运算）。

先将向量模平方改写成向量与自身的内积: $l = (X\boldsymbol{w} - \boldsymbol{y})^T(X\boldsymbol{w} - \boldsymbol{y})$

1. 求微分:$dl = (Xd\boldsymbol{w})^T(X\boldsymbol{w} - \boldsymbol{y}) + (X\boldsymbol{w} - \boldsymbol{y})^T(Xd\boldsymbol{w}) = 2(X\boldsymbol{w} - \boldsymbol{y})^T Xd\boldsymbol{w}$。

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

例 4【方差的最大似然估计】: 样本 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim N(\boldsymbol{\mu}, \Sigma)$，求方差 Σ 的最大似然估计。写成数学式是: $l = \log|\Sigma| + \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$，求 $\frac{\partial l}{\partial \Sigma}$ 的零点。其中 $\boldsymbol{x}_i$ 是 $m \times 1$ 列向量，$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ 是样本均值，Σ 是 $m \times m$ 对称正定矩阵，l 是标量。

解: 1. 作微分: 第一项是 $d\log|\Sigma| = |\Sigma|^{-1}d|\Sigma| = \mathrm{tr}(\Sigma^{-1}d\Sigma)$，第二项是 $\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T d\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = -\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$。

2. 作迹运算:$\mathrm{tr}\left(\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right) = \frac{1}{n}\sum_{i=1}^n \mathrm{tr}((\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) = \frac{1}{n}\sum_{i=1}^n \mathrm{tr}\left(\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\right)$
$= \mathrm{tr}(\Sigma^{-1}S\Sigma^{-1}d\Sigma)$, 定义 $S = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$ 为样本方差矩阵。得到 $dl = \mathrm{tr}\left((\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1})\, d\Sigma\right)$。

3. 对照导数与微分的联系，有 $\frac{\partial l}{\partial \Sigma} = (\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1})^T$，其零点即 Σ 的最大似然估计为 $\Sigma = S$。

矩阵 F 对矩阵 X 的导数

一般而言，标量就是 1×1 的矩阵，如果我们能推导出矩阵对矩阵的导数，标量对矩阵的导数不是自然的么，不应该可以统一进来么，那为啥还要大费周章地先写标量对矩阵的导数。原因是这两者不完全相同，并不能很简单地统一起来。

应该怎么定义矩阵对矩阵的导数。回答这个问题不是随意的，为了满足两个要求，我们对矩阵对矩阵的定义有严格的要求。我们的两个要求是:

1. 矩阵 $F \in \mathbb{R}^{p\times q}$ 对矩阵 $X \in \mathbb{R}^{m\times n}$ 的导数应包含所有 $mnpq$ 个偏导数 $\frac{\partial F_{kl}}{\partial X_{ij}}$，从而不损失信息。

2. 在标量对矩阵求导的地方，我们发现导数与微分有简明的联系。这里我们仍希望他们之间存在某种联系。

为此，我们先定义向量 $\boldsymbol{f}(p\times1)$ 对向量 $\boldsymbol{x}(m\times1)$ 的导数

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}=\begin{bmatrix}\frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_p}{\partial x_1}\\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_p}{\partial x_2}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial f_1}{\partial x_m} & \frac{\partial f_2}{\partial x_m} & \cdots & \frac{\partial f_p}{\partial x_m}\end{bmatrix}(m\times p)$$

此时，可以证明，$d\boldsymbol{f}=\sum_{i,j}\frac{\partial f_i}{\partial x_j}dx_j=\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}^T d\boldsymbol{x}$, 这个定义满足我们的两个要求，所以我们现在作好了了向量对向量的导数。

再定义矩阵的（按列优先）向量化：

$$vec(X)=[X_{11},\ldots,X_{m1},X_{12},\ldots,X_{m2},\ldots,X_{1n},\ldots,X_{mn}]^T(mn\times1)$$

并定义矩阵 F 对矩阵 X 的导数 $\frac{\partial F}{\partial X}=\frac{\partial vec(F)}{\partial vec(X)}(mn\times pq)$。此时，可以证明，导数与微分有联系 $vec(dF)=\frac{\partial F}{\partial X}^T vec(dX)$，这样，我们作好了满足要求的矩阵关于矩阵的导数。

列向量化运算法则

1. 线性：$vec(A+B)=vec(A)+vec(B)$。

2.矩阵乘法：$vec(AXB)=(B^T\otimes A)vec(X)$，其中 $\otimes$ 表示 Kronecker 积，$A(m\times n)$ $B(p\times q)$ 的 Kronecker 积是 $A\otimes B=[A_{ij}B](mp\times nq)$。此式证明见张贤达《矩阵分析与应用》第 107-108 页。

3. 转置：$vec(A^T)=K_{mn}vec(A)$，A 是 $m\times n$ 矩阵，其中 $K_{mn}(mn\times mn)$ 是换位矩阵 (commutation matrix)(就是一些初等换位矩阵的乘积)。

4. 逐元素乘法：$vec(A\odot X)=diag(A)vec(X)$，其中 $diag(A)(mn\times mn)$ 是用 A 的元素（按列优先）排成的对角阵。

一些 Kronecker 积和交换矩阵相关的恒等式

1.$(A\otimes B)^T=A^T\otimes B^T$。

2.$vec(\boldsymbol{a}\boldsymbol{b}^T)=\boldsymbol{b}\otimes\boldsymbol{a}$。

3.$(A\otimes B)(C\otimes D)=(AC)\otimes(BD)$。可以对 $F=D^TB^TXAC$ 求导来证明，一方面，直接求导得到 $\frac{\partial F}{\partial X}=(AC)\otimes(BD)$；另一方面，引入 $Y=B^TXA$，有 $\frac{\partial F}{\partial Y}=C\otimes D$, $\frac{\partial Y}{\partial X}=A\otimes B$，用链式法则得到 $\frac{\partial F}{\partial X}=(A\otimes B)(C\otimes D)$。

4.$K_{mn} = K_{nm}^T, K_{mn}K_{nm} = I$，所以换位矩阵是正交矩阵。

5.$K_{pm}(A \otimes B)K_{nq} = B \otimes A$，$A$ 是 $m \times n$ 矩阵，B 是 $p \times q$ 矩阵。可以对 AXB^T 做向量化来证明，一方面，$vec(AXB^T) = (B \otimes A)vec(X)$；另一方面，$vec(AXB^T) = K_{pm}vec(BX^TA^T) = K_{pm}(A \otimes B)vec(X^T) = K_{pm}(A \otimes B)K_{nq}vec(X)$。

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？从导数与微分的联系入手，$vec(dF) = \frac{\partial F}{\partial Y}^T vec(dY) = \frac{\partial F}{\partial Y}^T \frac{\partial Y}{\partial X}^T vec(dX)$，可以推出链式法则 $\frac{\partial F}{\partial X} = \frac{\partial Y}{\partial X}\frac{\partial F}{\partial Y}$

矩阵对矩阵的一般求导步骤

1. 对矩阵值函数 F 两端作微分，利用微分运算法则化简

2. 对两端作列向量化运算，利用列向量化法则化简，注意看列向量里面是什么形式，就用什么公式，如列向量里面是两个矩阵相乘，就想办法凑进去一个单位矩阵，并使得 $vecx$ 在中间，然后利用 vec 的矩阵乘法公式

3. 利用微分和矩阵的联系 $vec(dF) = \frac{\partial F}{\partial X}^T vec(dX)$, 求 $\frac{\partial f}{\partial X}$

一些例子

例 1：$F = AX$，X 是 $m \times n$ 矩阵，求 $\frac{\partial F}{\partial X}$。

解：1. 作微分：$dF = AdX$

2. 列向量化，使用矩阵乘法的技巧，注意在 dX 右侧添加单位阵：$vec(dF) = vec(AdX) = (I_n \otimes A)vec(dX)$

3. 对照导数与微分的联系得到 $\frac{\partial F}{\partial X} = I_n \otimes A^T$。

特例：如果 X 退化为向量，即 $\boldsymbol{f} = A\boldsymbol{x}$，则根据向量的导数与微分的关系 $\mathrm{d}\boldsymbol{f} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}^T d\boldsymbol{x}$，得到 $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = A^T$

$df(\mathbf{X}, \mathbf{Y}) = tr(\frac{\partial f}{\partial \mathbf{X}}^T d\mathbf{X}) + tr(\frac{\partial f}{\partial \mathbf{Y}}^T d\mathbf{Y})$

例 2：$f = \log|X|$ ，X 是 n×n 矩阵，求 $\nabla_X^2 f$。

解：1. 求微分：$d\nabla_X f = -(X^{-1}dXX^{-1})^T$

2. 列向量化，$vec(d\nabla_X f) = -K_{nn}vec(X^{-1}dXX^{-1}) = -K_{nn}(X^{-T} \otimes X^{-1})vec(dX)$，

3. 对照导数与微分的联系，得到 $\nabla_X^2 f = -K_{nn}(X^{-T} \otimes X^{-1})$，注意它是对称矩阵。

例 3：$F = A\exp(XB)$ A 是 $l \times m$ 矩阵，X 是 $m \times n$ 矩阵，B 是 $n \times p$ 矩阵，exp 为逐元素函数，求 $\frac{\partial F}{\partial X}$。

解：1. 求微分：$dF = A(\exp(XB) \odot (dXB))$

2. 列向量化:$vec(dF) = (I_p \otimes A)vec(\exp(XB) \odot (dXB)) = (I_p \otimes A)diag(\exp(XB))vec(dXB) = (I_p \otimes A)diag(\exp(XB))(B^T \otimes I_m)vec(dX)$。

3. 对照导数与微分的联系得到 $\frac{\partial F}{\partial X} = (B \otimes I_m)diag(\exp(XB))(I_p \otimes A^T)$。

注解

1. 一般而言，这套方法就是为了矩阵对矩阵求导而引入的，由于这里是利用列向量定义的导数，所以直接应用在标量对矩阵 $X \in \mathbb{R}^{m \times n}$ 的导数上，会得到一个 $mn \times 1$ 的列向量，这与我们一般定义的标量对矩阵的导数相悖，所以一般标量对矩阵的导数，我们还是利用上一节的方法。当然，若将上一节定义的标量 $f(X) \in \mathbb{R}^1$ 对矩阵 $X \in \mathbb{R}^{m \times n}$ 的导数用记号 $\nabla_X f \in \mathbb{R}^{m \times n}$ 来表示，则这里定义的 $\frac{\partial f}{\partial X} = vec(\nabla_X f)$, 在牢记这一条的情况下，我们可以用本节的方法来解决标量对矩阵求导，只是没有上一节的方法方便。为了满足读者的好奇心，我们给出标量对矩阵求导的一个例子，并且用两种方法来解决。

2. 标量对矩阵的二阶导数，又称 Hessian 矩阵，定义为 $\nabla_X^2 f = \frac{\partial^2 f}{\partial X^2} = \frac{\partial \nabla_X f}{\partial X}(mn \times mn)$ 是对称矩阵，这个二阶导数分两次进行，第一次是标量对矩阵求导，第二次是矩阵对矩阵求导。

3. 如何理解 $K_{mn}(mn \times mn)$, 它是一个换位矩阵，根据 $vec(A^T) = K_{mn}$ $vec(A)$，它的作用是使的 $vec(A^T)$ 和 $vec(A)$ 的若干行对换位置。由 $[A]_{i,j} = [A]_{j,i} = [vec(A^T)]_{(i-1)n+j} = [vec(A)]_{(j-1)n+i}$，这里 $A \in \mathbb{R}^{m \times n}, 1 \le i \le m, 1 \le j \le n$，所以 K_{mn} 就是单位矩阵 (mn×mn) 交换 $(i-1)n+j$ 和 $(j-1)n+i$ 行得到的一个矩阵。

对两节内容的总结

我们发展了从整体出发的矩阵求导的技术，导数与微分的联系是计算的枢纽。

上一节中，我们了解了，标量对矩阵的导数与微分的联系是 $df = tr((\nabla_X f)^T dX)$，先对 f 求微分，再使用迹技巧可求得导数，特别地，标量对向量的导数与微分的联系是 $df = (\nabla_{\boldsymbol{x}} f)^T d\boldsymbol{x}$

下一节中，我们了解了，矩阵对矩阵的导数与微分的联系是 $vec(dF) = \frac{\partial F}{\partial X}^T vec(dX)$，先对 F 求微分，再使用列向量化的技巧可求得导数，特别地，向量对向量的导数与微分的联系是 $d\boldsymbol{f} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}^T d\boldsymbol{x}$。

reference

如何理解矩阵对矩阵求导？-知乎-猪猪专业户
矩阵求导术 (上)-知乎-长躯鬼侠
矩阵求导术 (下)-知乎-长躯鬼侠

4 Lagrange duality

4.1 application

applied on:
- 最大熵模型
- SVM(support vector machine)

4.2 primal problem

Set $f(\boldsymbol{x}), c_i(\boldsymbol{x}), h_j(\boldsymbol{x})$ are continuously differentiable function over $bmRn$, consider optimization problem with constraints

$$\begin{aligned} \min_{\boldsymbol{x} \in \boldsymbol{R}^n} \ & f(\boldsymbol{x}) \\ s.t. \ & c_i(\boldsymbol{x}) \leq 0, \ i = 1, 2, \cdots, k \\ & h_j(\boldsymbol{x}) = 0, \ j = 1, 2, \cdots, l \end{aligned}$$

4.3 generalized Lagrange function

$$L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{x}) + \sum_{i=1}^{k} a_i c_i(\boldsymbol{x}) + \sum_{j=1}^{l} \beta_j h_j(\boldsymbol{x})$$

where, $\boldsymbol{x} = (x^1, x^2, \ldots, x^n)^T \in \boldsymbol{R}^n$,$\alpha_i, \beta_j$ are Lagrange multiplier, $\alpha_i \geq 0$
After introduced generalized Lagrange function, primal problem is equal to

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$
$$s.t.\ \alpha_i \geq 0,\ i = 1, 2, \ldots, k$$

4.4 dual problem

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$
$$s.t.\ \alpha_i \geq 0,\ i = 1, 2, \ldots, k$$

4.5 KKT(Karush-Kuhn-Tucker)condition

$$\begin{aligned}
\nabla_x L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= 0 \\
\nabla_\alpha L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= 0 \\
\nabla_\beta L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= 0 \\
\alpha_i c_i(\boldsymbol{x}) &= 0,\ i = 1, 2, \ldots, k \\
c_i(\boldsymbol{x}) &\leq 0,\ i = 1, 2, \ldots, k \\
\alpha_i &\geq 0, i = 1, 2, \ldots, k \\
h_j(\boldsymbol{x}) &= 0, j = 1, 2 \ldots, l
\end{aligned}$$

Theorem 1 *if $f(\boldsymbol{x})$ and $c_i(\boldsymbol{x})$ are convex function, $h_j(\boldsymbol{x})$ are affine function*[1]*, and inequation constrains $c_i(\boldsymbol{x})$ strictly hold, that is, exist $\boldsymbol{x}$, s.t. for any i, hold $c_i(\boldsymbol{x}) < 0$, then, there must be $\boldsymbol{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ are the optimal solution of primal problem as well as dual problem and satisfy KKT condition at $\boldsymbol{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$.*

Remark: so, when the prerequisites are satisfied, we can use KKT condition to find the optimal solution $\boldsymbol{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$.

[1] $f(x)$ is called affine function, when it holds that $f(x) = \boldsymbol{a} \cdot \boldsymbol{x} + b, \boldsymbol{a} \in \boldsymbol{R}^n, b \in \boldsymbol{R}, \boldsymbol{x} \in R^n$