第零章、一些数学准备

赵涵

2022 年 1 月 19 日

1 高等数学

这一部分取自[1]。

1.1 导数和微分

导数的定义：

$$f'(x_0) = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x_0} \tag{1}$$

或者

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \tag{2}$$

对于一个一元函数来讲，可以从左边两边分别定义导数，即左导数与右导数，分别有如下的定义，其中左导数：

$$f'_-(x_0) = \lim_{\Delta x \to 0^-} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} \tag{3}$$

右导数：

$$f'_+(x_0) = \lim_{\Delta x \to 0^+} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} \tag{4}$$

在定义完左右导数后，一个函数的导数性质就可以给出了。说一个函数在某一点可导，就意味着该函数在该点左右导数都存在且相等。关于函数的可导性的一些定理：

定理1：函数$f(x)$在x_0处可导⇔函数$f(x)$ 在x_0处可微。

定理2：函数$f(x)$在x_0处可导⇒函数$f(x)$ 在x_0处可导连续，反之不成立。

定理3：$f'(x_0) = A \Leftrightarrow f'_-(x_0) = f'_+(x_0) = A$。

1.2 平面曲线的切线和法线

对于一个平面曲线，函数在x_0导数为该点函数切线的斜率，可以定义该切线的切线方程，当切线方程给出后，与切线垂直的直线，称为法线，两条直线方程的斜率相乘等于-1。所以可以通过解析几何给出切线方程和法线方程的表达式：

$$y - y_0 = f'(x_0)(x - x_0) \quad Tangent \tag{5}$$

$$y - y_0 = -\frac{1}{f'(x - x_0)}(x - x_0) \quad Normal \tag{6}$$

1.3 函数的求导规则

若函数$u=u(x),v=v(x)$在点x出可导，则：

$$(u\pm v)' = u'+v' \tag{7}$$
$$(uv)' = u'v+uv' \tag{8}$$
$$(\frac{u}{v})' = \frac{u'v-uv'}{v^2} \tag{9}$$

通过以上的函数求导法则，可以求函数的导数了。在此处给出一些基本函数的导数供读者查阅。

$$y=const. \rightarrow y'=0 \tag{10}$$
$$y=x^{\alpha}(\alpha\in\mathbb{R}) \rightarrow y'=\alpha x^{\alpha-1} \tag{11}$$
$$y=a^{x}(a\in\mathbb{R}) \rightarrow y'=a^{x}\ln a \tag{12}$$
$$y=\log_a x \rightarrow y'=\frac{1}{x\ln a} \tag{13}$$
$$y=\sin x \rightarrow y'=\cos x \tag{14}$$
$$y=\cos x \rightarrow y'=-\sin x \tag{15}$$
$$y=\tan x \rightarrow y'=\frac{1}{\cos^2 x} \tag{16}$$
$$y=\cot x \rightarrow y'=-\frac{1}{\sin^2 x} \tag{17}$$
$$y=\arcsin x \rightarrow y'=\frac{1}{\sqrt{1-x^2}} \tag{18}$$
$$y=\arccos x \rightarrow y'=-\frac{1}{\sqrt{1-x^2}} \tag{19}$$
$$y=shx \rightarrow y'=chx \tag{20}$$
$$y=chx \rightarrow y'=shx \tag{21}$$

对于机器学习这门课程来说，最重要的就是复合函数的求导法则了。若$u=g(x)$在x出可导，并且$y=f(u)$在对应点u也可导，则复合函数的求导法则为：

$$\frac{dy}{dx}=\frac{dy}{du}\frac{du}{dx}=f'(u)g'(x) \tag{22}$$

请注意，虽然等式最后一步都是带一撇的符号，但是微分变量不一样，所以对于复合函数的求导，一般使用等式中间的写法。

一些基本函数，具有高阶导数，即多次求导，先给出一些常用的高阶导数公式：

$$(a^x)^{(n)} = a^x\ln^n a \quad (a>0) \tag{23}$$
$$(\sin kx)^{(n)} = k^n\sin(kx+n\frac{\pi}{2}) \tag{24}$$
$$(\cos kx)^{(n)} = k^n\cos(kx+n\frac{\pi}{2}) \tag{25}$$
$$(x^m)^{(n)} = m(m-1)\cdots(m-n+1)x^{m-n} \tag{26}$$
$$(\ln x)^{(n)} = (-1)^{n-1}\frac{(n-1)!}{x^n} \tag{27}$$

1.4 泰勒公式

泰勒公式是对一个函数的近似，一般在函数很复杂时，对函数进行泰勒展开，取前几项，舍去其余项，用来逼近原函数。在x_0 处的泰勒展开为：

$$f(x)=\sum_{i=0}^{n}\frac{f^{(i)}(x_0)}{i!}(x-x_0)^i+R_n(x) \tag{28}$$

其中R_n称为$f(x)$在x_0处的n阶泰勒余项,一般情况下，我们不会去讨论高阶无穷小项的性质。感兴趣的读者可以去查阅相关的专业书籍。若把x_0 设置为0，则函数在零点展开，称为麦克劳林展开。以下给出一些常用函数的麦克劳林展开公式：

$$e^x = 1 + x + \frac{1}{2!}x^2 \cdots + \frac{1}{n!}x^n + R_n(x) \tag{29}$$

$$\sin x = x - \frac{1}{3!}x^3 \cdots + \frac{x^n}{n!}\sin\frac{n\pi}{2} + R_n(x) \tag{30}$$

$$\cos x = 1 - \frac{1}{2!}x^2 + \frac{x^n}{n!}\cos\frac{n\pi}{2} + R_n(x) \tag{31}$$

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \cdots + (-1)^{n-1}\frac{x^n}{n} + R_n(x) \tag{32}$$

$$(1+x)^m = 1 + mx + \frac{m(m-1)}{2!} + \cdots + \frac{m(m-1)\cdots(m-n+1)}{(n+1)!}x^{n+1} + R_n(x) \tag{33}$$

1.5 函数的性质

1.5.1 单调性

函数的单调性指的是函数在一个区间内的变化趋势。对于函数单调性的判定，有如下的定理：

定理1：设函数$f(x)$在(a,b)区间内可导，如果对$\forall x \in (a,b)$，都有$f'(x) > 0$（或$f'(x) < 0$），则函数$f(x)$在(a,b)内是单调增加的（或单调减少）。当函数的一阶导数在x点处取0时，则函数可能在x处取极值。对于取极值的必要条件，有如下的定理：

定理2：设函数$f(x)$在x_0处可导，且在x_0处取极值，则$f'(x) = 0$。

定理3：设函数$f(x)$在x_0的某一邻域内可微，且$f'(x) = 0$ （或$f(x)$在x_0处连续，但$f'(x_0)$不存在）。

（1）若当x经过x_0时，$f'(x)$由”$+$”变”$-$” 则$f(x_0)$为极大值；

（2）若当x经过x_0时，$f'(x)$由”$-$”变”$+$” 则$f(x_0)$为极大值；

（3）若当x经过x_0时，$f'(x)$不变号，则$f(x_0)$不是极值。

定理4：设$f(x)$在点x_0处有$f''(x) \neq 0$，且$f'(x_0) = 0$，则当$f''(x_0) < 0$时，$f(x_0)$为极大值；当$f''(x_0) > 0$时，$f(x_0)$ 为极小值。

1.5.2 凹凸性

函数的凹凸性在不同的教科书里，有不同的定义规则，为了统一起见，我们再此给出凹凸性的定义，之后的章节，都按照此定义来规定函数的凹凸性。

定理1：若在定义域(a,b)上，$f''(x) < 0$ （或$f''(x) > 0$），则$f(x)$在(a,b)上是凸函数（或是凹函数）。

定理2：若在x_0处$f''(x) = 0$，（或$f''(x)$不存在），当x 变动经过x_0时。$f''(x)$变号，则$(x_0, f(x_)))$ 为拐点。

定理2说明了拐点的定义为函数在该点处，凹凸性发生了改变。

定理3：设$f(x)$在x_0点的某邻域内有三阶导数，且$f''(x) = 0$，$f'''(x) \neq 0$，则$(x_0, f(x_0)))$ 为拐点。

1.6 多元函数的微分规则

对于函数的自变量不止一个时，即两个或两个以上时，会出现全导数和偏导数的区别。下面简单介绍一下多元函数的导数定义，关于更细致的性质，放在线性代数里面的矩阵微分规则处。

二元函数关于x的偏导，只需要模仿一元函数导数的定义即可。这里把y 看成常量。对于更多自变量的偏导数定义可仿照下式。对x的偏导数：

$$\frac{\partial f(x,y)}{\partial x} = f_x(x,y) = \lim_{\Delta x \to 0}\frac{f(x+\Delta x, y) - f(x,y)}{\Delta x} \tag{34}$$

同理，大家可以得出$f(x,y)$关于y的偏导。再有了偏导数的概念后，我们可以继续定义方向导数，即函数在某一个方向上的变化率，对于给定方向L，该方向上有两点A 和B。我们对该方向求变化率，有如下的计算公式：

$$\frac{\partial f}{\partial L} = f_x \cos\alpha + f_y \cos\beta \tag{35}$$

其中$\cos\alpha = \frac{\overrightarrow{AB_x}}{|\overrightarrow{AB}|}$，$\cos\beta = \frac{\overrightarrow{AB_y}}{|\overrightarrow{AB}|}$，$|\overrightarrow{AB}|$为向量$|\overrightarrow{AB}|$的模。
对于方向导数，总有一个方向，是沿着该方向，函数的变化率最大，我们把这个方向上的变化率称为梯度，梯度的方向是沿着函数增加最快的方向。对于一个三元函数，可用如下的记号来表示梯度：

$$\nabla f(x,y,z) \equiv grad\ f(x,y,z) \equiv (f_x, f_y, f_z) \tag{36}$$

其中∇为梯度算符，在不同的坐标系下，有不同的表示式，再此我们给出在三维直角坐标系下的表达式$\nabla = \frac{\partial}{\partial x}\overrightarrow{i} + \frac{\partial}{\partial y}\overrightarrow{j} + \frac{\partial}{\partial z}\overrightarrow{k}$，可以看出梯度是一个向量，作用在标量函数上，给出标量函数的最大变化率方向。
对于一个多元函数，还可以定义全微分，对于三元函数，全微分的定义如下：

$$df(x,y,z) = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz \tag{37}$$

最后我们要指出，对于多元函数二阶混合偏导数，不同的微分变量，可以相互交换次序而不影响偏导数的值。即

$$\frac{\partial^2 f(x,y)}{\partial x \partial y} = \frac{\partial^2 f(x,y)}{\partial y \partial x} \tag{38}$$

同样以二元函数为例，相关的证明推导，我们给出来，感兴趣的可以看一下。

$$\begin{aligned}
\frac{\partial^2 f(x,y)}{\partial x \partial y} &= \frac{\partial}{\partial x}\frac{\partial f}{\partial y} && (39)\\
&= \frac{\partial}{\partial x} f'_y && (40)\\
&= \lim_{\Delta x \to 0} \frac{f'_y(x+\Delta x, y) - f'_y(x,y)}{\Delta x} && (41)\\
&= \lim_{\Delta x \to 0} \frac{\lim_{\Delta y \to 0} \frac{f(x+\Delta x, y+\Delta y) - f(x+\Delta x, y)}{\Delta y} - \lim_{\Delta y \to 0} \frac{f(x, y+\Delta y) - f(x,y)}{\Delta y}}{\Delta x} && (42)\\
&= \lim_{\Delta x \to 0} \lim_{\Delta y \to 0} \frac{[f(x+\Delta x, y+\Delta y) - f(x+\Delta x, y)] - [f(x, y+\Delta y) - f(x,y)]}{\Delta x \Delta y} && (43)\\
&= \lim_{\Delta x \to 0} \lim_{\Delta y \to 0} \frac{[f(x+\Delta x, y+\Delta y) - f(x, y+\Delta y)] - [f(x+\Delta x, y) - f(x,y)]}{\Delta x \Delta y} && (44)\\
&= \lim_{\Delta y \to 0} \frac{\lim_{\Delta x \to 0} \frac{f(x+\Delta x, y+\Delta y) - f(x, y+\Delta y)}{\Delta x} - \lim_{\Delta x \to 0} \frac{f(x+\Delta x, y) - f(x,y)}{\Delta x}}{\Delta y} && (45)\\
&= \lim_{\Delta y \to 0} \frac{f'_x(x, y+\Delta y) - f'_x(x,y)}{\Delta y} && (46)\\
&= \frac{\partial}{\partial y} f'_x && (47)\\
&= \frac{\partial^2 f(x,y)}{\partial y \partial x} && (48)
\end{aligned}$$

1.7 指示函数

这里介绍一种在机器学习当中常用的特殊函数，我们定义一个映射，当一个逻辑表达式的值为真时，映射为1，否则为0，表达式如下：

$$\mathbb{I}\{f(x) = y\} = 1 \tag{49}$$

这里$\mathbb{I}$是指示函数的记号，当$f(x)=y$时，指示函数取1，否则为0。

2 线性代数

线性代数部分取自斯坦福课程CS229的讲义[2]，摘取了其中比较重要的，和比较常用的一些知识，来进行回顾一下。主要包括四个方面：基本概念与记号，矩阵的运算规则，矩阵的一些操作与性质，矩阵的微分规则。

2.1 基本概念与记号

对于一个矩阵来说，我们一般使用大写字母来表示，例如使用$A \in \mathbb{R}^{m\times n}$来表示一个矩阵属于实数域，即矩阵的每一个元素都是实数，其中$m \times n$表示矩阵A有m行和n列。
矩阵可以看成是一个二维数组，矩阵的特殊情况是，只有一个维度，我们把它称为向量，记为$x \in \mathbb{R}^n$，用它来表示一个n维向量，不同的文献对向量有不同的记法，我们再此为了教学方便起见，默认向量都是**列向量**。列向量的转置是**行向量**，记为x^T。对于向量里面任意一个元素用下角标来表示，x_i表示向量x的第i个元素。

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{50}$$

对于矩阵里面的元素，一般用A_{ij}或a_{ij}来表示，其中i表示行数，j 表示列数。

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m1} & \cdots & a_{mn} \end{bmatrix} \tag{51}$$

在生活中计数都从1开始，但是在计算机语言里，大部分从0 开始，比如python，c，cpp，java等是从0开始计数，而$matlab$语言是从1 开始。我们的课程使用python语言进行模型和算法的展示，所以在编程时要注意与生活区分。对于矩阵的第j列，使用a^j或者$A_{:,j}$来表示：

$$A = \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^n \\ | & | & & | \end{bmatrix} \tag{52}$$

对于矩阵的某一行，使用a_i^T或者$A_{i,:}$来表示：

$$A = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_n^T & - \end{bmatrix} \tag{53}$$

我们再次强调，不同的教科书，记号会有差别，所以在阅读论文时，要格外注意。

2.2 矩阵运算规则

首先定义矩阵相乘，对于一个$A \in \mathbb{R}^{m\times n}$和$B \in \mathbb{R}^{n\times p}$，有如下的定义：

$$C = AB \in \mathbb{R}^{m\times p} \tag{54}$$

其中矩阵C的元素$C_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}$给出，通俗来讲，就是前面的矩阵用第$i$行的元素与后面的矩阵第$i$ 列的元素，一一对应相乘后求和。

对于向量来说，同样可以定义乘法，向量的乘法称为内积（inner product），又叫点乘（dot product）。对于$x, y \in \mathbb{R}^n$ 有如下的表达式：

$$x^T y \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i \tag{55}$$

与内积相对应的是外积，外积操作不需要两个向量的维度相等，对于两个向量$x \in \mathbb{R}^m, y \in \mathbb{R}^n$，外积有如下的表达式：

$$xy^T \in \mathbb{R}^{m\times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_1 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix} \tag{56}$$

作为一个例子，假如有一个矩阵，它的每一列都是相同的，我们写成这个矩阵的一列与一个全是1的行向量相乘，把这个全是1的向量记为$\mathbf{1} \in \mathbb{R}^n$，具体表达式如下：

$$A = \begin{bmatrix} | & | & & | \\ x & x & \cdots & x \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_1 & x_1 & \cdots & x_1 \\ x_2 & x_2 & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_m & x_m & \cdots & x_m \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} = x\mathbf{1}^T \tag{57}$$

对于矩阵与向量相乘，矩阵与矩阵相乘，都可以通过矩阵的行向量与列向量表示，这样做的好处是，直接在向量层面上进行数值计算，替代了在元素上进行的计算。具体的细节再此不再赘述，感兴趣的同学可以去查看CS229的讲义。

矩阵的乘法，满足结合律与分配律，不满足交换律，给出如下的公式进行说明：

$$(AB)C = A(BC) \tag{58}$$

$$A(B+C) = AB + AC \tag{59}$$

$$AB \neq BA \tag{60}$$

2.3 矩阵的操作与性质

2.3.1 单位矩阵与对角矩阵

单位矩阵$I \in \mathbb{R}^{n\times n}$给出如下的定义:

$$I_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{61}$$

对于任意一个矩阵，与单位矩阵相乘都等于它本身。

$$AI = IA = A \tag{62}$$

对角矩阵的定义为只有对角元存在非零元素，其余都为零元素。一般写成$D = diag(d_1, d_2, ..., d_n)$，也可写成如下的表达式：

$$D_{ij} = \begin{cases} d_i, & i = j \\ 0, & i \neq j \end{cases} \tag{63}$$

特殊的情况下，$I = diag(1, 1, ..., 1)$。

2.3.2 矩阵转置

所谓转置（transpose），是对矩阵的一种操作，把矩阵沿着对角线进行翻转。一个$A \in \mathbb{R}^{m \times n}$，它的转置定义为：

$$(A^T)_{ij} = A_{ji} \tag{64}$$

转置之后，矩阵的维度发生对调，即$A^T \in \mathbb{R}^{n \times m}$。转置操作有如下的性质：

$$(A^T)^T = A \tag{65}$$

$$(AB)^T = B^T A^T \tag{66}$$

$$(A + B)^T = A^T + B^T \tag{67}$$

2.3.3 对称矩阵

方阵的定义为，行与列的维数相同。即$A \in \mathbb{R}^{n \times n}$。在方阵的基础上，可以定义对称阵（symmetric）与反对称阵（anti-symmetric）。对称阵为$A^T = A$，反对称阵为$A^T = -A$，通过定义可以看出，反对称阵的对角元一定为0。对于任意一个方阵，都可以改写成一个对称阵和一个反对称的和，即：

$$A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T) \tag{68}$$

对于一个对称矩阵，我们一般写成$A \in \mathbb{S}^n$这种形式，$\mathbb{S}$代表“对称”这个单词的首字母。通过这种分解，可以简化运算。

2.3.4 矩阵的迹

一个方阵的迹（trace）的定义为：

$$trA = \sum_{i=1}^{n} A_{ii} \tag{69}$$

根据定义，对于方阵的求迹运算，有如下的性质：

$$trA = trA^T \tag{70}$$

$$tr(A + B) = trA + trB \tag{71}$$

$$tr(\alpha A) = \alpha trA \tag{72}$$

$$trAB = trBA \tag{73}$$

$$trABC = trBCA = trCAB \tag{74}$$

因为一个实数的迹就等于它本身，对实数求迹，然后运用轮换关系可以简化计算。

2.3.5 矩阵的范数

所谓范数，可以理解为一个向量的长度，这个长度是广义上的长度。特殊的，一个向量的二范数ℓ_2为

$$||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{75}$$

可以看出向量x的二范数的平方，对应着向量自己的内积，即$||x||_2^2 = x^T x$。
范数的定义，本质上是一个映射，即$f : \mathbb{R}^n \to \mathbb{R}$要求满足如下的四个性质：

- $\forall x \in \mathbb{R}^n, f(x) \geq 0$.非负性。
- $f(x) = 0$ only and only if $x = 0$.确定性。
- $\forall x \in \mathbb{R}^n, t \in \mathbb{R}, f(tx) = tf(x)$.同质性。
- $\forall x, y \in \mathbb{R}^n, f(x + y) \leq f(x) + f(y)$.三角不等式。

经常使用的还有一范数ℓ_1为：

$$||x||_1 = \sum_{i=1}^{n} |x_i| \tag{76}$$

还有无穷范数ℓ_∞为：

$$||x||_\infty = \max_i |x_i| \tag{77}$$

通过以上的例子，我们可以给出一般范数的定义，对于一个p范数（$p \geq 1$）的定义为：

$$||x||_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p} \tag{78}$$

对于矩阵，同样可以定义范数，在数学上，对矩阵的范数定义有很多，其中包括谱范数，核范数，Frobenius范数，与机器学习相关最常用的为Frobenius范数，定义如下：

$$||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2} = \sqrt{tr(A^T A)} \tag{79}$$

2.3.6 线性独立与秩

在一个向量的集合$\{x_1, x_2, ...x_n\} \in \mathbb{R}^m$中，如果任意一个向量都不能通过剩余的向量通过线性组合表示出来，则我们称这个集合内的向量彼此**线性无关**。若可以表示，则称为**线性有关**。可以写成：

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i \tag{80}$$

其中$\alpha_i \in \mathbb{R}$。
所谓的秩表示的是线性无关向量的最大个数。对于一个矩阵来说，秩又分为列秩与行秩。顾名思义，列秩表达的是一个这个矩阵的所有列，最多有多少列是线性无关的，行秩表达的是这个矩阵的所有行，最多有多少行是线性无关的。对于任意一个矩阵$A \in \mathbb{R}^{m \times n}$，可以证明，行秩与列秩相等。对于秩有如下的一些基本性质：

- $\forall A \in \mathbb{R}^{m \times n}$，$rank(A) \leq \min(m, n)$. 如果$rank(A) = \min(m, n)$，我们称矩阵$A$是满秩的。
- $\forall A \in \mathbb{R}^{m \times n}$，$rank(A) = rank(A^T)$.
- $\forall A \in \mathbb{R}^{m \times n}$，$\forall B \in \mathbb{R}^{n \times p}$,$rank(AB) \leq \min(rank(A), rank(B))$.
- $\forall A, B \in \mathbb{R}^{m \times n}$，$rank(A + B) \leq rank(A) + rank(B)$.

2.3.7 方阵的逆

一个方阵$A \in \mathbb{R}^{n \times n}$的逆，标记为$A^{-1}$。并且方阵的逆是唯一的，方阵与它自己的逆相乘为单位矩阵：

$$A^{-1}A = I = AA^{-1} \tag{81}$$

并不是所有的矩阵都具有逆。首先，对于一般的矩阵，不是方阵的矩阵，是没有逆的。其次方阵也可能没有逆，对于没有逆的方阵，我们把它称为不可逆方阵，或者奇异方阵。一个方阵要想有逆，充分必要条件是方阵满秩。当然还有很多条件来限制方阵的逆，再此我们就不在赘述了。下面我们看一下方阵的逆的一些性质：

- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1}$，我们一般把矩阵同时转置和取逆记为A^{-T}

以上的运算前提条件是$A, B \in \mathbb{R}^{n \times n}$方阵的逆都存在。

2.3.8 正交矩阵

先说一下两个向量的正交，所谓的正交，一种情况是，在平面直角坐标系下两个向量垂直，即两个向量点乘$\overrightarrow{a}\overrightarrow{b} = 0$，垂直记为$\perp$。在线性代数中，同样有正交的概念，即两个向量做内积，结果是0，公式表达为$x^T y = 0$。另一个概念是归一化，所谓的归一化，就是单位化，对于一个向量来说，就是向量的二范数为1，即$||x||_2 = 1$。正交矩阵的定义为，对于一个方阵$U \in \mathbb{R}^{n \times n}$，它的每列都是归一化的，并且任意两列向量都是正交的。根据这个定义我们可以立刻写出如下的等式：

$$U^T U = I = UU^T \tag{82}$$

换句话说，对于正交矩阵，它的逆等于它的转置，$U^{-1} = U^T$。如果这个矩阵不是方阵，但是它的任意两列都是归一且正交的，这样的矩阵满足如下的性质：$U^T U = I$，但是$UU^T \neq I$。我们一般只讨论矩阵是方阵的情况。

正交矩阵有一个很好的性质，对于任意的向量$x \in \mathbb{R}^n, U \in \mathbb{R}^{n \times n}$正交矩阵乘以一个向量，并不会改变该向量的二范数，即：

$$||Ux||_2 = ||x||_2 \tag{83}$$

2.3.9 矩阵的非零空间与范围

一组向量$\{x_1, x_2, ...x_n\}$构成的集合，我们把它们的线性组合，称为这组向量的扩张空间（span space）。定义如下：

$$span(\{x_1, x_2, ...x_n\}) = \{v : c = \sum_{i=1}^{n} \alpha_i x_i, \alpha_i \in \mathbb{R}\} \tag{84}$$

可以证明，如果这组向量是全是线性无关的，那么任意一个向量$v \in \mathbb{R}^n$都可以用$\{x_1, x_2, ...x_n\}$进行展开。如果这组向量里部分向量线性相关，那么我们可以定义一个向量的投影（projection）。投影的意思是对于一个向量$y \in \mathbb{R}^m$，从$\{x_1, x_2, ...x_n\}$构成的扩张空间里（$x_i \in \mathbb{R}^m$），找到一个向量v，使得v与y之间的欧式距离最短。我们给投影操作如下的记号：

$$Proj(y; \{x_1, x_2, ...x_n\}) = \arg \min_{v \in span(\{x_1, x_2, ...x_n\})} ||y - v||_2 \tag{85}$$

对于一个矩阵$A \in \mathbb{R}^{m \times n}$,它的范围（range），有时候也称为矩阵的列空间（columnspace），记为$R(A)$。实质就是用矩阵的列向量张成的一个空间，定义如下：

$$R(A) = \{v \in \mathbb{R}^m : v = Ax, x \in \mathbb{R}^n\} \tag{86}$$

如果矩阵A是一个满秩矩阵，那么任意一个向量y在矩阵A的列空间构成的投影，有如下的表达式：

$$Proj(y; A) = \arg\min_{v \in R(A)} ||v - y||_2 = A(A^T A)^{-1} A^T y \tag{87}$$

这个式子在推导线性回归的正规化方程时，我们会给出论证。特别地，当矩阵A 是一个向量时，我们可以简化投影操作的运算：

$$Proj(y; a) = \frac{aa^T}{a^T a} y \tag{88}$$

一个矩阵的零空间（nullspace）是这个矩阵构成的齐次方程组的解的集合，定义如下：

$$N(A) = \{x \in \mathbb{R}^n : Ax = 0\} \tag{89}$$

2.3.10 行列式

一个方阵的行列式（determinant）是一个映射：$\mathbb{R}^{n \times n} \to \mathbb{R}$，记号写为$|A|$或者$detA$，对于行列式，我们先给一个直觉上的认识，在给出它的计算公式。
给定一个方阵$A \in \mathbb{R}^{2 \times 2}$，假设它的取值为：

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \tag{90}$$

这个矩阵的每个列向量在二维平面直角坐标系下，构成了一个平行四边形，这个平行四边形的面积，通过计算可以得知是1。假如是一个三维的方阵，那么在空间直角坐标系下，会构成一个平行四面体，这个平行四面体的体积我们也是可以计算得到的。推广下去，对于一个n维的方阵，会存在一个n维的坐标空间，在这个空间下，方阵的每一个列向量，会构成一个超平行四面体，我们无法通过作图得到，但是我们知道，这个超平行四面体，同样具备所谓的体积属性（从量纲的角度来说，不是体积）。
从面积，体积的角度考虑，我们认为行列式应该具备面积或者体积的特征，所以，

- 对于单位立方体，它的行列式的值应该为1，即$|I| = 1$。
- 对于任意一个超平行四面体，扩大它的边k倍，它的体积属性也会随着扩大k倍。即：

$$\left| \begin{bmatrix} - & ka_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} \right| = k|A| \tag{91}$$

- 如果交换了两个向量的位置，计算得到的行列式的值，应该会和原来的相比差一个符号。即：

$$\left| \begin{bmatrix} - & a_2^T & - \\ - & a_1^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} \right| = -|A| \tag{92}$$

除了以上三个很直觉的性质外，行列式还具有如下的性质：

- $\forall A \in \mathbb{R}^{n\times n}, |A| = |A^T|$。
- $\forall A, B \in \mathbb{R}^{n\times n}, |AB| = |A||B|$。
- $\forall A \in \mathbb{R}^{n\times n}, |A| = 0$当且仅当$A$是奇异方阵。
- $\forall A \in \mathbb{R}^{n\times n}$并且$A$是满秩矩阵，则$|A^{-1}| = 1/|A|$。

在定义一个矩阵的行列式的计算公式前，我们先给一个记号，用来标记对一个抽掉它的第i行和第j列，记为$A_{\backslash i,\backslash j} \in \mathbb{R}^{(n-1)\times(n-1)}$。对于任意一个方阵的行列式，计算方法如下：

$$|A| = \sum_{i=1}^{n}(-1)^{i+j}a_{ij}|A_{\backslash i,\backslash j}| \ (for\ any\ j \in 1, 2...n) \tag{93}$$

$$= \sum_{j=1}^{n}(-1)^{i+j}a_{ij}|A_{\backslash i,\backslash j}| \ (for\ any\ i \in 1, 2...n) \tag{94}$$

在给出行列式的具体计算公式后，进而可以引入一个伴随矩阵（adjoint matrix）的概念，它是为了计算矩阵的逆引入的一个中间概念，具体的表达式如下：

$$adj(A) \in \mathbb{R}^{n\times n}, (adj(A))_{ij} = (-1)^{i+j}|A_{\backslash j,\backslash i}| \tag{95}$$

对于一个不是奇异的方阵，我们可以通过如下的公式，计算矩阵的逆：

$$A^{-1} = \frac{1}{|A|}adj(A) \tag{96}$$

2.3.11 二次型与半正定矩阵

给定一个方阵$A \in \mathbb{R}^{n\times n}$和一个向量$x \in \mathbb{R}^n$，标量值$x^TAx$被称为**二次型**。展开写，具有如下的形式：

$$x^TAx = \sum_{i=1}^{n}x_1(Ax)_i = \sum_{i=1}^{n}x_i(\sum_{j=1}^{n}A_{ij}x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij}x_ix_j \tag{97}$$

我们可以看到，

$$x^TAx = (x^TAx)^T = x^TA^Tx = x^T(\frac{1}{2}A + \frac{1}{2}A^T)x \tag{98}$$

第一个等号成立的条件是，对于一个标量来说，转置等于它本身，第二个等号是因为把转置操作吸收到括号里面，所以我们看到二次型的矩阵A只有对称部分才会提供有效的作用。一般默认情况下，我们认为矩阵A是对称矩阵。下面给这个矩阵A一些限制，在这些限制下，引出一些定义：

- 当对任意的非零向量x，都有$x^TAx > 0$，对称矩阵$A \in \mathbb{S}^n$是正定（positive definite）矩阵，经常记为$A \succ 0$或直接记为$A > 0$。对于所有正定矩阵构成的集合，用记号$\mathbb{S}^n_{++}$来表示。
- 当对任意的非零向量x，都有$x^TAx \geq 0$，对称矩阵$A \in \mathbb{S}^n$是半正定（positive semi-definite）矩阵，经常记为$A \succeq 0$ 或直接记为$A \geq 0$。对于所有正定矩阵构成的集合，用记号$\mathbb{S}^n_{+}$来表示。
- 当对任意的非零向量x，都有$x^TAx < 0$，对称矩阵$A \in \mathbb{S}^n$是负定（negative definite）矩阵，经常记为$A \prec 0$ 或直接记为$A < 0$。
- 当对任意的非零向量x，都有$x^TAx \leq 0$，对称矩阵$A \in \mathbb{S}^n$是半负定（negative semi-definite）矩阵，经常记为$A \preceq 0$ 或直接记为$A \leq 0$。
- 对称矩阵$A \in \mathbb{S}^n$是不定（indefinite definite）矩阵，也就是说，存在一些x向量使得$x^TAx > 0$，还存在一些向量使得$x^TAx < 0$。

我们可以看出如果A是正定的，那么$-A$是负定的，反之亦然。对于正定或者负定矩阵还有一个很重要的性质，就是它们都是满秩矩阵。我们可以采用反证法来进行证明，假设正定或负定矩阵A不是满秩矩阵，那么一定存在

$$a_j = \sum_{i \neq j} x_i a_i \tag{99}$$

存在$x_1, ...x_{j-1}, x_{j+1}, ...x_n \in \mathbb{R}$，我们令$x_j = -1$，那么有：

$$Ax = \sum_{i=1}^{n} x_i a_i = \sum_{i \neq j} x_i a_i + x_j a_j = a_j - a_j = 0 \tag{100}$$

这意味着存在一组系数，使得$x^T Ax = 0$，但是我们的前提条件是A是正定或者负定的，即$x^T Ax \neq 0$，所以我们论证了正定或者负定矩阵一定是满秩矩阵。
最后，正定矩阵使我们经常提到的，给任意一个矩阵$A \in \mathbb{R}^{m \times n}$，不要求这个矩阵是对称或者方阵，矩阵$G = A^T A$ 总是半正定的，矩阵G被称为Gram矩阵。进一步来说，若矩阵A的$m \geq n$且是满秩矩阵，那么矩阵A的Gram矩阵是正定的。

2.3.12 特征值与特征向量

给定一个方阵$A \in \mathbb{R}^{n \times n}$，我们说$\lambda \in \mathbb{C}$是特征值（eigenvalues），$x \in \mathbb{C}$是特征向量（eigenvectors），（$\mathbb{C}$是复数域），当满足如下的方程：

$$Ax = \lambda x, \ x \neq 0 \tag{101}$$

通过定义可以看出，对于特征向量可以进行放缩，比如特征向量$x \in \mathbb{C}$，有一个标量$t \in \mathbb{C}$，$A(tx) = tAx = t\lambda x = \lambda(tx)$，即$tx$依旧是矩阵$A$的特征向量，所以我们谈到特征向量一般认为是归一化的$||x||_2 = 1$。这里面还牵扯到符号的区别，我们认为符号已经吸收到特征值λ里面了。
把方程移项，我们得到如下的表达式：

$$(\lambda I - A)x = 0, \ x \neq 0 \tag{102}$$

我们说当方程存在非零解时，当且仅当$(\lambda I - A)$是非零空间，否则矩阵$(\lambda I - A)$就是奇异矩阵，奇异矩阵具有行列式为0的性质，即：

$$|(\lambda I - A)| = 0 \tag{103}$$

对于一个行列式展开定理进行展开，系数就是特征值λ，感兴趣的可以查阅相关的书籍。当解出来方程后，会得到$\lambda_1, \lambda_2, ...\lambda_n$共$n$个特征根，这$n$个特征根并不知道与哪个特征向量对应，所以还需要进一步解$(\lambda_i I - A)x = 0$解出来λ_i所对应的特征向量x。具体如何解这个方程，是一个很困难的问题，我们再此不在过多涉及。目前可以通过计算机语言（python，cpp等），使用一行代码，就轻松得到方程的解。
我们下面介绍一下特征值与特征向量的一些性质：

- 对矩阵A求迹，等价于A的特征值求和，$trA = \sum_{i=1}^{n} \lambda_i$。
- 对矩阵A求行列式的值，等价于A的特征值连乘，$detA = \prod_{i=1}^{n} \lambda_i$。
- 矩阵A的秩，为矩阵A的非零特征值的个数。
- 假定A是非奇异矩阵，那么我们有$1/\lambda$是矩阵A^{-1}的特征值，即$A^{-1}x = (1/\lambda)x$。（证明很简单，对原方程左乘A^{-1} 得到结果）
- 对于一个对角矩阵，$D = diag(d_1, ..., d_n)$，它的特征值就是对角元的元素。

2.3.13 对称矩阵的特征值与特征向量

一般情况下，对于一个矩阵一般的方阵，它的特征值和特征向量的结构是很微妙的（subtle）。但是幸运的，在机器学习里面，我们遇到一般都是对称矩阵的实矩阵，它的性质是很显然的。这一部分，我们假设矩阵A是一个实的对称阵，我们会得到如下的性质：

- 矩阵A的所有特征值都是实数，我们用$\lambda_1, \lambda_2, ...\lambda_n$来标记。
- $\forall i, Au_i = \lambda_i u_i, ||u_i||_2 = 1$，并且$\forall j \neq i, u^T u_j = 0$。（注意，会存在一种情况，就是特征值是重复的，那么它们对应的特征向量也是重复的，就不在满足上述的条件）

我们令U是一个正交矩阵，它的列向量用u_i来表示，即：

$$U = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{bmatrix} \tag{104}$$

令$\Lambda = diag(\lambda_1, ..., \lambda_n)$是对角矩阵，对角元是矩阵$A$的特征值。我们可以得到如下的等式：

$$AU = \begin{bmatrix} | & | & & | \\ Au_1 & Au_2 & \cdots & Au_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \lambda_1 u_1 & \lambda_2 u_2 & \cdots & \lambda_n u_n \\ | & | & & | \end{bmatrix} = U diag(\lambda_1, ..., \lambda_n) = U\Lambda \tag{105}$$

又因为矩阵U是正交矩阵，具有$U^T U = UU^T = I$的性质，所以上式等式两边同时右乘U^T，得到如下的方程：

$$A = AUU^T = U\Lambda U^T \tag{106}$$

矩阵A得到了一个新的表达式$U\Lambda U^T$，我们把这个方程，称为矩阵A的对角化，通过一个简单的对角矩阵来表示矩阵A，对于正交矩阵U，它的每一个列向量对应着矩阵A的特征向量，这中变化形式非常重要，不仅在机器学习里面，我们要对这样的变换求极值，而且在量子力学的矩阵表达里，也是举足轻重的。关于具体的应用，我们再此就不过多介绍了，读者可以去查看CS229的课程讲义，或者相关的线性代数的书籍。

2.4 矩阵微分规则

在高等数学部分，我们已经介绍了微分的相关知识，这一部分，由于我们介绍了线性代数，所以两者结合，就构成了矩阵微分。我们下面简单的介绍一些概念，方便我们后来的教学。

2.4.1 梯度

梯度的概念我们之前已经介绍过了，现在假设一个标量函数的自变量是一个矩阵，即存在着映射关系：$f : \mathbb{R}^{m \times n} \to \mathbb{R}$，那么我们把对函数$f$的梯度记为如下的表达式：

$$\nabla_A f(A) \in \mathbb{R}^{m \times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix} \tag{107}$$

换句话说，一个自变量为矩阵的标量函数，它的梯度为一个矩阵，矩阵元为：

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}} \tag{108}$$

通过定义可以看出，函数自变量的维度是多少，梯度后的维度也是多少。特别地，当矩阵A是一个向量时$x \in \mathbb{R}^n$，我们有：

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \tag{109}$$

我们要注意的是，我们只定义了对标量函数的梯度，即函数最终映射出的是一个实数。没有定义对矢量函数，或更高阶的张量函数的梯度，比如说，我们无法计算Ax 的梯度，因为$Ax \in \mathbb{R}^{n\times 1}$。
下面我给出两个基本性质：

- $\nabla_x(f(x) + g(x)) = \nabla_x f(x) + \nabla_x g(x)$
- $\forall t \in \mathbb{R}, \nabla_x(tf(x)) = t\nabla_x f(x)$

最后我们需要强调一点，梯度算符∇是一个微分操作，对于一个函数，可能会有不同的自变量，所以在使用∇时，一定要严谨，在右下角加上角标，到底是对谁求梯度。

2.4.2 海塞矩阵

当自变量是向量$x \in \mathbb{R}^n$时，我们不仅可以求函数的一阶梯度，还可以求函数的二阶梯度，我们把对标量函数的二阶梯度称作海塞矩阵（Hessian matrix），具有如下的表达式：

$$\nabla_x^2 f(x) \in \mathbb{R}^{n\times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} \tag{110}$$

换句话说，$\nabla_x^2 f(x) \in \mathbb{R}^{n\times n}$的矩阵元为：

$$(\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \tag{111}$$

由于在之前介绍高等数学部分，我们已经证明了函数的二阶偏微分是可以对微分元进行次序交换，所以海塞矩阵是一个对称矩阵，同样海塞矩阵也是定义在标量实值函数上的。关于海塞矩阵的推导，我们在这就不在涉及了，可以采用先求一阶导数，再求一阶导数的导数获得。

最后我们要强调，我们是在标量函数的自变量是向量的情况下定义的海塞矩阵。如果自变量是矩阵的情况下，我们依然可以定义海塞矩阵，当然矩阵的每一个元素应该是$\frac{\partial^2 f(A)}{\partial A_{ij} \partial A_{kl}}$，可以看出这是一个具有四个指标的量$ijkl$，一般把这样的量称为张量（tensor）。在课程当中，我们并不会去讨论它。

2.4.3 二次形的梯度与海塞矩阵

现在我们试图在一些简单的函数上计算函数的梯度与海塞矩阵。首先我们讨论一个线性函数，$\forall x \in \mathbb{R}^n$，设$f(x) = b^T x$并且向量$b \in \mathbb{R}^n$已知。那么函数用分量表达的形式如下：

$$f(x) = \sum_{i=1}^{n} b_i x_i \tag{112}$$

对函数的其中一个自变量求偏微分，得到如下的表达式：

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k \tag{113}$$

从上面的分量表达式就可以看出，$\nabla_x b^T x = b$。我们可以对应单变量的函数微分规则$\frac{\partial}{\partial x}(ax) = a$。
下面我们来考虑二次型函数$f(x) = x^T Ax$，其中矩阵$A \in \mathbb{S}^n$。同样先写出来元素的展开式：

$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j \tag{114}$$

我们先计算函数的一阶偏导数的表达式，给出如下的推导：

$$\begin{aligned}
\frac{\partial f(x)}{\partial x_k} &= \frac{\partial}{\partial x_k}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j & (115)\\
&= \frac{\partial}{\partial x_k}[\sum_{i\neq k}\sum_{j\neq k} A_{ij}x_i x_j + \sum_{i\neq k} A_{ik}x_i x_k + \sum_{j\neq k}\sum_{j\neq k} A_{kj}x_k x_j + A_{kk}x_k^2] & (116)\\
&= \sum_{i\neq k} A_{ik}x_i + \sum_{j\neq k} A_{kj}x_j + 2A_{kk}x_k & (117)\\
&= \sum_{i=1}^{n} A_{ik}x_i + \sum_{j=1}^{n} A_{kj}x_j = 2\sum_{i=1}^{n} A_{ki}x_i & (118)
\end{aligned}$$

最后一个等号成立，是因为我们已知矩阵A是对称矩阵。由分量表达式，进而可以写出整体的表达式，即$\nabla_x x^T Ax = 2Ax$。我们发现和自变量为标量的函数求导规则具有相同的形式：$\frac{\partial}{\partial x}(ax^2) = 2ax$。
我们继续计算二阶偏微分的元素值，即：

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_l} = \frac{\partial}{\partial x_k}[2\sum_{i=1}^{n} A_{li}x_i] = 2A_{lk} = 2A_{kl} \tag{119}$$

因此我们得到了$\nabla_x^2 x^T Ax = 2A$，同样的我们也可以去和自变量为标量的二次函数对照，形式依旧相同。我们总结一下：

- $\nabla_x b^T x = b$
- $\nabla_x x^T Ax = 2Ax$ (A是对称矩阵)
- $\nabla_x^2 x^T Ax = 2A$ (A是对称矩阵)

2.4.4 最小二乘法

最小二乘法（least square）其实是求一个平方函数的极值。给定一个矩阵$A \in \mathbb{R}^{m\times n}$，且$A$是一个满秩矩阵，还有一个向量$b \in \mathbb{R}^m$，且$b \notin R(A)$。一种情况下，当我们计算$Ax = b$时，我们不能得到$x$的精确解，我们只能期待$Ax$ 与b 尽可能的接近，我们采用欧几里得距离来进行测量两者之间的差距，即$||Ax - b||_2^2$。

我们先把这个平方项展开，得到如下的表达式：

$$||Ax - b||_2^2 = (Ax - b)^T(Ax - b) = x^T A^T Ax - 2b^T Ax + b^T b \tag{120}$$

对上面的方程，求变量x的梯度，得到：

$$\nabla_x(x^T A^T Ax - 2b^T Ax + b^T b) = \nabla_x x^T A^T Ax - \nabla_x 2b^T Ax + \nabla_x b^T b = 2A^T Ax - 2A^T b \tag{121}$$

最后，我们令梯度为0，求得函数的极值点，得到：

$$x = (A^T A)^{-1} A^T b \tag{122}$$

之后我们还会遇到相同的推导，就不在详细讲解了。

2.4.5 行列式的梯度

在已知一个矩阵$A \in \mathbb{R}^{n\times n}$，求这个矩阵行列式的梯度$\nabla_A|A|$，这不是一个容易的事情。首先我们回顾一下行列式的表达式：

$$|A| = \sum_{i=1}^{n}(-1)^{i+j}a_{ij}|A_{\setminus i,\setminus j}| \ (for\ any\ j \in 1,2...n) \tag{123}$$

所以，我们自然地可以得到：

$$\frac{\partial}{\partial A_{kl}}|A| = \frac{\partial}{\partial A_{kl}}\sum_{i=1}^{n}(-1)^{i+j}a_{ij}|A_{\setminus i,\setminus j}| = (-1)^{k+l}|A_{\setminus k,\setminus l}| = (adj(A))_{lk} \tag{124}$$

通过分量表达式，就可以立刻写出整体的表达式：

$$\nabla_A|A| = (adj(A))^T = |A|A^{-T} \tag{125}$$

现在我们考虑一种特殊情况，$f : \mathbb{S}^n_{++} \to \mathbb{R}$，$f(A) = \log|A|$。我们注意到，矩阵$A$是一个正定的对称阵，所以行列式的值恒为正数，所以在对数函数下，是有意义的。通过链式求导法则，我们有：

$$\frac{\partial \log|A|}{\partial A_{ij}} = \frac{\partial \log|A|}{\partial|A|}\frac{\partial|A|}{\partial A_{ij}} = \frac{1}{|A|}\frac{\partial|A|}{\partial A_{ij}} \tag{126}$$

我们把刚才的梯度带入上式，得到：

$$\nabla_A \log|A| = \frac{1}{|A|}\nabla_A|A| = A^{-1} \tag{127}$$

最后一个等号，我们运用了矩阵A是一个对称阵的性质。我们还可以发现，这个形式和我们的对数函数求导，具有相同的形式，即：$\partial/(\partial x)\log x = 1/x$。

2.4.6 拉格朗日乘子法

最后我们简单介绍一下拉格朗日（Lagrangian）乘子法求解带有约束条件的极值问题，即：

$$\max_{x\in\mathbb{R}^n} x^TAx\ subject\ to\ ||x||_2^2 = 1 \tag{128}$$

其中$A \in \mathbb{S}^n$。拉格朗日乘子法把带有约束条件的问题，转化为无约束的问题引入一系列参数，被称为拉格朗日乘子（Lagrangian multiplier）。我们定义如下的函数：

$$\mathcal{L}(x,\lambda) = x^TAx - \lambda x^Tx \tag{129}$$

此处的λ就是所谓的拉格朗日乘子。对函数$\mathcal{L}$求x的梯度，并令其为0。我们得到：

$$\nabla_x\mathcal{L}(x,\lambda) = \nabla_x(x^TAx - \lambda(x^Tx - 1)) = 2A^Tx - 2\lambda x = 0 \tag{130}$$

继续化简，我们得到线性方程：$Ax = \lambda x$。这就说明了带有约束条件的优化问题的解，就是矩阵A对应的特征向量。

3 概率论

概率理论是机器学习不可或缺的组成部分。我们将依赖概率的知识推导机器学习的算法,这一部分依旧取自CS229课程的补充材料[3]。概率理论是非常复杂的，所以我们仅仅回顾概率论里面最基本的一些概念与公式。没有介绍到的，会在课程当中进行补充。

3.1 概率的基本概念

首先介绍概率的基本概念。

- 样本空间（sample space）Ω：一个随机试验的所有结果构成的集合。集合里面每一个元素$\omega \in \Omega$是试验结果在真实世界状态的完整描述。
- 事件集合（set of events or event space）$\mathcal{F}$：事件集合是样本空间的子集，事件集合里面包含各种事件。$\mathcal{F}$具有三个性质：（1）$\emptyset \in \mathcal{F}$（2）$A \in \mathcal{F} \Rightarrow \Omega \backslash A \in \mathcal{F}$ （3）$A_1, A_2, ... \in \mathcal{F} \Rightarrow \cup_i A_i \in \mathcal{F}$。
- 概率测量（probability measure）：一个函数$P : \mathcal{F} \to \mathbb{R}$满足如下的性质：（1）$P(A) \geq 0, \forall A \in \mathcal{F}$（2）$P(\Omega) = 1$（3）如果$A_1, A_2, ...$是互斥事件（$A_i \cap A_j = \emptyset$当$i \neq j$），满足$P(\cup_i A_i) = \sum_i P(A_i)$。

以上三条，构成了概率论的三条公理。我们举一个简单的例子：一个质地均匀的骰子。样本空间为$\Omega = \{1, 2, 3, 4, 5, 6\}$。比如最简单的事件集合为$\mathcal{F} = \{\emptyset, \Omega\}$，其余的事件集合都是样本空间的子集。接下来在讨论一下概率映射，对于上面的事件集合，具有$P(\emptyset) = 0$，$P(\Omega) = 1$。对于一个质地均匀的骰子，每个面的概率应该是均等的，举个例子，$P(\{1, 2, 3, 4\}) = 4/6$，或者$P(\{1, 2, 6\}) = 3/6$。
下面给出概率的一些常用的性质：

- 当$A \subseteq B \Rightarrow P(A) \leq P(B)$;
- $P(A \cap B) \leq \min(P(A), P(B))$;
- $P(A \cup B) \leq P(A) + P(B)$;
- $P(\Omega \backslash A) = 1 - P(A)$;
- 全概率公式：如果$A_1, A_2, ...A_k$这些互斥事件，满足$\cup_{i=1}^k A_i = \Omega$，那么$\sum_{i=1}^k P(A_k) = 1$。

3.1.1 条件概率

我们令事件B具有非零的概率。那么在事件B发生的前提下，事件A发生的条件概率，具有如下的定义：

$$P(A|B) = \frac{P(A \cap B)}{B} \tag{131}$$

等式左边$P(A|B)$表示在事件B发生的条件下，事件A发生的概率。等式右边，分子表示的是事件A和B同时发生的概率，两个事件独立的条件是当且仅当（if and only if）$P(A \cap B) = P(AB)$或者等价条件$P(A|B) = P(A)$。分母表示事件B单独发生的概率。

3.2 随机变量

我们考虑一个试验，扔10次硬币，我们想知道有多少次是正面朝上的。对于这个试验，样本空间里的一个元素对应了一个长度为10 的轨迹，我们假设这个元素为$\omega_0 = \{H, H, T, T, H, T, T, H, H, T\} \in \Omega$。然而，在实际生活中，我们并不关心这条轨迹的顺序，取而代之的是，我们更关心最后的结果，那么我们说这个结果就是所谓的**随机变量**（random variables）。
更一般的来说，随机变量是一个映射：$X : \Omega \to \mathbb{R}$。我们一般用大写字母$X(\omega)$来表示随机变量，有时候会简写为$X$来表示，但是我们应该指导，它依赖于样本空间的输出$\omega$。随机变量的取值一般采用小写字母来表示$x$。接下来我们举两个例子。
第一个例子还是抛硬币。当抛十次硬币，正面朝上的次数是从0到10，不可能取值5.5。我们把这样的随机变量叫做**离散型（discrete）随机变量**，所以概率的取值，取决于随机变量的取值，即：

$$P(X = k) := P(\{\omega : X(w) = k\}), \tag{132}$$

第二个例子是随机变量为粒子的放射性衰变时间。这个时间是可以连续取值的，所以他有无穷多个取值，我们把这种随机变量称为**连续型（continuous）随机变量**。此时的概率，我们给出如下的定义：

$$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\}) \tag{133}$$

3.2.1 累计分布函数

所谓累积分布函数（cumulative distribution function），是一个函数映射，即：$F_X : \mathbb{R} \to [0,1]$，它表示对概率的测量。我们给出如下的定义：

$$F_X \equiv P(X \leq x) \tag{134}$$

通过这个定义，我不仅可以看出CDF函数是单调递增的，而且我们可以计算事件集合$\mathcal{F}$的概率。我们给出CDF函数的一些性质：

- $0 \leq F_X(x) \leq 1.$
- $\lim_{x \to -\infty} = 0.$
- $\lim_{x \to \infty} = 1.$
- $x \leq y \Rightarrow F_X(x) \leq F_X(y).$

3.2.2 概率质量函数

当一个随机变量X只能取有限个值，也就是说，随机变量X是一个离散随机变量。有一种更简单的方法来表示对随机变量X的测量，我们把这种映射称为概率质量函数（probability mass function）$\Omega \to \mathbb{R}$，具有如下的形式：

$$p_X(x) \equiv P(X = x) \tag{135}$$

对于一个离散随机变量，我们常用$Val(X)$来表示这个随机变量的所有可能取值。我们拿投十次硬币，正面朝上的概率举例，$Val(X) = \{0,1,2,3,4,5,6,7,8,9,10\}$。同样PMF具有如下三条性质：

- $0 \leq p_X(x) \leq 1.$
- $\sum_{x \in Val(X)} P_X(x) = 1.$
- $\sum_{x \in A} P_X(x) = P(X \in A).$

3.2.3 概率密度函数

对于有一些连续型随机变量，CDF函数是处处可导的。在这种情况下，把CDF的导数，定义为**概率密度函数**（Probability Density Function or PDF），具有如下的定义式：

$$f_X(x) \equiv \frac{dF_X(x)}{dx} \tag{136}$$

我们应该要注意，当有一些连续型随机变量的CDF不是处处可导时，PDF不存在。
基于我们对PDF的定义，我们可以得到在一段非常小的单元Δx 内，对应的概率值为：

$$P(x \leq X \leq x + \Delta x) \approx f_X(x)\Delta x \tag{137}$$

CDFs和PDFs都可以去计算不同事件的概率。但是我们应该注意，PDF在给定x时，得到的结果不是概率，也就是说$f_X(x) \neq P(X = x)$。举一个例子来说，$f(x)$可以取值比1大，但是$f(x)$对所有实空间的积分至多为1。同样，我们给出PDF 的三个性质：

- $f_X(x) \geq 0$.
- $\int_{-\infty}^{+\infty} f(x) = 1$.
- $\int_{x \in A} f_X(x)dx = P(X \in A)$.

3.2.4 期望

我们分为两种情况，第一种是离散型随机变量，我们有PMF，即$p_X(x)$。当存在一个映射$g : \mathbb{R} \to \mathbb{R}$，我们可以认为是随机变量$x$通过映射，得到了一个函数值$g(x)$。那么我们定义函数$g(x)$**期望**（expectation）或者**期望值**（expected value）为：

$$E[g(X)] \equiv \sum_{x \in Val(X)} g(x)p_X(x) \tag{138}$$

第二种情况为连续型随机变量，对应的是PDF$f_X(x)$。那么函数$g(X)$的期望定义为：

$$E[g(X)] \equiv \int_{-\infty}^{+\infty} g(x)f_X(x)dx \tag{139}$$

我们可以举一个例子，当$g(x) = x$时，那么$g(x)$的期望$E[g(x)]$就是随机变量X的均值。关于期望，有如下四条常用的性质：

- 对于任意的常数a，$\forall a \in \mathbb{R}, E[a] = a$.
- 对于任意的常数a，有$E[af(X)] = aE[f(X)]$.
- 期望的线性叠加性：$E[f(X) + g(X)] = E[f(X)] + E[g(X)]$.
- 对于离散型随机变量X，$E[1\{X = k\}] = P(X = k)$.

3.2.5 方差

随机变量X的方差，是对随机变量X在均值$E[X]$ 变动的一种测量。正式的定义，我们给出如下的表达式：

$$Var[X] \equiv E[(X - E(X))^2] \tag{140}$$

我们可以把这个表达式展开，得到一个方差更精简的表达式：

$$\begin{aligned} E[(X - E(X))^2] &= E[X^2 - 2E[X]X + E[X]^2] && (141)\\ &= E[X^2] - 2E[X]E[X] + E[X]^2 && (142)\\ &= E[X^2] - E[X]^2 && (143) \end{aligned}$$

第二个等号，我们利用了期望的线性叠加性质，我们可以看到，一个随机变量的方差为随机变量的平方均值减去随机变量均值的平方。方差具有两个常用的性质：

- 对于任意常数$a \in \mathbb{R}$，$Var[a] = 0$.
- 对于任意常数$a \in \mathbb{R}$，$Var[af(X)] = a^2Var[f(X)]$.

我们可以给出两个简单的例子：

第一个例子，当PDF为均匀分布时，$f_X(x) = 1, \ \forall \ x \in [0, 1]$，其他情况下恒为0。我们计算均匀分布的期望与方差：

- $E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx = \int_0^1 x dx = \frac{1}{2}$.
- $E[X^2] = \int_{-\infty}^{+\infty} x^2 f_X(x) dx = \int_0^1 x^2 dx = \frac{1}{3}$.
- $Var[X] = E[X^2] - E[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$.

第二个例子，我们计算一个$g(x) = 1\{x \in A\}$函数的期望：

- 当随机变量是离散型时，$E[g(X)] = \sum_{x \in Val(X)} 1\{x \in A\} P_X(x) dx = \sum_{x \in A} P_X(x) dx = P(x \in A)$.
- 当随机变量是连续型时，$E[g(X)] = \int_{-\infty}^{+\infty} 1\{x \in A\} f_X(x) dx = \int_{x \in A} f_X(x) dx = P(x \in A)$.

3.2.6 常见的概率分布

离散型随机变量：

- $X \sim Bernoulli(p)$，这里$0 \leq p \leq 1$，PDF的形式如下：

$$p(x) = \begin{cases} p, & if\ x = 1 \\ 1 - p, & if\ x = 0 \end{cases} \tag{144}$$

- $X \sim Binomial(p)$，这里$0 \leq p \leq 1$，PDF的形式如下：

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{145}$$

- $X \sim Geometric(p)$，这里$0 \leq p \leq 1$，PDF的形式如下：

$$p(x) = p(1-p)^{x-1} \tag{146}$$

- $X \sim Poission(\lambda)$，这里$0 \leq \lambda$，PDF的形式如下：

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!} \tag{147}$$

连续型随机变量：

- $X \sim Uniform(a, b)$，这里$a < b$，PDF的形式如下：

$$f(x) = \begin{cases} \frac{1}{b-a}, & if\ a \leq x \leq b \\ 0, & otherwise \end{cases} \tag{148}$$

- $X \sim Exponential(\lambda)$，这里$\lambda > 0$，PDF的形式如下：

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & if\ x \geq 0 \\ 0, & otherwise \end{cases} \tag{149}$$

- $X \sim Normal(\mu, \sigma^2)$，这是我们将要在课上着重讨论的概率分布，又叫高斯分布（Gaussian distribution）。PDF的形式如下：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{150}$$

我们把以上的内容，总结为一个表，并把均值和方差一并给出，作为一个参考：

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & if\ x=1 \\ 1-p, & if\ x=0 \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$ for $0 \le k \le n$	np	npq
$Geometric(p)$	$p(1-p)^{x-1}$ for $k=1,2,...$	$1/p$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda x}$ for $k=1,2,...$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}$ for $\forall\ x \in (a,b)$	$(a+b)/2$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ for $x \ge 0, \lambda \ge 0$	$1/\lambda$	$1/\lambda^2$

3.3 二元随机变量

以上我们都是在讨论一个随机变量，但是更多的情况下，是多元随机变量，所以我们接下来，先讨论一下二元随机变量。

3.3.1 联合与边缘累积分布

假定我们有两个随机变量X和Y。首先我们引入X 和Y的**联合累积概率分布**，有如下的定义：

$$F_{XY}(x,y) = P(X \le x, Y \le y) \tag{151}$$

可以证明，通过联合累积概率分布函数，可以分别求得X和Y 各自的累积概率分布函数。那么联合CDF和各自的CDF的关系如下：

$$F_X(x) = \lim_{y\to+\infty} F_{XY}(x,y) \tag{152}$$

$$F_Y(y) = \lim_{x\to+\infty} F_{XY}(x,y) \tag{153}$$

这里，我们把$F_X(x)$和$F_Y(y)$称为边缘累积概率分布函数。联合累积概率分布函数具有如下的性质：

- $0 \le F_{XY}(x,y) \le 1.$
- $\lim_{x,y\to+\infty} F_{XY}(x,y) = 1.$
- $\lim_{x,y\to-\infty} F_{XY}(x,y) = 0$
- $F_X(x) = \lim_{y\to+\infty} F_{XY}(x,y).$

3.3.2 联合与边缘密度质量分布

如果X和Y是离散型随机变量，那么就可以定义联合概率质量函数$P_{XY} : \mathbb{R} \times \mathbb{R} \to [0,1]$，具有如下的定义：

$$P_{XY}(x,y) = P(X=x, Y=y) \tag{154}$$

对于所有的x,y，我们都有$0 \le P_{X,Y}(x,y) \le 1$，并且具有概率归一性：$\sum_{x\in Val(X)}\sum_{y\in Val(Y)} P_{XY}(x,y) = 1$。

对于如何从联合概率质量函数，求得边缘概率质量函数，就是通过求和把其中一个随机变量给消掉，即：

$$P_X(x) = \sum_y P_{X,Y}(x,y). \tag{155}$$

我们把等号左边的$P_X(x)$称作边缘概率质量函数，在统计学上，一般把多元随机变量构成的联合分布通过求和或者积分给消掉，只剩余一个随机变量的函数的过程，称为“边缘化”（marginalization）。

3.3.3 联合与边缘概率密度分布

上面我们讨论了离散型随机变量，现在我们看一下连续型随机变量。假设X和Y是连续型随机变量，那么它们具有联合分布函数$F_{XY}(x,y)$，它们对自变量的二阶偏导数定义为联合概率密度函数（joint probability density function）：

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} \tag{156}$$

和单变量的pdf一样，它的取值不是概率，即$f_{XY}(x,y) \neq P(X=x, Y=y)$。而是：

$$\int\int_{x\in A} f_{XY}(x,y)dxdy = P((X,Y)\in A). \tag{157}$$

同样，我们说$f_{XY}(x,y)$总是非负的，并且可能存在在某一处的取值大于1，但是依旧要满足概率归一条件$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{XY}(x,y) = 1$。
与离散型的类似，我们可以通过积分消掉其中的一个随机变量，定义：

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy \tag{158}$$

为随机变量X的边缘概率密度分布（marginal probability density function），相同的方法，可以求出$f_Y y$。

3.3.4 条件分布

二元随机变量，我们同样可以定义条件概率。首先来看离散型随机变量，在$X=x$的情况下，Y发生的概率是多少，通过两个随机变量的PMF进行定义：

$$f_{Y|X}(y|x) = \frac{P_{XY}(x,y)}{P_X(x)},\ P_X(x) \neq 0. \tag{159}$$

对于连续型随机变量，会存在一个问题，就是边缘概率密度分布，会存在在某处的值取0。一个更准确的定义，是通过联合累积概率函数的方法得到，再此我们不去考虑这些证明的细节，通过对离散型随机变量的类比，简单地给出连续型随机变量条件概率密度分布的形式：

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)},\ f_X(x) \neq 0. \tag{160}$$

3.3.5 贝叶斯规则

在定义了二元随机变量的条件概率分布后，我们就可以给出概率论里面一个很重要的公式了，贝叶斯规则$Bayes'srule$。离散型的具有如下的形式：

$$P_{Y|X}(y|x) = \frac{P_{XY}(x,y)}{P_X(x)} = \frac{P_{X|Y}(x|y)P_Y(y)}{\sum_{y'\in Val(Y)} P_{X|Y}(x|y')P_Y(y')} \tag{161}$$

连续型随机变量，具有如下的形式：

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{+\infty} f_{X|Y}(x|y')f_Y(y')dy'} \tag{162}$$

3.3.6 独立性

两个随机变量独立，意味着$\forall x, y \in \mathbb{R},\ F_{XY}(x,y) = F_X(x)F_Y(y)$。下面给出四个等价条件：

- 离散型随机变量，$\forall x \in Val(X), y \in Val(Y),\ P_{XY}(x,y) = P_X(x)P_Y(y).$
- 离散型随机变量，$\forall y \in Val(Y)$，且$P_X(x) \neq 0$，$P_{Y|X}(y|x) = P_Y(y).$
- 连续型随机变量，$\forall x, y \in \mathbb{R},\ f_{XY}(x,y) = f_X(x)f_Y(y).$
- 连续型随机变量，$\forall y \in \mathbb{R}$，且$f_X(x) \neq 0$，$f_{Y|X}(y|x) = f_Y(y).$

还有一种不是太正式的说法，两个随机变量各自互不影响，也可以认为是独立的。即只需要知道$f_X(x)$和$f_Y(y)$就可以知道它们的联合概率密度函数了。下面我们给出一个引理：**引理1**：如果对于任意的子集$A, B \subseteq \mathbb{R}$，$X$和$Y$是独立的，我们有：

$$P(x \in A, Y \in B) = P(X \in A)P(Y \in B). \tag{163}$$

通过上面的引理，可以证明X和Y之间的独立性。

3.3.7 期望与方差

二元随机变量的期望与方差。我们还是先讨论离散型的随机变量，假设存在一个函数，具有映射$g : \mathbf{R}^2 \to \mathbf{R}$，那么对于函数$g$的期望，有如下的定义：

$$E[g(X,Y)] \equiv \sum_{x \in Val(X)} \sum_{y \in Val(Y)} g(x,y)P_{XY}(x,y). \tag{164}$$

连续型随机变量，把求和号改成积分号，具有如下的形式：

$$E[g(X,Y)] \equiv \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y)f_{XY}(x,y)dxdy \tag{165}$$

接下来定义两个随机变量X和Y的协方差：

$$Cov[X,Y] \equiv E[(X - E[X])(Y - E[Y])] \tag{166}$$

我们可以仿照单元随机变量的形式，进行展开，得到如下的表达式：

$$\begin{aligned} Cov[X,Y] &= E[(X - E[X])(Y - E[Y])] && (167)\\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] && (168)\\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] && (169)\\ &= E[XY] - E[X]E[Y] && (170) \end{aligned}$$

当两个随机变量之间的协方差是0时，我们说X和Y是**不相关的**[1]（uncorrelated）。
方差与期望具有如下几个常用的性质：

- $E[f(X,Y) + g(X,Y)] = E[f(X,Y)] + E[g(X,Y)].$
- $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X,Y].$
- 如果X与Y之间独立，那么$Cov[X,Y] = 0.$
- 如果X与Y之间独立，那么$E[f(X)g(Y)] = E[f(X)]E[g(Y)].$

[1]然而，不相关和独立不是相同的意思。比如，如果$X \sim Uniform(-1,1)$，且$Y = X^2$，我们可以证明X和Y是不相关的，但是很明显X和Y是不独立的。

4 多元随机变量

多元随机变量，是在二元随机变量的基础上，推广而来。具体来说，我们有n个连续型随机变量，$X_1(\omega), X_2(\omega), ...X_n(\omega)$。这一部分，我们就仅关注连续型随机变量，对于离散型仅把符号进行一些调整即可得到对应的公式。

4.1 基本概念

我们首先给出多元随机变量一起组成的联合分布函数（joint distribution function），联合概率密度函数（joint probability density function），边缘概率密度函数（marginal probability density function），条件概率密度函数（conditional probability density function）：

$$
\begin{aligned}
F_{X_1,X_2,...,X_n}(x_1,x_2,...,x_n) &= P(X_1 \le x_1, X_2 \le x_2, ..., X_n \le x_n) && (171)\\
f_{X_1,X_2,...,X_n}(x_1,x_2,...,x_n) &= \frac{\partial^2 F_{X_1,X_2,...,X_n}(x_1,x_2,...,x_n)}{\partial x_1 ... \partial x_n} && (172)\\
f_{X_1}(x_1) &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{X_1,X_2,...,X_n}(x_1,x_2,...x_n) dx_1 dx_2 ... dx_n && (173)\\
f_{X_1|X_2,...,X_n}(x_1|x_2,...x_n) &= \frac{f_{X_1,X_2,...,X_n}(x_1,x_2,...,x_n)}{f_{X_2,...X_n}(x_2,...x_n)} && (174)
\end{aligned}
$$

如果计算一个事件$A \subseteq \mathbb{R}^n$的概率，我们有：

$$
P((x_1,x_2,...,x_n) \in A) = \int_{(x_1,x_2,...,x_n)\in A} f_{X_1,X_2,...,X_n}(x_1,x_2,...,x_n) dx_1 dx_2 ... dx_n \quad (175)
$$

根据条件概率的定义，我们可以得到多元概率密度函数分布的**链式法则**（Chain rule）：

$$
\begin{aligned}
f(x_1,x_2,...,x_n) &= f(x_n|x_1,x_2,...,x_{n-1}) f(x_1,x_2,...,x_{n-1}) && (176)\\
&= f(x_n|x_1,x_2,...,x_{n-1}) f(x_{n-1}|x_1,x_2,...,x_{n-2}) f(x_1,x_2,...,x_{n-1}) && (177)\\
&= f(x_1) \prod_{i=2}^{n} f(x_i|x_1,...,x_{i-1}). && (178)
\end{aligned}
$$

对于多个事件，先假设有事件$A_1, A_2, ..., A_k$，对于任意的子集$\mathbb{S} \subseteq \{1,2,...,k\}$，当他们彼此独立时，我们有：

$$
P(\cap_{i\in\mathbb{S}} A_i) = \prod_{i\in\mathbb{S}} P(A_i). \quad (179)
$$

与此相似，我们说多个随机变量$X_1, X_2, ..., X_n$是彼此独立的，有：

$$
f(x_1,...,x_n) = f(x_1) f(x_2) \cdots f(x_n). \quad (180)
$$

我们可以看到，对于多元随机变量的独立性，是从二元随机变量的独立性那里，推广而来。
随机变量的独立性，是在机器学习里面非常重要的一个环节。因为我们假定一个样本集里面的每个样本都来自于同一个不知道的分布产生。假设我们第一次采样得到的样本为$(x^{(1)}, y^{(1)})$，其余$m-1$个样本是第一个样本的副本，那么这m个样本构成的训练集，我们知道：

$$
P((x^{(1)},y^{(1)}),...,(x^{(m)},y^{(m)})) \neq \prod_{i=1}^{m} P(x^{(i)},y^{(i)}) \quad (181)
$$

尽管训练集的大小是m，但是由于它们不是彼此独立的，所以实力上对训练机器学习算法来说，有效的训练集大小是小于m的，所谓的训练集的有效规模（effective size）。

4.2 随机向量

假设我们有n个随机变量。我们把这n个随机变量叠在一起，构成一个列向量，即$X = [X_1, X_2, ...X_n]^T$。我们把这个列向量称为**随机向量**（random vector）。从这一角度出发，我们之前说过的多元随机变量得到的结果，对于随机向量也同样适用。

4.2.1 期望

考虑任意一个函数$g : \mathbb{R}^n \to \mathbb{R}$. 该函数的期望值的定义为：

$$E[g(X)] = \int_{\mathbb{R}^n} g(x_1, x_2, ..., x_n) f_{X_1, X_2, ..., X_n}(x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_n \tag{182}$$

积分限从负无穷到正无穷。如果函数$g : \mathbb{R}^n \to \mathbb{R}^m$，具体来说，$g(x)$的形式为：

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix} \tag{183}$$

那么期望值具有如下的形式：

$$E[g(x)] = \begin{bmatrix} E[g_1(x)] \\ E[g_2(x)] \\ \vdots \\ E[g_m(x)] \end{bmatrix} \tag{184}$$

4.2.2 协方差

对于给定的一个随机向量$X : \Omega \to \mathbb{R}^n$，它的协方差矩阵$\Sigma$是一个$n \times n$的对称方阵，里面的元素由$\Sigma_{ij} = Cov[X_i, X_j]$。从定义出发，我们可以写出方阵的具体表达式：

$$\Sigma = \begin{bmatrix} Cov[X_1, X_1] & \cdots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \cdots & Cov[X_n, X_n] \end{bmatrix} \tag{185}$$

$$= \begin{bmatrix} E[X_1^2] - E[X_1]E[X_1] & \cdots & E[X_1 X_n] - E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n X_1] - E[X_n]E[X_1] & \cdots & E[X_n^2] - E[X_n]E[X_n] \end{bmatrix} \tag{186}$$

$$= \begin{bmatrix} E[X_1^2] & \cdots & E[X_1 X_n] \\ \vdots & \ddots & \vdots \\ E[X_n X_1] & \cdots & E[X_n^2] \end{bmatrix} - \begin{bmatrix} E[X_1]E[X_1] & \cdots & E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n]E[X_1] & \cdots & E[X_n]E[X_n] \end{bmatrix} \tag{187}$$

$$= E[XX^T] - E[X]E[X]^T = ... = E[(X - E[X])(X - E[X])^T]. \tag{188}$$

最后我们得到了期望的定义式。同样协方差矩阵具有以下几个常见的性质：

- $\Sigma \succeq 0$，意味着协方差矩阵是一个半正定矩阵。
- $\Sigma = \Sigma^T$是对称的。

4.3 多元高斯分布

一个特别重要的概率分布是多元高斯分布（multivariate Gaussian distribution）。一个随机向量$X \in \mathbb{R}^n$具有均值为$\mu \in \mathbb{R}^n$，协方差为$\Sigma \in \mathcal{S}_{++}^n$。具体的表达式如下：

$$f_{X_1.X_2,...,X_n}(x_1, x_2, ..., x_n; \mu, \sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \tag{189}$$

我们写成$X \sim N(\mu, \Sigma)$。注意到，当$n = 1$时，这个公式退化为简单的一元高斯分布，此时对应的均值为μ_1，方差为Σ_{11}。

References

[1] Tongji University Department of mathematics. *Advanced mathematics,Seventh Edition.* Higher Education Press, 2014.

[2] Zico Kolter. *Linear Algebra Review and Reference.* Standford University, 2020.

[3] Arian Maleki and Tom Do. *Review of Probablity Theory.* Standford University, 2020.