矩阵求导补充

写这篇是为了让读者能看懂原书的一条公式，关于那个公式原书略过了一些细节，对新手比较致命。

那个公式就是线性回归的解析解，其实读者完全可以跳过那个解析解，当然进而也可以跳过本节这么一大坨，但……来测吧！

这篇 theory 主要就是补充一下矩阵求导的细节，比如像一元微积分中对多项式求导的结论，可以加快计算的，再比如加法求导法则、链式求导法则之类的。

首先要先对矩阵求导有个基本印象，这里可以看我写的 矩阵求导 (isirin1131.github.io)。

其次，我们不需要对完整的矩阵求导建立代数系统，而只需要对向量求导建立理论和得出并记住一些基本结果，这对理解线性回归这一节中的公式已经足够了。（其实是我搞不定更高的理论）

约定：依据分子布局，所有的向量都是列向量

向量求导形如： 对于向量 $\mathbb{x}$，其长度为 n，形如 $\{x_1, x_2, \cdots, x_n\}$，对于向量 $\mathbb{y}$，也就是 $\mathbb{f}(\mathbb{x})$，其长度为 m，形如 $\{f_1(\mathbb{x}), f_2(\mathbb{x}), \cdots, f_m(\mathbb{x})\}$，我们要求 $\dfrac{\partial}{\partial \mathbb{x}}\mathbb{y}$，得到的结果是个 $m \times n$ 的矩阵，并且如果 $\mathbb{y}\mathbb{x}^T$ 这个 $m \times n$ 矩阵的第 (i, j) 项是 $f_i(\mathbb{x})\mathbb{x}_j$，那么 $\dfrac{\partial}{\partial \mathbb{x}}\mathbb{y}$ 的第 (i, j) 项就是 $\dfrac{\partial}{\partial x_j} f_i(\mathbb{x})$。

几个基础的例子：

1. $\dfrac{\partial}{\partial \mathbb{x}}\mathbb{a}$ 等于 $\mathbb{0}$（$\mathbb{a}$ 是与 $\mathbb{x}$ 无关的常数向量，$\mathbb{0}$ 是全零矩阵），这个很容易推
2. $\dfrac{\partial}{\partial \mathbb{x}}\mathbb{x}$ 等于 $\mathbb{1}$（$\mathbb{1}$ 是一个对角线为 1，其余地方为 0 的方阵），这个也比较容易推
3. $\dfrac{\partial}{\partial \mathbb{x}}\mathbb{A}\mathbb{x}$ 等于 $\mathbb{A}$（$\mathbb{A}$ 是个列数为 n 的矩阵，行数不做要求），这个稍微有点难，就稍微讲讲，首先 $\mathbb{A}\mathbb{x}$ 是个列向量，而且 $\mathbb{A}\mathbb{x}$ 的每个分量都是 $\mathbb{A}$ 的某个行向量与 $\mathbb{x}$ 的点积，往后就显然了
4. $\dfrac{\partial}{\partial \mathbb{x}}\mathbb{x}^T\mathbb{A}$ 等于 A^T，这个跟上一个存在某种对应的关系，读者不需要管，硬推就行了。
5. $\dfrac{\partial}{\partial \mathbb{x}}\mathbb{x}^T\mathbb{A}\mathbb{x}$，这个稍微复杂，首先 $\mathbb{x}^T\mathbb{A}\mathbb{x}$ 是个标量，具体地，形如 $\mathbb{x}^T\mathbb{a}_1x_1 + \mathbb{x}^T\mathbb{a}_2x_2 + \cdots + \mathbb{x}^T\mathbb{a}_nx_n$（$\mathbb{a}_i$ 是列向量），然后 $\dfrac{\partial}{\partial \mathbb{x}}\mathbb{x}^T\mathbb{A}\mathbb{x}$ 就是行向量，但有点复杂，比如它第一项是 $\mathbb{x}^T\mathbb{a}_1 + a_{1,1}x_1 + a_{2,1}x_2 + \cdots + a_{n,1}x_1$，第二项是 $\mathbb{x}^T\mathbb{a}_2 + a_{1,2}x_2 + a_{2,2}x_2 + \cdots + a_{n,2}x_2$，到这个地步我们也能看出来了，$\dfrac{\partial}{\partial \mathbb{x}}\mathbb{x}^T\mathbb{A}\mathbb{x}$ 等于 $\mathbb{x}^T(\mathbb{A} + \mathbb{A}^T)$

6. $\frac{\partial}{\partial \mathbb{x}} \mid \mathbb{x} \mid^2$ 等于 $2\mathbb{x}^T$（$\mid \mathbb{x} \mid^2$ 意思是向量长度的平方，数值上等于 $\sum_{i=1}^n (x_i)^2$，其实是个标量，但我们也可以把它当成向量处理），这个比较显然

这向量求导的这些性质怎么这么像标量求导的那些性质？那既然如此，四则运算求导法则和链式求导法则也存在吗？

存在。

首先，$\frac{\partial}{\partial \mathbb{x}} \alpha \mathbb{y} = \alpha \frac{\partial}{\partial \mathbb{x}} \mathbb{y}$，其中 α 是标量，这个很好验证；$\frac{\partial}{\partial \mathbb{x}} \mathbb{A} \mathbb{y} = \mathbb{A} \frac{\partial}{\partial \mathbb{x}} \mathbb{y}$(其中 $\mathbb{A}$ 是个列数为 m 的矩阵，行数不做要求)，这个看着复杂，实际上一推就有。

其次，$\frac{\partial}{\partial \mathbb{x}} (\mathbb{u} + \mathbb{v}) = \frac{\partial}{\partial \mathbb{x}} \mathbb{u} + \frac{\partial}{\partial \mathbb{x}} \mathbb{v}$，其中 $\mathbb{u}, \mathbb{v}$ 都长为 m，形如 $\{f_1(\mathbb{x}), f_2(\mathbb{x}), \cdots, f_m(\mathbb{x})\}$。这个也是一推就有。

乘法法则没有，不会推（

（ps：其实更弱的情况，比如标量对向量求导，是可以有乘法法则的，但这里没有。）

最后，链式法则也是成立的，也就是 $\frac{\partial}{\partial \mathbb{x}} \mathbb{y} = \frac{\partial}{\partial \mathbb{u}} \mathbb{y} \frac{\partial}{\partial \mathbb{x}} \mathbb{u}$

证明这个挺简单，矩阵乘一下就明白了，但是要先知道一个理论（不然证不了），也就是原书里提到过的“多变量函数的链式法则”，如下：

> 让我们先考虑单变量函数。假设函数$y = f(u)$和$u = g(x)$都是可微的，根据链式法则：$\frac{dy}{dx} = \frac{dy}{du} \frac{du}{dx}$.现在考虑一个更一般的场景，即函数具有任意数量的变量的情况。假设可微分函数y有变量$u_1, u_2, \ldots, u_m$，其中每个可微分函数u_i都有变量$x_1, x_2, \ldots, x_n$。注意，y是x_1, x_2， $\ldots, x_n$的函数。对于任意$i = 1, 2, \ldots, n$，链式法则给出：$\frac{\partial y}{\partial x_i} = \frac{\partial y}{\partial u_1} \frac{\partial u_1}{\partial x_i} + \frac{\partial y}{\partial u_2} \frac{\partial u_2}{\partial x_i} + \cdots + \frac{\partial y}{\partial u_m} \frac{\partial u_m}{\partial x_i}$

唉，这个多变量函数的链式法则实际上是多变量微积分的内容，我不会证，读者自己了解吧。

我们主要关注的还是向量求导的链式法则。

完结，不写了！