数学基础笔记（V2）

你不是一个人在战斗！

haiguang2000@qq.com

最后修改：2020-04-24

目录

CS229 机器学习课程复习材料-线性代数

这部分是斯坦福大学 CS 229 机器学习课程的基础材料，原始文件下载

原文作者：Zico Kolter，修改：Chuong Do， Tengyu Ma

翻译：黄海广

1. 基础概念和符号

线性代数提供了一种紧凑地表示和操作线性方程组的方法。 例如，以下方程组：

$$4x_1 - 5x_2 = -13$$
$$-2x_1 + 3x_2 = 9$$

这是两个方程和两个变量，正如你从高中代数中所知，你可以找到 x_1 和 x_2 的唯一解（除非方程以某种方式退化，例如，如果第二个方程只是第一个的倍数，但在上面的情况下，实际上只有一个唯一解）。 在矩阵表示法中，我们可以更紧凑地表达：

$$Ax = b$$
$$\text{with } A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$$

我们可以看到，这种形式的线性方程有许多优点（比如明显地节省空间）。

1.1 基本符号

我们使用以下符号：

- $A \in \mathbb{R}^{m \times n}$，表示 A 为由实数组成具有m行和n列的矩阵。
- $x \in \mathbb{R}^n$，表示具有n个元素的向量。 通常，向量x将表示列向量：即，具有n行和1列的矩阵。 如果我们想要明确地表示行向量：具有 1 行和n列的矩阵 - 我们通常写x^T（这里$x^T x$的转置）。
- x_i表示向量x的第i个元素

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

- 我们使用符号 a_{ij}（或$A_{ij}, A_{i,j}$等）来表示第 i 行和第j列中的 A 的元素：

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

- 我们用a^j或者$A_{:,j}$表示矩阵A的第j列：

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

- 我们用a_i^T或者$A_{i,:}$表示矩阵A的第i行：

$$A = \begin{bmatrix} -a_1^T - \\ -a_2^T - \\ \vdots \\ -a_m^T - \end{bmatrix}$$

- 在许多情况下，将矩阵视为列向量或行向量的集合非常重要且方便。 通常，在向量而不是标量上操作在数学上（和概念上）更清晰。只要明确定义了符号，用于矩阵的列或行的表示方式并没有通用约定。

2. 矩阵乘法

两个矩阵相乘，其中 $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ ，则：

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

其中：

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

请注意，为了使矩阵乘积存在，A中的列数必须等于B中的行数。有很多方法可以查看矩阵乘法，我们将从检查一些特殊情况开始。

2.1 向量-向量乘法

给定两个向量$x, y \in \mathbb{R}^n$,$x^T y$通常称为**向量内积**或者**点积**，结果是个**实数**。

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

注意：$x^T y = y^T x$ 始终成立。

给定向量 $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ (他们的维度是否相同都没关系)，$xy^T \in \mathbb{R}^{m \times n}$叫做**向量外**

积 ，当 $(xy^T)_{ij} = x_i y_j$ 的时候，它是一个矩阵。

$$xy^T \in \mathbb{R}^{m\times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix}$$

举一个外积如何使用的一个例子：让$\mathbf{1} \in R^n$表示一个n维向量，其元素都等于 1，此外，考虑矩阵$A \in R^{m\times n}$，其列全部等于某个向量 $x \in R^m$。 我们可以使用外积紧凑地表示矩阵 A:

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

2.2 矩阵-向量乘法

给定矩阵 $A \in \mathbb{R}^{m\times n}$，向量 $x \in \mathbb{R}^n$ ，它们的积是一个向量 $y = Ax \in R^m$。有几种方法可以查看矩阵向量乘法，我们将依次查看它们中的每一种。

如果我们按行写A，那么我们可以表示Ax为：

$$y = Ax = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}$$

换句话说，第i个y是A的第i行和x的内积，即：$y_i = y_i = a_i^T x$。

同样的， 可以把 A 写成列的方式，则公式如下：，

$$y = Ax = \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_1 \end{bmatrix} x_1 + \begin{bmatrix} a_2 \end{bmatrix} x_2 + \cdots + \begin{bmatrix} a_n \end{bmatrix} x_n$$

换句话说，y是A的列的线性组合，其中线性组合的系数由x的元素给出。

到目前为止，我们一直在右侧乘以列向量，但也可以在左侧乘以行向量。 这是写的，$y^T = x^T A$ 表示$A \in \mathbb{R}^{m\times n}$，$x \in \mathbb{R}^m$，$y \in \mathbb{R}^n$。 和以前一样，我们可以用两种可行的方式表达$y^T$，这取决于我们是否根据行或列表达$A$.

第一种情况，我们把A用列表示：

$$y^T = x^T A = x^T \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x^T a^1 & x^T a^2 & \ldots & x^T a^n \end{bmatrix}$$

这表明y^T的第i个元素等于x和A的第i列的内积。

最后，根据行表示A，我们得到了向量-矩阵乘积的最终表示:

$$y^T = x^T A = [x_1 \quad x_2 \quad \cdots \quad x_n] \begin{bmatrix} -a_1^T - \\ -a_2^T - \\ \vdots \\ -a_m^T - \end{bmatrix} = x_1[-a_1^T -] + x_2[-a_2^T -] + \cdots + x_n[-a_n^T -]$$

所以我们看到y^T是A的行的线性组合，其中线性组合的系数由x的元素给出。

2.3 矩阵-矩阵乘法

有了这些知识，我们现在可以看看四种不同的（形式不同，但结果是相同的）矩阵-矩阵乘法：也就是本节开头所定义的$C = AB$的乘法。

首先，我们可以将矩阵 - 矩阵乘法视为一组向量-向量乘积。 从定义中可以得出：最明显的观点是C的(i, j)元素等于A的第i行和B的的j列的内积。 如下面的公式所示：

$$C = AB = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & & \vdots \\ - & a_m^T & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \cdots & a_1^T b_p \\ a_2^T b_1 & a_2^T b_2 & \cdots & a_2^T b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b_1 & a_m^T b_2 & \cdots & a_m^T b_p \end{bmatrix}$$

这里的$A \in \mathbb{R}^{m \times n}$ ，$B \in \mathbb{R}^{n \times p}$， $a_i \in \mathbb{R}^n$ ，$b^j \in \mathbb{R}^{n \times p}$， 这里的$A \in \mathbb{R}^{m \times n}$， $B \in \mathbb{R}^{n \times p}$，$a_i \in \mathbb{R}^n$，$b^j \in \mathbb{R}^{n \times p}$，所以它们可以计算内积。我们用通常用行表示$A$而用列表示$B$。或者，我们可以用列表示$A$，用行表示$B$，这时$AB$是求外积的和。公式如下：

$$C = AB = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & b_1^T & - \\ - & b_2^T & - \\ & & \vdots \\ - & b_n^T & - \end{bmatrix} = \sum_{i=1}^{n} a_i \, b_i^T$$

换句话说，AB等于所有的A的第i列和B第i行的外积的和。因此，在这种情况下， $a_i \in \mathbb{R}^m$和$b_i \in \mathbb{R}^p$， 外积$a^i b_i^T$的维度是$m \times p$，与C的维度一致。

其次，我们还可以将矩阵 - 矩阵乘法视为一组矩阵向量积。如果我们把B用列表示，我们可以将C的列视为A和B的列的矩阵向量积。公式如下：

$$C = AB = A \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ Ab_1 & Ab_2 & \cdots & Ab_p \\ | & | & & | \end{bmatrix}$$

这里C的第i列由矩阵向量乘积给出，右边的向量为$c_i = Ab_i$。 这些矩阵向量乘积可以使用前一小节中给出的两个观点来解释。 最后，我们有类似的观点，我们用行表示A，C的行作为A和C行之间的矩阵向量积。公式如下：

$$C = AB = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} B = \begin{bmatrix} - & a_1^T B & - \\ - & a_2^T B & - \\ & & \vdots \\ - & a_m^T B & - \end{bmatrix}$$

这里第i行的C由左边的向量的矩阵向量乘积给出：$c_i^T = a_i^T B$

将矩阵乘法剖析到如此大的程度似乎有点过分，特别是当所有这些观点都紧跟在我们在本节开头给出的初始定义（在一行数学中）之后。

这些不同方法的直接优势在于它们允许您**在向量的级别/单位而不是标量上进行操作**。为了完全理解线性代数而不会迷失在复杂的索引操作中，关键是要用尽可能多的概念进行操作。

实际上所有的线性代数都处理某种矩阵乘法，花一些时间对这里提出的观点进行直观的理解是非常必要的。

除此之外，了解一些更高级别的矩阵乘法的基本属性是很有必要的：

- 矩阵乘法结合律：$(AB)C = A(BC)$
- 矩阵乘法分配律：$A(B + C) = AB + AC$
- 矩阵乘法通常不是可交换的；也就是说，通常$AB \neq BA$。（例如，假设$A \in \mathbb{R}^{m \times n}$，$B \in \mathbb{R}^{n \times p}$，如果$m$和$q$不相等，矩阵乘积$BA$甚至不存在！）

如果您不熟悉这些属性，请花点时间自己验证它们。 例如，为了检查矩阵乘法的相关性，假设$A \in \mathbb{R}^{m \times n}$， $B \in \mathbb{R}^{n \times p}$，$C \in \mathbb{R}^{p \times q}$。 注意$AB \in \mathbb{R}^{m \times p}$，所以$(AB)C \in \mathbb{R}^{m \times q}$。 类似地，$BC \in \mathbb{R}^{n \times q}$，所以$A(BC) \in \mathbb{R}^{m \times q}$。 因此，所得矩阵的维度一致。 为了表明矩阵乘法是相关的，足以检查$(AB)C$的第(i, j)个元素是否等于$A(BC)$的第(i, j)个元素。 我们可以使用矩阵乘法的定义直接验证这一点：

$$\begin{aligned} ((AB)C)_{ij} &= \sum_{k=1}^{p} (AB)_{ik} C_{kj} = \sum_{k=1}^{p} \left(\sum_{l=1}^{n} A_{il} B_{lk} \right) C_{kj} \\ &= \sum_{k=1}^{p} \left(\sum_{l=1}^{n} A_{il} B_{lk} C_{kj} \right) = \sum_{l=1}^{n} \left(\sum_{k=1}^{p} A_{il} B_{lk} C_{kj} \right) \\ &= \sum_{l=1}^{n} A_{il} \left(\sum_{k=1}^{p} B_{lk} C_{kj} \right) = \sum_{l=1}^{n} A_{il} (BC)_{lj} = (A(BC))_{ij} \end{aligned}$$

3 运算和属性

在本节中，我们介绍矩阵和向量的几种运算和属性。希望能够为您复习大量此类内容，这些笔记可以作为这些主题的参考。

3.1 单位矩阵和对角矩阵

单位矩阵，$I \in \mathbb{R}^{n\times n}$，它是一个方阵，对角线的元素是 1，其余元素都是 0：

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

对于所有$A \in \mathbb{R}^{m\times n}$，有：

$$AI = A = IA$$

注意，在某种意义上，单位矩阵的表示法是不明确的，因为它没有指定I的维数。通常，I的维数是从上下文推断出来的，以便使矩阵乘法成为可能。 例如，在上面的等式中，$AI = A$中的 I 是$n \times n$矩阵，而$A = IA$中的I是$m \times m$矩阵。

对角矩阵是一种这样的矩阵：对角线之外的元素全为 0。对角阵通常表示为：$D = diag(d_1, d_2, \ldots, d_n)$，其中：

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

很明显：单位矩阵$I = diag(1,1,\ldots,1)$。

3.2 转置

矩阵的转置是指翻转矩阵的行和列。

给定一个矩阵：

$A \in \mathbb{R}^{m\times n}$，它的转置为$n \times m$的矩阵$A^T \in \mathbb{R}^{n\times m}$，其中的元素为：

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

事实上，我们在描述行向量时已经使用了转置，因为列向量的转置自然是行向量。

转置的以下属性很容易验证：

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

3.3 对称矩阵

如果$A = A^T$，则矩阵$A \in \mathbb{R}^{n\times n}$是对称矩阵。 如果$A = -A^T$，它是反对称的。 很容易证明，对于任何矩阵$A \in \mathbb{R}^{n\times n}$，矩阵$A + A^T$是对称的，矩阵$A - A^T$是反对称的。 由此得出，任何方矩阵$A \in \mathbb{R}^{n\times n}$可以表示为对称矩阵和反对称矩阵的和，所以：

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

上面公式的右边的第一个矩阵是对称矩阵，而第二个矩阵是反对称矩阵。 事实证明，对称矩阵在实践中用到很多，它们有很多很好的属性，我们很快就会看到它们。 通常将大小为n的所有对称矩阵的集合表示为$\mathbb{S}^n$，因此$A \in \mathbb{S}^n$意味着A是对称的$n \times n$矩阵;

3.4 矩阵的迹

方矩阵$A \in \mathbb{R}^{n\times n}$的迹，表示为$\mathrm{tr}(A)$（或者只是$\mathrm{tr}A$，如果括号显然是隐含的），是矩阵中对角元素的总和：

$$\mathrm{tr}A = \sum_{i=1}^{n} A_{ii}$$

如 **CS229** 讲义中所述，迹具有以下属性（如下所示）：

- 对于矩阵$A \in \mathbb{R}^{n\times n}$，则：$\mathrm{tr}A = \mathrm{tr}A^T$
- 对于矩阵$A, B \in \mathbb{R}^{n\times n}$，则：$\mathrm{tr}(A + B) = \mathrm{tr}A + \mathrm{tr}B$
- 对于矩阵$A \in \mathbb{R}^{n\times n}$，$t \in \mathbb{R}$，则：$\mathrm{tr}(tA) = t\mathrm{tr}A.$
- 对于矩阵 A，B，AB 为方阵，则：$\mathrm{tr}AB = \mathrm{tr}BA$
- 对于矩阵 A，B，C，ABC为方阵，则：$\mathrm{tr}ABC = \mathrm{tr}BCA = \mathrm{tr}CAB$，同理，更多矩阵的积也是有这个性质。

作为如何证明这些属性的示例，我们将考虑上面给出的第四个属性。 假设$A \in \mathbb{R}^{m\times n}$和$B \in \mathbb{R}^{n\times m}$（因此$AB \in \mathbb{R}^{m\times m}$是方阵）。 观察到$BA \in \mathbb{R}^{n\times n}$也是一个方阵，因此对它们进行迹的运算是有意义的。 要证明$\mathrm{tr}AB = \mathrm{tr}BA$，请注意：

$$\begin{aligned}\mathrm{tr}AB &= \sum_{i=1}^{m}(AB)_{ii} = \sum_{i=1}^{m}\left(\sum_{j=1}^{n}A_{ij}\,B_{ji}\right)\\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}\,B_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{m}B_{ji}\,A_{ij}\\ &= \sum_{j=1}^{n}\left(\sum_{i=1}^{m}B_{ji}\,A_{ij}\right) = \sum_{j=1}^{n}(BA)_{jj} = \mathrm{tr}BA\end{aligned}$$

这里，第一个和最后两个等式使用迹运算符和矩阵乘法的定义，重点在第四个等式，使用标量乘法的可交换性来反转每个乘积中的项的顺序，以及标量加法的可交换性和相关性，以便重新排列求和的顺序。

3.5 范数

向量的范数$\|x\|$是非正式度量的向量的“长度”。例如，我们有常用的欧几里德或ℓ_2范数，

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

注意：$\|x\|_2^2 = x^Tx$

更正式地，范数是满足 4 个属性的函数（$f:\mathbb{R}^n \to \mathbb{R}$）：

对于所有的 $x \in \mathbb{R}^n$, $f(x) \geq 0$(非负).

当且仅当$x = 0$ 时，$f(x) = 0$ (明确性).

对于所有$x \in \mathbb{R}^n, t \in \mathbb{R}$，则 $f(tx) = |t|f(x)$ (正齐次性).

对于所有 $x, y \in \mathbb{R}^n$, $f(x+y) \leq f(x) + f(y)$ (三角不等式)

其他范数的例子是ℓ_1范数:

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

和ℓ_∞范数：

$$\|x\|_\infty = \max_i |x_i|$$

事实上，到目前为止所提出的所有三个范数都是ℓ_p范数族的例子，它们由实数$p \geq 1$参数化，并定义为：

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

也可以为矩阵定义范数，例如 **Frobenius** 范数:

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

许多其他更多的范数，但它们超出了这个复习材料的范围。

3.6 线性相关性和秩

一组向量$x_1, x_2, \cdots x_n \in \mathbb{R}$，如果没有向量可以表示为其余向量的线性组合，则称称该向量是线性无相关的。 相反，如果属于该组的一个向量可以表示为其余向量的线性组合，则称该向量是线性相关的。 也就是说，如果：

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

对于某些标量值$\alpha_1, \cdots \alpha_n - 1 \in \mathbb{R}$，要么向量$x_1, x_2, \cdots x_n$是线性相关的; 否则，向量是线性无关的。 例如，向量：

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

是线性相关的，因为：$x_3 = -2x_1 + x_2$。

矩阵$A \in \mathbb{R}^{m \times n}$的**列秩**是构成线性无关集合的$A$的最大列子集的大小。 由于术语的多样性，这通常简称为A的线性无关列的数量。同样，行秩是构成线性无关集合的A的最大行数。对于任何矩阵$A \in \mathbb{R}^{m \times n}$，事实证明$A$的列秩等于$A$的行秩（尽管我们不会证明这一点），因此两个量统称为A的**秩**，用 $\mathrm{rank}(A)$表示。 以下是秩的一些基本属性：

- 对于 $A \in \mathbb{R}^{m \times n}$，$\mathrm{rank}(A) \leq min(m, n)$，如果$ \text(A) = \text{min} (m, n)$，则： A 被称作**满秩**。
- 对于 $A \in \mathbb{R}^{m \times n}$， $\mathrm{rank}(A) = \mathrm{rank}(A^T)$
- 对于 $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}$, $\mathrm{rank}(AB) \leq \min(\mathrm{rank}(A), \mathrm{rank}(B))$
- 对于 $A, B \in \mathbb{R}^{m \times n}$，$\mathrm{rank}(A + B) \leq \mathrm{rank}(A) + \mathrm{rank}(B)$

3.7 方阵的逆

方阵$A \in \mathbb{R}^{n\times n}$的倒数表示为$A^{-1}$，并且是这样的独特矩阵:

$$A^{-1}A = I = AA^{-1}$$

请注意，并非所有矩阵都具有逆。 例如，非方形矩阵根据定义没有逆。 然而，对于一些方形矩阵A，可能仍然存在A^{-1}可能不存在的情况。 特别是，如果A^{-1}存在，我们说A是**可逆**的或**非奇异**的，否则就是**不可逆**或**奇异**的。为了使方阵 A 具有逆A^{-1}，则A必须是满秩。我们很快就会发现，除了满秩之外，还有许多其它的充分必要条件。 以下是逆的属性; 假设$A, B \in \mathbb{R}^{n\times n}$，而且是非奇异的：

- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1}$因此，该矩阵通常表示为A^{-T}。 作为如何使用逆的示例，考虑线性方程组，$Ax = b$，其中$A \in \mathbb{R}^{n\times n}$，$x, b \in \mathbb{R}$， 如果$A$是非奇异的（即可逆的），那么$x = A^{-1}b$。 （如果$A \in \mathbb{R}^{m\times n}$不是方阵，这公式还有用吗？）

3.8 正交阵

如果 $x^T y = 0$，则两个向量$x, y \in \mathbb{R}^n$ 是**正交**的。如果$\| x \|_2 = 1$，则向量$x \in \mathbb{R}^n$ 被归一化。如果一个方阵$U \in \mathbb{R}^{n\times n}$的所有列彼此正交并被归一化（这些列然后被称为正交），则方阵U是正交阵（注意在讨论向量时的意义不一样）。

它可以从正交性和正态性的定义中得出: $A-1A = I = AA-1$

$$U^T U = I = UU^T$$

换句话说，正交矩阵的逆是其转置。 注意，如果U不是方阵 :即，$U \in \mathbb{R}^{m\times n}$，$n < m$ ，但其列仍然是正交的，则$U^T U = I$，但是$UU^T \neq I$。我们通常只使用术语"正交"来描述先前的情况 ，其中U是方阵。 正交矩阵的另一个好的特性是在具有正交矩阵的向量上操作不会改变其欧几里德范数，即:

$$\| Ux \|_2 = \| x \|_2$$

对于任何 $x \in \mathbb{R}$ ，$U \in \mathbb{R}^n$是正交的。

3.9 矩阵的值域和零空间

一组向量$\{x_1, \dots x_n\}$是可以表示为$\{x_1, \dots x_n\}$的线性组合的所有向量的集合。 即：

$$\text{span}(\{x_1, \dots x_n\}) = \left\{v: v = \sum_{i=1}^{n} \alpha_i x_i, \quad \alpha_i \in \mathbb{R}\right\}$$

可以证明，如果$\{x_1, \dots x_n\}$是一组n个线性无关的向量，其中每个$x_i \in \mathbb{R}^n$，则$\text{span}(\{x_1, \dots x_n\}) = \mathbb{R}^n$。 换句话说，任何向量$v \in \mathbb{R}^n$都可以写成$x_1$到$x_n$的线性组合。

向量$y \in \mathbb{R}^m$投影到$\{x_1, \dots x_n\}$（这里我们假设$x_i \in \mathbb{R}^m$）得到向量$v \in \text{span}(\{x_1, \dots, x_n\})$，由欧几里德范数$\| v - y \|_2$可以得知，这样$v$尽可能接近$y$。

我们将投影表示为$\text{Proj}(y; \{x_1, \dots x_n\})$，并且可以将其正式定义为:

$$\text{Proj}(y; \{x_1, \dots x_n\}) = \text{argmin}_{v \in \text{span}(\{x_1, \dots, x_n\})} \| y - v \|_2$$

矩阵$A \in \mathbb{R}^{m \times n}$的值域（有时也称为列空间），表示为$\mathcal{R}(A)$，是$A$列的跨度。换句话说，

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

做一些技术性的假设（即A是满秩且$n < m$），向量$y \in \mathbb{R}^m$到A的范围的投影由下式给出:

$$\text{Proj}(y; A) = \text{argmin}_{v \in \mathcal{R}(A)} \| v - y \|_2 = A(A^T A)^{-1} A^T y$$

这个最后的方程应该看起来非常熟悉，因为它几乎与我们在课程中（我们将很快再次得出）得到的公式：用于参数的最小二乘估计一样。 看一下投影的定义，显而易见，这实际上是我们在最小二乘问题中最小化的目标（除了范数的平方这里有点不一样，这不会影响找到最优解），所以这些问题自然是非常相关的。

当A只包含一列时，$a \in \mathbb{R}^m$，这给出了向量投影到一条线上的特殊情况：

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

一个矩阵$A \in \mathbb{R}^{m \times n}$的零空间 $\mathcal{N}(A)$ 是所有乘以A时等于 0 向量的集合，即：

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

注意，$\mathcal{R}(A)$中的向量的大小为m，而 $\mathcal{N}(A)$ 中的向量的大小为n，因此$\mathcal{R}(A^T)$和 $\mathcal{N}(A)$中的向量的大小均为$\mathbb{R}^n$。 事实上，还有很多例子。 证明：

$$\{w: w = u + v, u \in \mathcal{R}(A^T), v \in \mathcal{N}(A)\} = \mathbb{R}^n \text{ and } \mathcal{R}(A^T) \cap \mathcal{N}(A) = \{\mathbf{0}\}$$

换句话说，$\mathcal{R}(A^T)$和 $\mathcal{N}(A)$ 是不相交的子集，它们一起跨越$\mathbb{R}^n$的整个空间。 这种类型的集合称为**正交补**，我们用$\mathcal{R}(A^T) = \mathcal{N}(A)^{\perp}$表示。

3.10 行列式

一个方阵$A \in \mathbb{R}^{n\times n}$的行列式是函数det：$\mathbb{R}^{n\times n} \to \mathbb{R}^n$，并且表示为$|A|$。 或者det$A$（有点像迹运算符，我们通常省略括号）。 从代数的角度来说，我们可以写出一个关于A行列式的显式公式。 因此，我们首先提供行列式的几何解释，然后探讨它的一些特定的代数性质。

给定一个矩阵：

$$\begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_n^T & - \end{bmatrix}$$

考虑通过采用A行向量$a_1, \dots a_n \in \mathbb{R}^n$的所有可能线性组合形成的点$S \subset \mathbb{R}^n$的集合，其中线性组合的系数都在 0 和 1 之间; 也就是说，集合S是span$(\{a_1, \dots a_n\})$受到系数$a_1, \dots a_n$的限制的线性组合，$\alpha_1, \cdots, \alpha_n$满足$0 \le \alpha_i \le 1, i = 1, \dots, n$。从形式上看，

$$S = \left\{ v \in \mathbb{R}^n : v = \sum_{i=1}^{n} \alpha_i a_i \text{ where } 0 \le \alpha_i \le 1, i = 1, \dots, n \right\}$$

事实证明，A的行列式的绝对值是对集合S的“体积”的度量。

比方说：一个2×2的矩阵(4)：

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}$$

它的矩阵的行是：

$$a_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad a_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

对应于这些行对应的集合S如图 1 所示。对于二维矩阵，S通常具有平行四边形的形状。在我们的例子中，行列式的值是$|A| = -7$（可以使用本节后面显示的公式计算），因此平行四边形的面积为 7。（请自己验证！）

在三维中，集合S对应于一个称为平行六面体的对象（一个有倾斜边的三维框，这样每个面都有一个平行四边形）。行定义S的3×3矩阵 S 的行列式的绝对值给出了平行六面体的三维体积。在更高的维度中，集合S是一个称为n维平行切的对象。

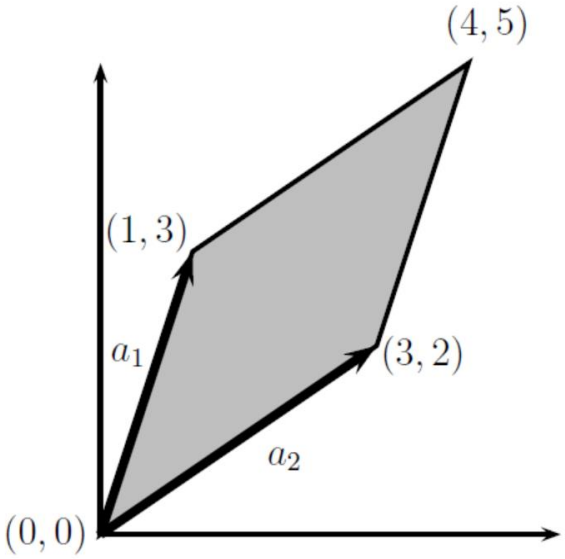


图 1：给出的2×2矩阵A的行列式的图示。 这里，a_1和a_2是对应于A行的向量，并且集合S对应于阴影区域（即，平行四边形）。 这个行列式的绝对值，$|\det A| = 7$，即平行四边形的面积。

在代数上，行列式满足以下三个属性（所有其他属性都遵循这些属性，包括通用公式）：

1. 恒等式的行列式为 1，$|I| = 1$（几何上，单位超立方体的体积为 1）。

2. 给定一个矩阵 $A \in \mathbb{R}^{n \times n}$，如果我们将$A$中的一行乘上一个标量$t \in \mathbb{R}$，那么新矩阵的行列式是$t|A|$

$$\begin{vmatrix} - & ta_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ & a_m^T & - \end{vmatrix} = t|A|$$

几何上，将集合S的一个边乘以系数t，体积也会增加一个系数t。

1. 如果我们交换任意两行在a_i^T和a_j^T，那么新矩阵的行列式是$-|A|$，例如：

$$\begin{vmatrix} - & a_2^T & - \\ - & a_1^T & - \\ & \vdots & \\ - & a_m^T & - \end{vmatrix} = -|A|$$

你一定很奇怪，满足上述三个属性的函数的存在并不多。事实上，这样的函数确实存在，而且是唯一的（我们在这里不再证明了）。

从上述三个属性中得出的几个属性包括：

- 对于 $A \in \mathbb{R}^{n \times n}$，$|A| = |A^T|$

- 对于 $A, B \in \mathbb{R}^{n \times n}$，$|AB| = |A||B|$

- 对于 $A \in \mathbb{R}^{n\times n}$，有且只有当$A$是奇异的（比如不可逆），则：$|A| = 0$
- 对于 $A \in \mathbb{R}^{n\times n}$ 同时，A为非奇异的，则：$|A^{-1}| = 1/|A|$

在给出行列式的一般定义之前，我们定义，对于$A \in \mathbb{R}^{n\times n}$，$A_{\setminus i,\setminus j} \in \mathbb{R}^{(n-1)\times(n-1)}$是由于删除第$i$行和第$j$列而产生的矩阵。 行列式的一般（递归）公式是：

$$\begin{aligned}|A| &= \sum_{i=1}^{n}(-1)^{i+j}a_{ij}\left|A_{\setminus i,\setminus j}\right| \quad (\text{ for any } j \in 1,\dots,n)\\ &= \sum_{j=1}^{n}(-1)^{i+j}a_{ij}\left|A_{\setminus i,\setminus j}\right| \quad (\text{ for any } i \in 1,\dots,n)\end{aligned}$$

对于 $A \in \mathbb{R}^{1\times 1}$，初始情况为$|A| = a_{11}$。如果我们把这个公式完全展开为 $A \in \mathbb{R}^{n\times n}$，就等于$n!$（$n$阶乘）不同的项。因此，对于大于$3\times 3$的矩阵，我们几乎没有明确地写出完整的行列式方程。然而，3×3大小的矩阵的行列式方程是相当常见的，建议好好地了解它们：

$$|[a_{11}]| = a_{11}$$

$$\left|\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}\right| = a_{11}a_{22} - a_{12}a_{21}$$

$$\left|\begin{bmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{bmatrix}\right| = \begin{array}{l}a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}\\ -a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}\end{array}$$

矩阵$A \in \mathbb{R}^{n\times n}$的经典伴随矩阵（通常称为伴随矩阵）表示为$\text{adj}(A)$，并定义为：

$$\text{adj}(A) \in \mathbb{R}^{n\times n}, \quad (\text{adj}(A))_{ij} = (-1)^{i+j}\left|A_{\setminus j,\setminus i}\right|$$

（注意索引$A_{\setminus j,\setminus i}$中的变化）。可以看出，对于任何非奇异$A \in \mathbb{R}^{n\times n}$，

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

虽然这是一个很好的“显式”的逆矩阵公式，但我们应该注意，从数字上讲，有很多更有效的方法来计算逆矩阵。

3.11 二次型和半正定矩阵

给定方矩阵$A \in \mathbb{R}^{n\times n}$和向量$x \in \mathbb{R}^{n}$，标量值$x^{T}Ax$被称为二次型。 写得清楚些，我们可以看到：

$$x^{T}Ax = \sum_{i=1}^{n}x_i\,(Ax)_i = \sum_{i=1}^{n}x_i\left(\sum_{j=1}^{n}A_{ij}\,x_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij}\,x_i x_j$$

注意：

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

第一个等号的是因为是标量的转置与自身相等，而第二个等号是因为是我们平均两个本身相等的量。 由此，我们可以得出结论，只有A的对称部分有助于形成二次型。 出于这个原因，我们经常隐含地假设以二次型出现的矩阵是对称阵。 我们给出以下定义：

- 对于所有非零向量$x \in \mathbb{R}^n$，$x^TAx > 0$，对称阵$A \in \mathbb{S}^n$为**正定**（positive definite,PD）。这通常表示为$A \succ 0$（或$A > 0$），并且通常将所有正定矩阵的集合表示为$\mathbb{S}^n_{++}$。

- 对于所有向量$x^TAx \geq 0$，对称矩阵$A \in \mathbb{S}^n$是**半正定**(positive semidefinite ,PSD)。这写为（或$A \succcurlyeq 0$仅$A \geq 0$），并且所有半正定矩阵的集合通常表示为$\mathbb{S}^n_{+}$。

- 同样，对称矩阵$A \in \mathbb{S}^n$是**负定**（negative definite,ND），如果对于所有非零$x \in \mathbb{R}^n$，则$x^TAx < 0$表示为$A \prec 0$（或$A < 0$）。

- 类似地，对称矩阵$A \in \mathbb{S}^n$是**半负定**(negative semidefinite,NSD），如果对于所有$x \in \mathbb{R}^n$，则$x^TAx \leq 0$表示为$A \preccurlyeq 0$（或$A \leq 0$）。

- 最后，对称矩阵$A \in \mathbb{S}^n$是**不定**的，如果它既不是正半定也不是负半定，即，如果存在$x_1, x_2 \in \mathbb{R}^n$，那么$x_1^TAx_1 > 0$且$x_2^TAx_2 < 0$。

很明显，如果A是正定的，那么$-A$是负定的，反之亦然。同样，如果A是半正定的，那么$-A$是是半负定的，反之亦然。如果果A是不定的，那么$-A$是也是不定的。

正定矩阵和负定矩阵的一个重要性质是它们总是满秩，因此是可逆的。为了了解这是为什么，假设某个矩阵$A \in \mathbb{S}^n$不是满秩。然后，假设A的第j列可以表示为其他$n-1$列的线性组合：

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

对于某些$x_1, \cdots x_{j-1}, x_{j+1}, \cdots, x_n \in \mathbb{R}$。设$x_j = -1$，则：

$$Ax = \sum_{i \neq j} x_i \, a_i = 0$$

但这意味着对于某些非零向量x，$x^TAx = 0$，因此A必须既不是正定也不是负定。如果A是正定或负定，则必须是满秩。 最后，有一种类型的正定矩阵经常出现，因此值得特别提及。 给定矩阵$A \in \mathbb{R}^{m \times n}$（不一定是对称或偶数平方），矩阵$G = A^TA$（有时称为 **Gram 矩**

阵）总是半正定的。 此外，如果$m \geq n$（同时为了方便起见，我们假设A是满秩），则$G = A^T A$是正定的。

3.12 特征值和特征向量

给定一个方阵$A \in \mathbb{R}^{n \times n}$，我们认为在以下条件下，$\lambda \in \mathbb{C}$是$A$的**特征值**，$x \in \mathbb{C}^n$是相应的**特征向量**：

$$Ax = \lambda x, x \neq 0$$

直观地说，这个定义意味着将A乘以向量x会得到一个新的向量，该向量指向与x相同的方向，但按系数λ缩放。值得注意的是，对于任何特征向量$x \in \mathbb{C}^n$和标量$t \in \mathbb{C}$，$A(cx) = cAx = c\lambda x = \lambda(cx)$，$cx$也是一个特征向量。因此，当我们讨论与$\lambda$相关的**特征向量**时，我们通常假设特征向量被标准化为长度为 1（这仍然会造成一些歧义，因为x和$-x$都是特征向量，但我们必须接受这一点）。

我们可以重写上面的等式来说明(λ, x)是A的特征值和特征向量的组合：

$$(\lambda I - A)x = 0, x \neq 0$$

但是$(\lambda I - A)x = 0$只有当$(\lambda I - A)$有一个非空零空间时，同时$(\lambda I - A)$是奇异的，x才具有非零解，即：

$$|(\lambda I - A)| = 0$$

现在，我们可以使用行列式的先前定义将表达式$|(\lambda I - A)|$扩展为λ中的（非常大的）多项式，其中，λ的度为n。它通常被称为矩阵A的特征多项式。

然后我们找到这个特征多项式的n（可能是复数）根，并用$\lambda_1, \cdots, \lambda_n$表示。这些都是矩阵$A$的特征值，但我们注意到它们可能不明显。为了找到特征值λ_i对应的特征向量，我们只需解线性方程$(\lambda I - A)x = 0$，因为$(\lambda I - A)$是奇异的，所以保证有一个非零解（但也可能有多个或无穷多个解）。

应该注意的是，这不是实际用于数值计算特征值和特征向量的方法（记住行列式的完全展开式有$n!$项），这是一个数学上的争议。

以下是特征值和特征向量的属性（所有假设在$A \in \mathbb{R}^{n \times n}$具有特征值$\lambda_1, \cdots, \lambda_n$的前提下）：

- A的迹等于其特征值之和

$$\mathrm{tr}A = \sum_{i=1}^{n} \lambda_i$$

- A的行列式等于其特征值的乘积

$$|A| = \prod_{i=1}^{n} \lambda_i$$

- A的秩等于A的非零特征值的个数

- 假设A非奇异，其特征值为λ和特征向量为x。那么$1/\lambda$是具有相关特征向量x的A^{-1}的特征值，即$A^{-1}x = (1/\lambda)x$。（要证明这一点，取特征向量方程，$Ax = \lambda x$，两边都左乘A^{-1}）

- 对角阵的特征值$d = diag(d_1, \cdots, d_n)$实际上就是对角元素$d_1, \cdots, d_n$

3.13 对称矩阵的特征值和特征向量

通常情况下，一般的方阵的特征值和特征向量的结构可以很细微地表示出来。 值得庆幸的是，在机器学习的大多数场景下，处理对称实矩阵就足够了，其处理的对称实矩阵的特征值和特征向量具有显着的特性。

在本节中，我们假设A是实对称矩阵, 具有以下属性：

1. A的所有特征值都是实数。 我们用用$\lambda_1, \cdots, \lambda_n$表示。

2. 存在一组特征向量$u_1, \cdots u_n$，对于所有i，u_i是具有特征值λ_i和b的特征向量。$u_1, \cdots u_n$是单位向量并且彼此正交。

设U是包含u_i作为列的正交矩阵：

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

设$\Lambda = diag(\lambda_1, \cdots, \lambda_n)$是包含$\lambda_1, \cdots, \lambda_n$作为对角线上的元素的对角矩阵。 使用 2.3 节的方程（2）中的矩阵 - 矩阵向量乘法的方法，我们可以验证：

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

考虑到正交矩阵U满足$UU^T = I$，利用上面的方程，我们得到：

$$A = AUU^T = U\Lambda U^T$$

这种A的新的表示形式为$U\Lambda U^T$，通常称为矩阵A的对角化。术语对角化是这样来的：通

过这种表示，我们通常可以有效地将对称矩阵A视为对角矩阵 ，这更容易理解。关于由特征向量U定义的基础， 我们将通过几个例子详细说明。

背景知识：代表另一个基的向量。

任何正交矩阵$U=\begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{bmatrix}$定义了一个新的属于$\mathbb{R}^n$的基（坐标系），意义如下：对于任何向量$x \in \mathbb{R}^n$都可以表示为$u_1$，$\cdots u_n$的线性组合，其系数为$x_1, \cdots x_n$：

$$x = \hat{x}_1 u_1 + \cdots + \cdots \hat{x}_n u_n = U\hat{x}$$

在第二个等式中，我们使用矩阵和向量相乘的方法。 实际上，这种$\hat{x}$是唯一存在的:

$$x = U\hat{x} \Leftrightarrow U^T x = \hat{x}$$

换句话说，向量$\hat{x} = U^T x$可以作为向量x的另一种表示，与U定义的基有关。

“对角化”矩阵向量乘法。 通过上面的设置，我们将看到左乘矩阵A可以被视为左乘以对角矩阵关于特征向量的基。 假设x是一个向量，$\hat{x}$表示U的基。设$z = Ax$为矩阵向量积。现在让我们计算关于U的基z： 然后，再利用$UU^T = U^T = I$和方程$A = AUU^T = U\Lambda U^T$，我们得到：

$$\hat{z} = U^T z = U^T Ax = U^T U\Lambda U^T x = \Lambda\hat{x} = \begin{bmatrix} \lambda_1 \hat{x}_1 \\ \lambda_2 \hat{x}_2 \\ \vdots \\ \lambda_n \hat{x}_n \end{bmatrix}$$

我们可以看到，原始空间中的左乘矩阵A等于左乘对角矩阵Λ相对于新的基，即仅将每个坐标缩放相应的特征值。 在新的基上，矩阵多次相乘也变得简单多了。例如，假设$q = AAAx$。根据A的元素导出q的分析形式，使用原始的基可能是一场噩梦，但使用新的基就容易多了：

$$\hat{q} = U^T q = U^T AAAx = U^T U\Lambda U^T U\Lambda U^T U\Lambda U^T x = \Lambda^3\hat{x} = \begin{bmatrix} \lambda_1^3 \hat{x}_1 \\ \lambda_2^3 \hat{x}_2 \\ \vdots \\ \lambda_n^3 \hat{x}_n \end{bmatrix}$$

“对角化”二次型。作为直接的推论，二次型$x^T Ax$也可以在新的基上简化。

$$x^T Ax = x^T U\Lambda U^T x = \hat{x}\,\Lambda\,\hat{x} = \sum_{i=1}^{n} \lambda_i \hat{x}_i^2$$

(回想一下，在旧的表示法中，$x^T Ax = \sum_{i=1,j=1}^{n} x_i\, x_j A_{ij}$涉及一个$n^2$项的和，而不是上面等式中的$n$项。)利用这个观点，我们还可以证明矩阵$A$的正定性完全取决于其特征值的符号：

1. 如果所有的$\lambda_i > 0$，则矩阵A正定的，因为对于任意的$\hat{x} \neq 0, x^T Ax = \sum_{i=1}^{n} \lambda_i \hat{x}_i^2 > 0$

2. 如果所有的$\lambda_i \geq 0$，则矩阵A是为正半定，因为对于任意的$\hat{x}, x^T Ax = \sum_{i=1}^n \lambda_i \hat{x}_i^2 \geq 0$

3. 同样，如果所有$\lambda_i < 0$或$\lambda_i \leq 0$，则矩阵A分别为负定或半负定。

4. 最后，如果A同时具有正特征值和负特征值，比如 ｌ$\lambda_i > 0$和$\lambda_j < 0$，那么它是不定的。这是因为如果我们让$\hat{x}$满足$\hat{x}_i = 1$和$\hat{x}_k = 0$，同时所有的$k \neq i$，那么$x^T Ax = \sum_{i=1}^n \lambda_i \hat{x}_i^2 > 0$,我们让$\hat{x}$满足$\hat{x}_i = 1$和$\hat{x}_k = 0$，同时所有的$k \neq i$，那么$x^T Ax = \sum_{i=1}^n \lambda_i \hat{x}_i^2 < 0$

特征值和特征向量经常出现的应用是最大化矩阵的某些函数。特别是对于矩阵$A \in \mathbb{S}^n$，考虑以下最大化问题：

$$\max_{x \in \mathbb{R}^n} x^T Ax = \sum_{i=1}^n \lambda_i \hat{x}_i^2 \quad \text{subject to} \quad \| x \|_2^2 = 1$$

也就是说，我们要找到（范数 1）的向量，它使二次型最大化。假设特征值的阶数为$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n$，此优化问题的最优值为$\lambda_1$，且与$\lambda_1$对应的任何特征向量$u_1$都是最大值之一。（如果$\lambda_1 > \lambda_2$，那么有一个与特征值$\lambda_1$对应的唯一特征向量，它是上面那个优化问题的唯一最大值。） 我们可以通过使用对角化技术来证明这一点：注意，通过公式$\| Ux \|_2 = \| x \|_2$推出$\| x \|_2 = \| \hat{x} \|_2$，并利用公式：

$x^T Ax = x^T U \Lambda U^T x = \hat{x} \Lambda \hat{x} = \sum_{i=1}^n \lambda_i \hat{x}_i^2$，我们可以将上面那个优化问题改写为：

$$\max_{\hat{x} \in \mathbb{R}^n} \hat{x}^T \Lambda \hat{x} = \sum_{i=1}^n \lambda_i \hat{x}_i^2 \quad \text{subject to} \quad \| \hat{x} \|_2^2 = 1$$

然后，我们得到目标的上界为λ_1：

$$\hat{x}^T \Lambda \hat{x} = \sum_{i=1}^n \lambda_i \hat{x}_i^2 \leq \sum_{i=1}^n \lambda_1 \hat{x}_i^2 = \lambda_1$$

此外，设置$\hat{x} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$可让上述等式成立，这与设置$x = u_1$相对应。

4. 矩阵微积分

虽然前面章节中的主题通常包含在线性代数的标准课程中，但似乎很少涉及（我们将广泛使用）的一个主题是微积分扩展到向量设置展。尽管我们使用的所有实际微积分都是相对微不足道的，但是符号通常会使事情看起来比实际困难得多。 在本节中，我们将介绍矩阵

微积分的一些基本定义，并提供一些示例。

4.1 梯度

假设$f:\mathbb{R}^{m\times n}\to\mathbb{R}$是将维度为$m\times n$的矩阵$A\in\mathbb{R}^{m\times n}$作为输入并返回实数值的函数。然后$f$的梯度（相对于$A\in\mathbb{R}^{m\times n}$）是偏导数矩阵，定义如下：

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

即，$m\times n$矩阵:

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

请注意，$\nabla_A f(A)$的维度始终与A的维度相同。特殊情况，如果A只是向量$A\in\mathbb{R}^n$，则

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

重要的是要记住，只有当函数是实值时，即如果函数返回标量值，才定义函数的梯度。例如，$A\in\mathbb{R}^{m\times n}$相对于$x$，我们不能取$Ax$的梯度，因为这个量是向量值。 它直接从偏导数的等价性质得出：

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

原则上，梯度是偏导数对多变量函数的自然延伸。然而，在实践中，由于符号的原因，使用梯度有时是很困难的。例如，假设$A\in\mathbb{R}^{m\times n}$是一个固定系数矩阵，假设$b\in\mathbb{R}^m$是一个固定系数向量。设$f:\mathbb{R}^{m\times n}\to\mathbb{R}$为$f(z)=z^Tz$定义的函数，因此$\nabla_z f(z)=2z$。但现在考虑表达式，

$$\nabla f(Ax)$$

该表达式应该如何解释？ 至少有两种可能性：1.在第一个解释中，回想起$\nabla_z f(z)=2z$。在这里，我们将$\nabla f(Ax)$解释为评估点Ax处的梯度，因此:

$$\nabla f(Ax) = 2(Ax) = 2Ax \in \mathbb{R}^m$$

2.在第二种解释中，我们将数量$f(Ax)$视为输入变量x的函数。 更正式地说，设$g(x) = f(Ax)$。 然后在这个解释中:

$$\nabla f(Ax) = \nabla_x g(x) \in \mathbb{R}^n$$

在这里，我们可以看到这两种解释确实不同。 一种解释产生m维向量作为结果，而另一种解释产生n维向量作为结果！ 我们怎么解决这个问题？

这里，关键是要明确我们要区分的变量。 在第一种情况下，我们将函数f与其参数z进行区分，然后替换参数Ax。 在第二种情况下，我们将复合函数$g(x) = f(Ax)$直接与x进行微分。

我们将第一种情况表示为$\nabla z f(Ax)$，第二种情况表示为$\nabla x f(Ax)$。

保持符号清晰是非常重要的，以后完成课程作业时候你就会发现。

4.2 黑塞矩阵

假设$f: \mathbb{R}^n \to \mathbb{R}$是一个函数，它接受$\mathbb{R}^n$中的向量并返回实数。那么关于$x$的**黑塞矩阵**（也有翻译作海森矩阵），写做：$\nabla_x^2 f(Ax)$，或者简单地说，$H$是$n \times n$矩阵的偏导数：

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

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

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

注意：黑塞矩阵通常是对称阵：

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}$$

与梯度相似，只有当$f(x)$为实值时才定义黑塞矩阵。

很自然地认为梯度与向量函数的一阶导数的相似，而黑塞矩阵与二阶导数的相似（我们使用的符号也暗示了这种关系）。 这种直觉通常是正确的，但需要记住以下几个注意事项。首先，对于一个变量$f: \mathbb{R} \to \mathbb{R}$的实值函数，它的基本定义：二阶导数是一阶导数的导数，即：

$$\frac{\partial^2 f(x)}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial}{\partial x} f(x)$$

然而，对于向量的函数，函数的梯度是一个向量，我们不能取向量的梯度，即:

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

上面这个表达式没有意义。 因此，黑塞矩阵不是梯度的梯度。 然而，下面这种情况却这几乎是正确的：如果我们看一下梯度$(\nabla_x f(x))_i = \partial f(x)/\partial x_i$的第$i$个元素，并取关于于$x$的梯度我们得到：

$$\nabla_x \frac{\partial f(x)}{\partial x_i} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_i \partial x_1} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_i \partial x_n} \end{bmatrix}$$

这是黑塞矩阵第i行（列）,所以：

$$\nabla_x^2 f(x) = [\nabla_x(\nabla_x f(x))_1 \quad \nabla_x(\nabla_x f(x))_2 \quad \cdots \quad \nabla_x(\nabla_x f(x))_n]$$

简单地说：我们可以说由于：$\nabla_x^2 f(x) = \nabla_x(\nabla_x f(x))^T$，只要我们理解，这实际上是取$\nabla_x f(x)$的每个元素的梯度，而不是整个向量的梯度。

最后，请注意，虽然我们可以对矩阵$A \in \mathbb{R}^n$取梯度，但对于这门课，我们只考虑对向量$x \in \mathbb{R}^n$取黑塞矩阵。 这会方便很多（事实上，我们所做的任何计算都不要求我们找到关于矩阵的黑森方程），因为关于矩阵的黑塞方程就必须对矩阵所有元素求偏导数$\partial^2 f(A)/\left(\partial A_{ij} \partial A_{k\ell}\right)$，将其表示为矩阵相当麻烦。

4.3 二次函数和线性函数的梯度和黑塞矩阵

现在让我们尝试确定几个简单函数的梯度和黑塞矩阵。 应该注意的是，这里给出的所有梯度都是 **CS229** 讲义中给出的梯度的特殊情况。

对于$x \in \mathbb{R}^n$, 设$f(x) = b^T x$ 的某些已知向量$b \in \mathbb{R}^n$ ，则：

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

所以：

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

由此我们可以很容易地看出$\nabla_x b^T x = b$。这应该与单变量微积分中的类似情况进行比较，其中$\partial/(\partial x)ax = a$。 现在考虑$A \in \mathbb{S}^n$的二次函数$f(x) = x^T Ax$。 记住这一点：

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

为了取偏导数，我们将分别考虑包括x_k和x_2^k因子的项：

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

最后一个等式，是因为A是对称的（我们可以安全地假设，因为它以二次形式出现）。注意，$\nabla_x f(x)$的第k个元素是A和x的第k行的内积。 因此，$\nabla_x x^T Ax = 2Ax$。 同样，这应该提醒你单变量微积分中的类似事实，即$\partial/(\partial x)ax^2 = 2ax$。

最后，让我们来看看二次函数$f(x) = x^T Ax$黑塞矩阵（显然，线性函数$b^T x$的黑塞矩阵为零）。在这种情况下:

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_\ell} = \frac{\partial}{\partial x_k}\left[\frac{\partial f(x)}{\partial x_\ell}\right] = \frac{\partial}{\partial x_k}\left[2\sum_{i=1}^{n} A_{\ell i}\, x_i\right] = 2A_{\ell k} = 2A_{k\ell}$$

因此，应该很清楚$\nabla_x^2 x^T Ax = 2A$，这应该是完全可以理解的（同样类似于$\partial^2/(\partial x^2)ax^2 = 2a$的单变量事实）。

简要概括起来：

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

4.4 最小二乘法

让我们应用上一节中得到的方程来推导最小二乘方程。假设我们得到矩阵$A \in \mathbb{R}^{m\times n}$（为了简单起见，我们假设$A$是满秩）和向量$b \in \mathbb{R}^m$，从而使$b \notin \mathcal{R}(A)$。在这种情况下，我们将无法找到向量$x \in \mathbb{R}^n$，由于$Ax = b$，因此我们想要找到一个向量$x$，使得$Ax$尽可能接近 b，用欧几里德范数的平方$\| Ax - b \|_2^2$来衡量。

使用公式$\| x \|^2 = x^T x$，我们可以得到：

$$\begin{aligned} \| Ax - b \|_2^2 &= (Ax - b)^T (Ax - b) \\ &= x^T A^T Ax - 2b^T Ax + b^T b \end{aligned}$$

根据x的梯度，并利用上一节中推导的性质：

$$\begin{aligned} \nabla_x (x^T A^T Ax - 2b^T Ax + b^T b) &= \nabla_x x^T A^T Ax - \nabla_x 2b^T Ax + \nabla_x b^T b \\ &= 2A^T Ax - 2A^T b \end{aligned}$$

将最后一个表达式设置为零，然后解出x，得到了正规方程：

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

这和我们在课堂上得到的相同。

4.5 行列式的梯度

现在让我们考虑一种情况，我们找到一个函数相对于矩阵的梯度，也就是说，对于$A \in \mathbb{R}^{n\times n}$，我们要找到$\nabla_A |A|$。回想一下我们对行列式的讨论：

$$|A| = \sum_{i=1}^{n} (-1)^{i+j} A_{ij} \left| A_{\setminus i, \setminus j} \right| \quad (\text{ for any } j \in 1, \dots, n)$$

所以：

$$\frac{\partial}{\partial A_{k\ell}} |A| = \frac{\partial}{\partial A_{k\ell}} \sum_{i=1}^{n} (-1)^{i+j} A_{ij} \left| A_{\setminus i, \setminus j} \right| = (-1)^{k+\ell} \left| A_{\setminus k, \setminus \ell} \right| = (\text{adj}(A))_{\ell k}$$

从这里可以知道，它直接从伴随矩阵的性质得出：

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

现在我们来考虑函数$f: \mathbb{S}_{++}^n \to \mathbb{R}$，$f(A) = \log |A|$。注意，我们必须将$f$的域限制为正定矩阵，因为这确保了$|A| > 0$，因此$|A|$的对数是实数。在这种情况下，我们可以使用链式法则（没什么奇怪的，只是单变量演算中的普通链式法则）来看看：

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

从这一点可以明显看出：

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

我们可以在最后一个表达式中删除转置，因为A是对称的。注意与单值情况的相似性，其中$\partial/(\partial x)\log x = 1/x$。

4.6 特征值优化

最后，我们使用矩阵演算以直接导致特征值/特征向量分析的方式求解优化问题。 考虑以下等式约束优化问题：

$$\max_{x \in \mathbb{R}^n} x^T A x \quad \text{subject to} \quad \| x \|_2^2 = 1$$

对于对称矩阵$A \in \mathbb{S}^n$。求解等式约束优化问题的标准方法是采用**拉格朗日**形式，一种包含等式约束的目标函数，在这种情况下，拉格朗日函数可由以下公式给出：

$$\mathcal{L}(x, \lambda) = x^T A x - \lambda x^T x$$

其中，λ被称为与等式约束关联的拉格朗日乘子。可以确定，要使x^*成为问题的最佳点，拉格朗日的梯度必须在x^*处为零（这不是唯一的条件，但它是必需的）。也就是说，

$$\nabla_x \mathcal{L}(x, \lambda) = \nabla_x (x^T A x - \lambda x^T x) = 2A^T x - 2\lambda x = 0$$

请注意，这只是线性方程$Ax = \lambda x$。 这表明假设$x^T x = 1$，可能最大化（或最小化）$x^T A x$的唯一点是A的特征向量。

CS229 机器学习课程复习材料-概率论

这部分是斯坦福大学 CS229 机器学习课程的基础材料，原始文件下载

原文作者：Arian Maleki ， Tom Do

翻译：石振宇

审核和修改制作：黄海广

概率论是对不确定性的研究。通过这门课，我们将依靠概率论中的概念来推导机器学习算法。这篇笔记试图涵盖适用于 **CS229** 的概率论基础。概率论的数学理论非常复杂，并且涉及到“分析”的一个分支：测度论。在这篇笔记中，我们提供了概率的一些基本处理方法，但是不会涉及到这些更复杂的细节。

1. 概率的基本要素

为了定义集合上的概率，我们需要一些基本元素：

- 样本空间Ω：随机实验的所有结果的集合。在这里，每个结果 $w \in \Omega$ 可以被认为是实验结束时现实世界状态的完整描述。
- 事件集（事件空间）$\mathcal{F}$：元素 $A \in \mathcal{F}$ 的集合（称为事件）是 Ω 的子集（即每个 $A \subseteq \Omega$ 是一个实验可能结果的集合）。

  备注：$\mathcal{F}$需要满足以下三个条件：

  (1) $\emptyset \in \mathcal{F}$

  (2) $A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}$

  (3) $A_1, A_2, \cdots A_i \in \mathcal{F} \Rightarrow \cup_i A_i \in \mathcal{F}$
- 概率度量P：函数P是一个$\mathcal{F} \to \mathbb{R}$的映射，满足以下性质：
  - 对于每个 $A \in \mathcal{F}$，$P(A) \geq 0$,
  - $P(\Omega) = 1$
  - 如果$A_1, A_2, \cdots$ 是互不相交的事件（即 当$i \neq j$时，$A_i \cap A_j = \emptyset$ ），那么：

$$P\left(\bigcup_i A_i\right) = \sum_i P\left(A_i\right)$$

以上三条性质被称为**概率公理**。

举例：

考虑投掷六面骰子的事件。样本空间为$\Omega = \{1，2，3，4，5，6\}$。最简单的事件空间是平凡事件空间$\mathcal{F} = \{\emptyset, \Omega\}$.另一个事件空间是$\Omega$的所有子集的集合。对于第一个事件空间，满足上述要求的唯一概率度量由$P(\emptyset) = 0$，$p(\Omega) = 1$给出。对于第二个事件空间，一个有效的概率度量是将事件空间中每个事件的概率分配为$i/6$，这里i 是这个事件集合中元素的数量；例如$P(\{1,2,3,4\}) = 4/6$，$P(\{1,2,3\}) = 3/6$。

性质：

- 如果$A \subseteq B$，则：$P(A) \leq P(B)$
- $P(A \cap B) \leq min(P(A), P(B))$
- （布尔不等式）：$P(A \cup B) \leq P(A) + P(B)$
- $P(\Omega|A) = 1 - P(A)$
- （全概率定律）：如果A_1，…，A_k是一些互不相交的事件并且它们的并集是Ω，那么它们的概率之和是 1

1.1 条件概率和独立性

假设B是一个概率非 0 的事件，我们定义在给定B的条件下A 的条件概率为：

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}$$

换句话说，$P(A|B)$是度量已经观测到B事件发生的情况下A事件发生的概率，两个事件被称为独立事件当且仅当$P(A \cap B) = P(A)P(B)$（或等价地，$P(A|B) = P(A)$）。因此，独立性相当于是说观察到事件B对于事件A的概率没有任何影响。

2. 随机变量

考虑一个实验，我们翻转 10 枚硬币，我们想知道正面硬币的数量。这里，样本空间Ω的元素是长度为 10 的序列。例如，我们可能有$w_0 = \{H，H，T，H，T，H，H，T，T，T\} \in$

Ω。然而，在实践中，我们通常不关心获得任何特定正反序列的概率。相反，我们通常关心结果的实值函数，比如我们 10 次投掷中出现的正面数，或者最长的背面长度。在某些技术条件下，这些函数被称为**随机变量**。

更正式地说，随机变量X是一个的$\Omega \to \mathbb{R}$函数。通常，我们将使用大写字母$X(\omega)$或更简单的X(其中隐含对随机结果ω的依赖)来表示随机变量。我们将使用小写字母x来表示随机变量的值。

举例： 在我们上面的实验中，假设$X(\omega)$是在投掷序列ω中出现的正面的数量。假设投掷的硬币只有 10 枚，那么$X(\omega)$只能取有限数量的值，因此它被称为**离散随机变量**。这里，与随机变量X相关联的集合取某个特定值k的概率为：

$$P(X = k) := P(\{\omega: X(\omega) = k\})$$

举例： 假设$X(\omega)$是一个随机变量，表示放射性粒子衰变所需的时间。在这种情况下，$X(\omega)$具有无限多的可能值，因此它被称为**连续随机变量**。我们将X在两个实常数a和b之间取值的概率(其中$a < b$)表示为：

$$P(a \le X \le b) := P(\{\omega: a \le X(\omega) \le b\})$$

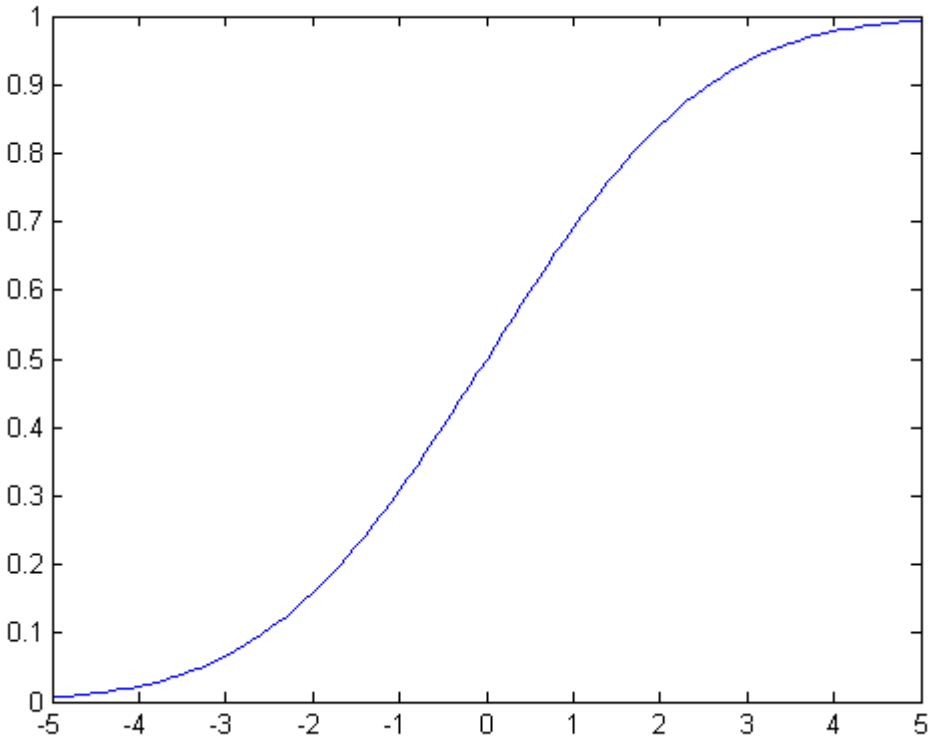
2.1 累积分布函数

为了指定处理随机变量时使用的概率度量，通常可以方便地指定替代函数(**CDF**、**PDF** 和 **PMF**)，在本节和接下来的两节中，我们将依次描述这些类型的函数。

累积分布函数(CDF)是函数$F_X: \mathbb{R} \to [0,1]$，它将概率度量指定为：

$$F_X(x) \triangleq P(X \le x)$$

通过使用这个函数，我们可以计算任意事件发生的概率。图 1 显示了一个样本 **CDF** 函数。



性质：

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

2.2 概率质量函数

当随机变量X取有限种可能值(即，X是离散随机变量)时，表示与随机变量相关联的概率度量的更简单的方法是直接指定随机变量可以假设的每个值的概率。特别地，概率质量函数(**PMF**)是函数 $p_X: \Omega \to \mathbb{R}$，这样：

$$p_X(x) \triangleq P(X = x)$$

在离散随机变量的情况下，我们使用符号$Val(X)$表示随机变量X可能假设的一组可能值。例如，如果$X(\omega)$是一个随机变量，表示十次投掷硬币中的正面数，那么$Val(X) = \{0，1，2，\ldots，10\}$。

性质：

- $0 \leq p_X(x) \leq 1$
- $\sum_{x \in V \text{ al } (X)} p_X(x) = 1$
- $\sum_{x \in A} p_X(x) = P(X \in A)$

2.3 概率密度函数

对于一些连续随机变量，累积分布函数$F_X(x)$处可微。在这些情况下，我们将**概率密度函数(PDF)**定义为累积分布函数的导数，即：

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

请注意，连续随机变量的概率密度函数可能并不总是存在的(即，如果它不是处处可微)。

根据微分的性质，对于很小的Δx，

$$P(x \leq X \leq x + \Delta x) \approx f_X(x)\Delta x$$

CDF 和 **PDF(**当它们存在时！)都可用于计算不同事件的概率。但是应该强调的是，任意给定点的**概率密度函数(PDF)**的值不是该事件的概率，即$f_X(x) \not= P(X = x)$。例如，$f_X(x)$可以取大于 1 的值(但是$f_X(x)$在$\mathbb{R}$的任何子集上的积分最多为 1)。

性质：

$$f_X(x) \geq 0$$

$$\int_{-\infty}^{\infty} f_X(x) = 1$$

$$\int_{x \in A} f_X(x)dx = P(X \in A)$$

2.4 期望

假设X是一个离散随机变量，其 **PMF** 为 $p_X(x)$，$g: \mathbb{R} \to \mathbb{R}$是一个任意函数。在这种情况下，$g(X)$可以被视为随机变量，我们将$g(X)$的期望值定义为：

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

如果X是一个连续的随机变量，其 **PDF** 为$f_X(x)$，那么$g(X)$的期望值被定义为：

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

直觉上，$g(X)$的期望值可以被认为是$g(x)$对于不同的x值可以取的值的“加权平均值”，

其中权重由$p_X(x)$或$f_X(x)$给出。作为上述情况的特例，请注意，随机变量本身的期望值，是通过令$g(x) = x$得到的，这也被称为随机变量的平均值。

性质：

- 对于任意常数 $a \in \mathbb{R}$，$E[a] = a$
- 对于任意常数 $a \in \mathbb{R}$，$E[af(X)] = aE[f(X)]$
- （线性期望）：$E[f(X) + g(X)] = E[f(X)] + E[g(X)]$
- 对于一个离散随机变量X，$E[1\{X = k\}] = P(X = k)$

2.5 方差

随机变量X的**方差**是随机变量X的分布围绕其平均值集中程度的度量。形式上，随机变量X的方差定义为：

$$\text{Var}[X] \triangleq E[(X - E(X))^2]$$

使用上一节中的性质，我们可以导出方差的替代表达式:

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

其中第二个等式来自期望的线性，以及$E[X]$相对于外层期望实际上是常数的事实。

性质：

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

举例：

计算均匀随机变量X的平均值和方差，任意$x \in [0，1]$，其 PDF 为 $p_X(x) = 1$，其他地方为 0。

$$E[X] = \int_{-\infty}^{\infty} x\, f_X(x)dx = \int_0^1 x\, dx = \frac{1}{2}$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2\, f_X(x)dx = \int_0^1 x^2\, dx = \frac{1}{3}$$

$$Var[X] = E[X^2] - E[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

举例：

假设对于一些子集$A \subseteq \Omega$，有$g(x) = 1\{x \in A\}$，计算$E[g(X)]$?

离散情况：

$$E[g(X)] = \sum_{x \in Val(X)} 1\{x \in A\} P_X(x) dx = \sum_{x \in A} P_X(x) dx = P(x \in A)$$

连续情况：

$$E[g(X)] = \int_{-\infty}^{\infty} 1\{x \in A\} f_X(x) dx = \int_{x \in A} f_X(x) dx = P(x \in A)$$

2.6 一些常见的随机变量

离散随机变量

- 伯努利分布：硬币掷出正面的概率为p（其中：$0 \leq p \leq 1$），如果正面发生，则为1，否则为0。

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

- 二项式分布：掷出正面概率为p（其中：$0 \leq p \leq 1$）的硬币n次独立投掷中正面的数量。

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

- 几何分布：掷出正面概率为p（其中：$p > 0$）的硬币第一次掷出正面所需要的次数。

- 泊松分布：用于模拟罕见事件频率的非负整数的概率分布（其中：$\lambda > 0$）。

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

连续随机变量

- 均匀分布：在a和b之间每个点概率密度相等的分布（其中：$a < b$）。

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

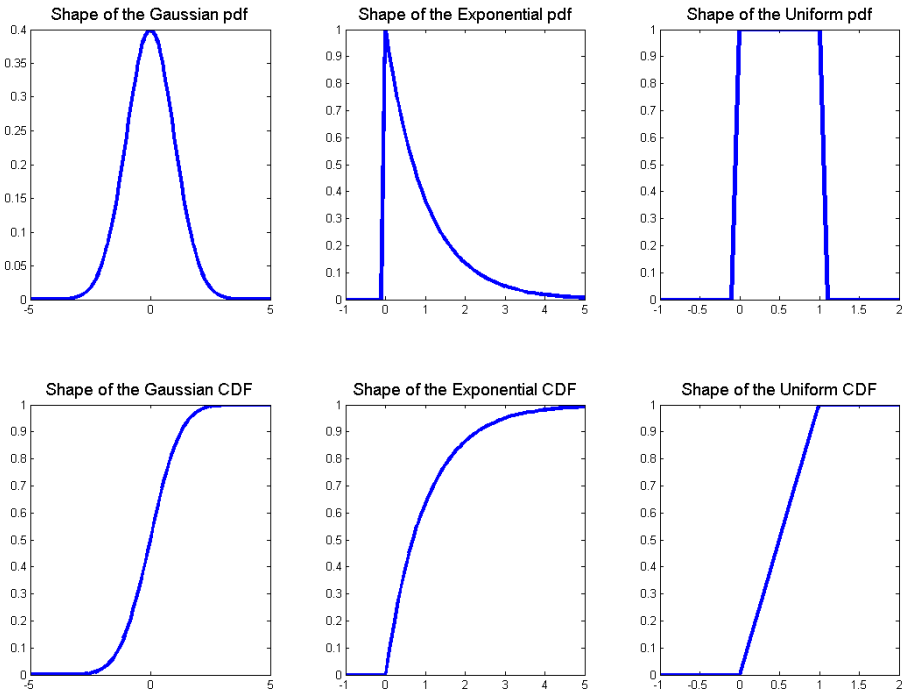
- 指数分布：在非负实数上有衰减的概率密度（其中：$\lambda > 0$）。

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

- 正态分布：又被称为高斯分布。

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

一些随机变量的概率密度函数和累积分布函数的形状如图 2 所示。



下表总结了这些分布的一些特性：

分布	概率密度函数(PDF)或者概率质量函数(**PMF**)	均值	方差
$Bernoulli(p)$(伯努利分布)	$\begin{cases} p & \text{if } x=1 \\ 1-p & \text{if } x=0 \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$(二项式分布)	$\binom{n}{k}p^k(1-p)^{n-k}$ 其中：$0 \le k \le n$	np	npq
$Geometric(p)$(几何分布)	$p(1-p)^{k-1}$ 其中：$k=1,2,\cdots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$(泊松分布)	$e^{-\lambda}\lambda^x/x!$ 其中：$k=1,2,\cdots$	λ	λ
$Uniform(a,b)$(均匀分布)	$\frac{1}{b-a}$ 存在$x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu,\sigma^2)$(高斯分布)	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	μ	σ^2
$Exponential(\lambda)$(指数分布)	$\lambda e^{-\lambda x}$ $x \ge 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

3. 两个随机变量

到目前为止，我们已经考虑了单个随机变量。然而，在许多情况下，在随机实验中，我

们可能有不止一个感兴趣的量。例如，在一个我们掷硬币十次的实验中，我们可能既关心$X(\omega)$ =出现的正面数量，也关心$Y(\omega)$ =连续最长出现正面的长度。在本节中，我们考虑两个随机变量的设置。

3.1 联合分布和边缘分布

假设我们有两个随机变量，一个方法是分别考虑它们。如果我们这样做，我们只需要$F_X(x)$和$F_Y(y)$。但是如果我们想知道在随机实验的结果中，X和Y同时假设的值，我们需要一个更复杂的结构，称为X和Y的**联合累积分布函数**，定义如下:

$$F_{XY}(x,y) = P(X \leq x, Y \leq y)$$

可以证明，通过了解联合累积分布函数，可以计算出任何涉及到X和Y的事件的概率。

联合 **CDF**: $F_{XY}(x,y)$和每个变量的联合分布函数$F_X(x)$和$F_Y(y)$分别由下式关联:

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

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

这里我们称$F_X(x)$和$F_Y(y)$为 $F_{XY}(x,y)$的**边缘累积概率分布函数**。

性质：

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

3.2 联合概率和边缘概率质量函数

如果X和Y是离散随机变量，那么**联合概率质量函数** $p_{XY}: \mathbb{R} \times \mathbb{R} \to [0,1]$由下式定义：

$$p_{XY}(x,y) = P(X = x, Y = y)$$

这里，对于任意x，y，$0 \leq P_{XY}(x,y) \leq 1$，并且 $\sum_{x\in Val(X)} \sum_{y\in Val(Y)} P_{XY}(x,y) = 1$

两个变量上的**联合 PMF** 分别与每个变量的概率质量函数有什么关系？事实上：

$$p_X(x) = \sum_y p_{XY}(x,y)$$

对于$p_Y(y)$类似。在这种情况下，我们称$p_X(x)$为X的边际概率质量函数。在统计学中，

将一个变量相加形成另一个变量的边缘分布的过程通常称为“边缘化”。

3.3 联合概率和边缘概率密度函数

假设X和Y是两个连续的随机变量，具有联合分布函数F_{XY}。在$F_{XY}(x,y)$在x和y中处处可微的情况下，我们可以定义**联合概率密度函数**：

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

如同在一维情况下，$f_{XY}(x,y)\not= P(X = x,Y = y)$，而是：

$$\iint_{x\in A} f_{XY}\,(x,y)dxdy=P((X,Y)\in A)$$

请注意，概率密度函数$f_{XY}(x,y)$的值总是非负的，但它们可能大于 1。尽管如此，可以肯定的是 $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}\,(x,y)=1$

与离散情况相似，我们定义:

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

作为X的**边际概率密度函数**(或**边际密度**)，对于$f_Y(y)$也类似。

3.4 条件概率分布

条件分布试图回答这样一个问题，当我们知道X必须取某个值x时，Y上的概率分布是什么？在离散情况下，给定Y的条件概率质量函数是简单的：

$$p_{Y|X}(y|x)=\frac{p_{XY}(x,y)}{p_X(x)}$$

假设分母不等于 0。

在连续的情况下，在技术上要复杂一点，因为连续随机变量的概率等于零。忽略这一技术点，我们通过类比离散情况，简单地定义给定$X=x$的条件概率密度为：

$$f_{Y|X}(y|x)=\frac{f_{XY}(x,y)}{f_X(x)}$$

假设分母不等于 0。

3.5 贝叶斯定理

当试图推导一个变量给定另一个变量的条件概率表达式时，经常出现的一个有用公式是

贝叶斯定理。

对于离散随机变量X和Y：

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

对于连续随机变量X和Y：

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

3.6 独立性

如果对于X和Y的所有值，$F_{XY}(x,y) = F_X(x)F_Y(y)$，则两个随机变量$X$和$Y$是独立的。等价地，

- 对于离散随机变量，对于任意$x \in Val(X)$，$y \in Val(Y)$，$p_{XY}(x,y) = p_X(x)p_Y(y)$。
- 对于离散随机变量，$p_Y|X(y|x) = p_Y(y)$当对于任意$y \in Val(Y)$且$p_X(x) \neq 0$。
- 对于连续随机变量，$f_{XY}(x,y) = f_X(x)f_Y(y)$ 对于任意 $x, y \in \mathbb{R}$。
- 对于连续随机变量，$f_{Y|X}(y|x) = f_Y(y)$，当$f_X(x) \neq 0$对于任意$y \in \mathbb{R}$。

非正式地说，如果“知道”一个变量的值永远不会对另一个变量的条件概率分布有任何影响，那么两个随机变量X和Y是独立的，也就是说，你只要知道$f(x)$和$f(y)$就知道关于这对变量(X, Y)的所有信息。以下引理将这一观察形式化:

引理 3.1

如果X和Y是独立的，那么对于任何$A, B \subseteq \mathbb{R}$，我们有：

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

利用上述引理，我们可以证明如果X与Y无关，那么X的任何函数都与Y的任何函数无关。

3.7 期望和协方差

假设我们有两个离散的随机变量X, Y并且$g: \mathbf{R}^2 \to \mathbf{R}$是这两个随机变量的函数。那么$g$的期望值以如下方式定义：

$$E[g(X,Y)] \triangleq \sum_{x \in Val(X)} \sum_{y \in Val(Y)} g(x,y)p_{XY}(x,y)$$

对于连续随机变量X，Y，类似的表达式是：

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

我们可以用期望的概念来研究两个随机变量之间的关系。特别地，两个随机变量的**协方差**定义为：

$$Cov[X,Y] \triangleq E[(X - E[X])(Y - E[Y])]$$

使用类似于方差的推导，我们可以将它重写为：

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

在这里，说明两种协方差形式相等的关键步骤是第三个等号，在这里我们使用了这样一个事实，即$E[X]$和$E[Y]$实际上是常数，可以被提出来。当$cov[X，Y] = 0$时，我们说X和Y不相关。

性质：

- （期望线性）$E[f(X,Y) + g(X,Y)] = E[f(X,Y)] + E[g(X,Y)]$
- $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X,Y]$
- 如果X和Y相互独立，那么 $Cov[X,Y] = 0$
- 如果X和Y相互独立，那么 $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$.

4. 多个随机变量

上一节介绍的概念和想法可以推广到两个以上的随机变量。特别是，假设我们有n个连续随机变量，$X_1(\omega), X_2(\omega), \cdots X_n(\omega)$。在本节中，为了表示简单，我们只关注连续的情况，对离散随机变量的推广工作类似。

4.1 基本性质

我们可以定义$X_1, X_2, \cdots, X_n$的**联合累积分布函数**、**联合概率密度函数**，以及给定$X_2, \cdots, X_n$时X_1的**边缘概率密度函数**为：

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n)$$

$$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots x_n)=\frac{\partial^n F_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots x_n)}{\partial x_1\ldots\partial x_n}$$

$$f_{X_1}(X_1)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots x_n)dx_2\ldots dx_n$$

$$f_{X_1|X_2,\ldots,X_n}(x_1|x_2,\ldots x_n)=\frac{f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots x_n)}{f_{X_2,\ldots,X_n}(x_1,x_2,\ldots x_n)}$$

为了计算事件$A\subseteq\mathbb{R}^n$的概率，我们有：

$$P\big((x_1,x_2,\ldots x_n)\in A\big)=\int_{(x_1,x_2,\ldots x_n)\in A}f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots x_n)dx_1dx_2\ldots dx_n$$

链式法则：

从多个随机变量的条件概率的定义中，可以看出：

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

独立性:对于多个事件，$A_1,\cdots,A_k$,我们说$A_1,\cdots,A_k$ 是相互独立的,当对于任何子集$S\subseteq\{1，2,\cdots,k\}$，我们有：

$$P\left(\bigcap_{i\in S}A_i\right)=\prod_{i\in S}P(A_i)$$

同样，我们说随机变量$X_1,X_2,\cdots,X_n$是独立的，如果：

$$f(x_1,\cdots,x_n)=f(x_1)f(x_2)\cdots f(x_n)$$

这里，相互独立性的定义只是两个随机变量独立性到多个随机变量的自然推广。

独立随机变量经常出现在机器学习算法中，其中我们假设属于训练集的训练样本代表来自某个未知概率分布的独立样本。为了明确独立性的重要性，考虑一个“坏的”训练集，我们首先从某个未知分布中抽取一个训练样本$(x^{(1)},y^{(1)})$，然后将完全相同的训练样本的$m-1$个副本添加到训练集中。在这种情况下，我们有：

$$P\left(\left(x^{(1)},y^{(1)}\right),\ldots.\left(x^{(m)},y^{(m)}\right)\right)\neq\prod_{i=1}^{m}P\left(x^{(i)},y^{(i)}\right)$$

尽管训练集的大小为m，但这些例子并不独立！虽然这里描述的过程显然不是为机器学习算法建立训练集的明智方法，但是事实证明，在实践中，样本的不独立性确实经常出现，并且它具有减小训练集的“有效大小”的效果。

4.2 随机向量

假设我们有 n 个随机变量。当把所有这些随机变量放在一起工作时，我们经常会发现把它们放在一个向量中是很方便的...我们称结果向量为随机向量(更正式地说，随机向量是从Ω到$\mathbb{R}^n$的映射)。应该清楚的是，随机向量只是处理n个随机变量的一种替代符号，因此联合概率密度函数和综合密度函数的概念也将适用于随机向量。

期望:

考虑$g: \mathbb{R}^n \to \mathbb{R}$中的任意函数。这个函数的期望值被定义为

$$\begin{aligned} E[g(X)] &= \int_{\mathbb{R}^n} g(x_1, x_2, \dots, x_n) f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots x_n) dx_1 dx_2 \dots dx_n E[g(X)] \\ &= \int_{\mathbb{R}^n} g(x_1, x_2, \dots, x_n) f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots x_n) dx_1 dx_2 \dots dx_n \end{aligned}$$

其中，$\int_{\mathbb{R}^n}$是从$-\infty$到∞的n个连续积分。如果g是从$\mathbb{R}^n$到$\mathbb{R}^m$的函数，那么g的期望值是输出向量的元素期望值，即，如果g是：

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

那么，

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

协方差矩阵：对于给定的随机向量$X: \Omega \to \mathbb{R}^n$，其协方差矩阵$\Sigma$是$n \times n$平方矩阵，其输入由$\Sigma_{ij} = Cov[X_i, X_j]$给出。从协方差的定义来看，我们有：

$$\Sigma = \begin{bmatrix} Cov[X_1, X_1] & \cdots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \cdots & Cov[X_n, X_n] \end{bmatrix}$$
$$= \begin{bmatrix} E[X_1^2] - E[X_1]E[X_1] & \cdots & E[X_1X_n] - E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_nX_1] - E[X_n]E[X_1] & \cdots & E[X_n^2] - E[X_n]E[X_n] \end{bmatrix}$$
$$= \begin{bmatrix} E[X_1^2] & \cdots & E[X_1X_n] \\ \vdots & \ddots & \vdots \\ E[X_nX_1] & \cdots & E[X_n^2] \end{bmatrix} - \begin{bmatrix} E[X_1]E[X_1] & \cdots & E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n]E[X_1] & \cdots & E[X_n]E[X_n] \end{bmatrix}$$
$$= E[XX^T] - E[X]E[X]^T = \ldots = E[(X - E[X])(X - E[X])^T]$$

其中矩阵期望以明显的方式定义。 协方差矩阵有许多有用的属性:

- $\Sigma \succcurlyeq 0$；也就是说，Σ是正半定的。
- $\Sigma = \Sigma^T$；也就是说，Σ是对称的。

4.3 多元高斯分布

随机向量上概率分布的一个特别重要的例子叫做多元高斯或多元正态分布。随机向量$X \in \mathbb{R}^n$被认为具有多元正态(或高斯)分布，当其具有均值$\mu \in \mathbb{R}^n$和协方差矩阵$\Sigma \in \mathbb{S}_{++}^n$(其中$\mathbb{S}_{++}^n$指对称正定$n \times n$矩阵的空间)

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

我们把它写成$X \sim \mathcal{N}(\mu, \Sigma)$。请注意，在$n = 1$的情况下，它降维成普通正态分布，其中均值参数为$\mu_1$，方差为$\Sigma_{11}$。

一般来说，高斯随机变量在机器学习和统计中非常有用，主要有两个原因：

首先，在统计算法中对“噪声”建模时，它们非常常见。通常，噪声可以被认为是影响测量过程的大量小的独立随机扰动的累积；根据中心极限定理，独立随机变量的总和将趋向于“看起来像高斯”。

其次，高斯随机变量便于许多分析操作，因为实际中出现的许多涉及高斯分布的积分都有简单的封闭形式解。我们将在本课程稍后遇到这种情况。

5. 其他资源

一本关于 **CS229** 所需概率水平的好教科书是谢尔顿·罗斯的《概率第一课》(*A First Course on Probability* by Sheldon Ross)。

机器学习的数学基础（国内教材）

高等数学

1.导数定义：

导数和微分的概念

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

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

2.左右导数导数的几何意义和物理意义

函数$f(x)$在x_0处的左、右导数分别定义为：

左导数：$$f'_-(x_0)=\lim_{\Delta x\to 0^-}\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}=\lim_{x\to x_0^-}\frac{f(x)-f(x_0)}{x-x_0},(x=x_0+\Delta x)$$

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

3.函数的可导性与连续性之间的关系

Th1：函数$f(x)$在x_0处可微⇔ $f(x)$在x_0处可导。

Th2:若函数在点x_0处可导，则$y=f(x)$在点x_0处连续，反之则不成立。即函数连续不一定可导。

Th3:$f'(x_0)$存在⇔ $f'_-(x_0)=f'_+(x_0)$

4.平面曲线的切线和法线

切线方程 ： $y-y_0=f'(x_0)(x-x_0)$

法线方程：$y-y_0=-\frac{1}{f'(x_0)}(x-x_0),f'(x_0)\neq 0$

5.四则运算法则

设函数$u=u(x)$，$v=v(x)$在点x可导，则：

(1) $(u \pm v)' = u' \pm v'$

(2) $(uv)' = uv' + vu'$ $\quad d(uv) = udv + vdu$

(3) $(\frac{u}{v})' = \frac{vu\prime - uv\prime}{v^2}(v \neq 0)$ $\quad d(\frac{u}{v}) = \frac{vdu - udv}{v^2}$

6. 基本导数与微分表

(1) $y = c$（常数） 则： $y' = 0$ $\quad dy = 0$

(2) $y = x^{\alpha}$(α为实数) 则： $y' = \alpha x^{\alpha-1}$ $\quad dy = \alpha x^{\alpha-1}dx$

(3) $y = a^x$ 则： $y' = a^x \ln a$ $\quad dy = a^x \ln a dx$ 特例：$(\mathrm{e}^x)' = \mathrm{e}^x$ $\quad d(\mathrm{e}^x) = \mathrm{e}^x dx$

(4) $y = \log_a x$ 则：

$y' = \frac{1}{x\ln a}$ $\quad dy = \frac{1}{x\ln a}dx$ 特例：$y = \ln x$ $\quad (\ln x)' = \frac{1}{x}$ $\quad d(\ln x) = \frac{1}{x}dx$

(5) $y = \sin x$ 则：$y' = \cos x$ $\quad d(\sin x) = \cos x dx$

(6) $y = \cos x$ 则：$y' = -\sin x$ $\quad d(\cos x) = -\sin x dx$

(7) $y = \tan x$ 则： $y' = \frac{1}{\cos^2 x} = \sec^2 x$ $\quad d(\tan x) = \sec^2 x dx$

(8) $y = \cot x$ 则：$y' = -\frac{1}{\sin^2 x} = -\csc^2 x$ $\quad d(\cot x) = -\csc^2 x dx$

(9) $y = \sec x$ 则：$y' = \sec x \tan x$ $\quad d(\sec x) = \sec x \tan x dx$

(10) $y = \csc x$ 则：$y' = -\csc x \cot x$ $\quad d(\csc x) = -\csc x \cot x dx$

(11) $y = \arcsin x$ 则：$y' = \frac{1}{\sqrt{1-x^2}}$ $\quad d(\arcsin x) = \frac{1}{\sqrt{1-x^2}}dx$

(12) $y = \arccos x$ 则：$y' = -\frac{1}{\sqrt{1-x^2}}$ $\quad d(\arccos x) = -\frac{1}{\sqrt{1-x^2}}dx$

(13) $y = \arctan x$ 则：$y' = \frac{1}{1+x^2}$ $\quad d(\arctan x) = \frac{1}{1+x^2}dx$

(14) $y = \operatorname{arccot} x$ 则：$y' = -\frac{1}{1+x^2}$ $\quad d(\operatorname{arccot} x) = -\frac{1}{1+x^2}dx$

(15) $y=shx$ 则：$y'=chx$　　$d(shx)=chxdx$

(16) $y=chx$ 则：$y'=shx$　　$d(chx)=shxdx$

7.复合函数，反函数，隐函数以及参数方程所确定的函数的微分法

(1) 反函数的运算法则：设$y=f(x)$在点x的某邻域内单调连续，在点x处可导且$f'(x)\neq 0$，则其反函数在点x所对应的y处可导，并且有$\frac{dy}{dx}=\frac{1}{\frac{dx}{dy}}$

(2) 复合函数的运算法则:若$\mu=\varphi(x)$在点x可导,而$y=f(\mu)$在对应点$\mu(\mu=\varphi(x))$可导,则复合函数$y=f(\varphi(x))$在点x可导,且$y'=f'(\mu)\cdot\varphi'(x)$

(3) 隐函数导数$\frac{dy}{dx}$的求法一般有三种方法：

1)方程两边对x求导，要记住y是x的函数，则y的函数是x的复合函数.例如$\frac{1}{y}$，y^2，lny，e^y等均是x的复合函数. 对x求导应按复合函数连锁法则做。

2)公式法.由$F(x,y)=0$知 $\frac{dy}{dx}=-\frac{F'_x(x,y)}{F'_y(x,y)}$,其中，$F'_x(x,y)$，　$F'_y(x,y)$分别表示$F(x,y)$对x和y的偏导数。

3)利用微分形式不变性

8.常用高阶导数公式

（1）$(a^x)^{(n)}=a^x\ln^n a\quad(a>0)$　　$(e^x)^{(n)}=e^x$

（2）$(\sin kx)^{(n)}=k^n\sin(kx+n\cdot\frac{\pi}{2})$

（3）$(\cos kx)^{(n)}=k^n\cos(kx+n\cdot\frac{\pi}{2})$

（4）$(x^m)^{(n)}=m(m-1)\cdots(m-n+1)x^{m-n}$

（5）$(\ln x)^{(n)}=(-1)^{(n-1)}\frac{(n-1)!}{x^n}$

（6）莱布尼兹公式：若$u(x),v(x)$均n阶可导，则：　$(uv)^{(n)}=\sum_{i=0}^{n}c_n^i u^{(i)}v^{(n-i)}$，其中$u^{(0)}=u$，$v^{(0)}=v$

9.微分中值定理，泰勒公式

Th1:(费马定理)

若函数$f(x)$满足条件：

(1)函数$f(x)$在x_0的某邻域内有定义，并且在此邻域内恒有 $f(x) \le f(x_0)$或$f(x) \ge f(x_0)$,

(2) $f(x)$在x_0处可导,则有 $f'(x_0) = 0$

Th2:(罗尔定理)

设函数$f(x)$满足条件：

(1)在闭区间$[a,b]$上连续； (2)在(a,b)内可导；(3)$f(\mathrm{a}) = f(b)$

则在(a,b)内∃一个ξ，使 $f'(\xi) = 0$

Th3:(拉格朗日中值定理)

设函数$f(x)$满足条件：

(1)在$[a,b]$上连续；(2)在(a,b)内可导；

则在(a,b)内存在一个ξ，使 $\frac{f(b)-f(a)}{b-a} = f'(\xi)$

Th4:(柯西中值定理)

设函数$f(x)$，$g(x)$满足条件：

(1) 在$[a,b]$上连续；(2) 在(a,b)内可导且$f'(x)$，$g'(x)$均存在，且$g'(x) \ne 0$

则在(a,b)内存在一个ξ，使 $\frac{f(b)-f(a)}{g(b)-g(a)} = \frac{f\prime(\xi)}{g\prime(\xi)}$

10.洛必达法则

法则Ⅰ($\frac{0}{0}$型不定式极限)

设函数$f(x),g(x)$满足条件： $\lim_{x\to x_0} f(x) = 0, \lim_{x\to x_0} g(x) = 0$; $f(x),g(x)$在x_0的邻域内可导

(在x_0处可除外)且$g'(x) \ne 0$;

$\lim\limits_{x\to x_0}\frac{f'(x)}{g'(x)}$存在(或∞)。

则：$\lim\limits_{x\to x_0}\frac{f(x)}{g(x)}=\lim\limits_{x\to x_0}\frac{f'(x)}{g'(x)}$

法则I′（$\frac{0}{0}$型不定式极限）

设函数$f(x),g(x)$满足条件：$\lim\limits_{x\to\infty}f(x)=0,\lim\limits_{x\to\infty}g(x)=0$;存在一个$X>0$, 当$|x|>X$时, $f(x),g(x)$可导, 且$g'(x)\neq 0$; $\lim\limits_{x\to x_0}\frac{f'(x)}{g'(x)}$存在(或∞)。

则：$\lim\limits_{x\to x_0}\frac{f(x)}{g(x)}=\lim\limits_{x\to x_0}\frac{f'(x)}{g'(x)}$.

法则Ⅱ（$\frac{\infty}{\infty}$型不定式极限）

设函数$f(x),g(x)$满足条件：$\lim\limits_{x\to x_0}f(x)=\infty,\lim\limits_{x\to x_0}g(x)=\infty$; $f(x),g(x)$在x_0 的邻域内可导(在x_0处可除外)且$g'(x)\neq 0$; $\lim\limits_{x\to x_0}\frac{f'(x)}{g'(x)}$存在(或∞)。

则：$\lim\limits_{x\to x_0}\frac{f(x)}{g(x)}=\lim\limits_{x\to x_0}\frac{f'(x)}{g'(x)}$.

同理法则II′($\frac{\infty}{\infty}$型不定式极限)仿法则I′可写出

11. 泰勒公式

设函数$f(x)$在点x_0处的某邻域内具有$n+1$阶导数，则对该邻域内异于x_0的任意点x，在x_0与x之间至少存在一个ξ，使得：

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+\frac{1}{2!}f''(x_0)(x-x_0)^2+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+R_n(x)$$

其中 $R_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}$称为$f(x)$在点$x_0$处的$n$阶泰勒余项。

令$x_0=0$，则n阶泰勒公式：

$$f(x)=f(0)+f'(0)x+\frac{1}{2!}f''(0)x^2+\cdots+\frac{f^{(n)}(0)}{n!}x^n+R_n(x)\cdots\cdots$$

(1) 其中 $R_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}x^{n+1}$，$\xi$在 0 与$x$之间。(1)式称为麦克劳林公式

常用五种函数在$x_0=0$处的泰勒公式 ：

1) $\mathrm{e}^x=1+x+\frac{1}{2!}x^2+\cdots+\frac{1}{n!}x^n+\frac{x^{n+1}}{(n+1)!}e^{\xi}$

或 $=1+x+\frac{1}{2!}x^2+\cdots+\frac{1}{n!}x^n+o(x^n)$

2) $\sin x=x-\frac{1}{3!}x^3+\cdots+\frac{x^n}{n!}\sin\frac{n\pi}{2}+\frac{x^{n+1}}{(n+1)!}\sin\left(\xi+\frac{n+1}{2}\pi\right)$

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

3) $\cos x=1-\frac{1}{2!}x^2+\cdots+\frac{x^n}{n!}\cos\frac{n\pi}{2}+\frac{x^{n+1}}{(n+1)!}\cos(\xi+\frac{n+1}{2}\pi)$

或 $=1-\frac{1}{2!}x^2+\cdots+\frac{x^n}{n!}\cos\frac{n\pi}{2}+o(x^n)$

4) $\ln(1+x)=x-\frac{1}{2}x^2+\frac{1}{3}x^3-\cdots+(-1)^{n-1}\frac{x^n}{n}+\frac{(-1)^n x^{n+1}}{(n+1)(1+\xi)^{n+1}}$

或 $=x-\frac{1}{2}x^2+\frac{1}{3}x^3-\cdots+(-1)^{n-1}\frac{x^n}{n}+o(x^n)$

5) $(1+x)^m=1+mx+\frac{m(m-1)}{2!}x^2+\cdots+\frac{m(m-1)\cdots(m-n+1)}{n!}x^n$

$+\frac{m(m-1)\cdots(m-n+1)}{(n+1)!}x^{n+1}(1+\xi)^{m-n-1}$

或 $(1+x)^m=1+mx+\frac{m(m-1)}{2!}x^2+\cdots+\frac{m(m-1)\cdots(m-n+1)}{n!}x^n+o(x^n)$

12. 函数单调性的判断

Th1：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）。

Th2：（取极值的必要条件）设函数$f(x)$在x_0处可导，且在x_0处取极值，则$f'(x_0)=0$.

Th3：（取极值的第一充分条件）设函数$f(x)$在x_0的某一邻域内可微，且$f'(x_0)=0$（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$f(x_0)$为极小值；

(3)若$f'(x)$经过$x=x_0$的两侧不变号，则$f(x_0)$不是极值。

Th4：(取极值的第二充分条件)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．注：如果$f''(x_0)0$，此方法失效。

13.渐近线的求法

(1)水平渐近线

若$\lim\limits_{x\to+\infty}f(x)=b$，或$\lim\limits_{x\to-\infty}f(x)=b$，则$y=b$ 称为函数$y=f(x)$的水平渐近线。

(2)铅直渐近线

若$\lim\limits_{x\to x_0^-}f(x)=\infty$，或$\lim\limits_{x\to x_0^+}f(x)=\infty$，则$x=x_0$ 称为$y=f(x)$的铅直渐近线。

(3)斜渐近线 若$a=\lim\limits_{x\to\infty}\frac{f(x)}{x}$， $b=\lim\limits_{x\to\infty}[f(x)-ax]$，则 $y=ax+b$称为$y=f(x)$的斜渐近线。

14.函数凹凸性的判断

Th1：(凹凸性的判别定理）若在Ⅰ上$f''(x)<0$（或$f''(x)>0$）， 则$f(x)$在Ⅰ上是凸的（或凹的）。

Th2：(拐点的判别定理 1)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$(x_0,f(x_0))$为拐点。

Th3：(拐点的判别定理 2)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。

15.弧微分

$dS=\sqrt{1+y'^2}dx$

16.曲率

曲线$y=f(x)$在点(x,y)处的曲率$k=\frac{|y''|}{(1+y'^2)^{3/2}}$. 对于参数方程：

$$\begin{cases} x = \varphi(t) \\ y = \psi(t) \end{cases}, k = \frac{|\varphi'(t)\psi''(t) - \varphi''(t)\psi'(t)|}{[\varphi'^2(t) + \psi'^2(t)]^{3/2}}$$

17. 曲率半径

曲线在点M处的曲率$k(k \neq 0)$与曲线在点M处的曲率半径ρ有如下关系：$\rho = \frac{1}{k}$

线性代数

行列式

1.行列式按行（列）展开定理

(1) 设$A=(a_{ij})_{n\times n}$，则：$a_{i1}A_{j1}+a_{i2}A_{j2}+\cdots+a_{in}A_{jn}=\begin{cases}|A|,i=j\\0,i\neq j\end{cases}$

或$a_{1i}A_{1j}+a_{2i}A_{2j}+\cdots+a_{ni}A_{nj}=\begin{cases}|A|,i=j\\0,i\neq j\end{cases}$

即 $AA^*=A^*A=|A|E$,其中：$A^*=\begin{pmatrix}A_{11}&A_{12}&\dots&A_{1n}\\A_{21}&A_{22}&\dots&A_{2n}\\\dots&\dots&\dots&\dots\\A_{n1}&A_{n2}&\dots&A_{nn}\end{pmatrix}=(A_{ji})=(A_{ij})^T$

$$D_n=\begin{vmatrix}1&1&\dots&1\\x_1&x_2&\dots&x_n\\\dots&\dots&\dots&\dots\\x_1^{n-1}&x_2^{n-1}&\dots&x_n^{n-1}\end{vmatrix}==\prod_{1\le j<i\le n}(x_i-x_j)$$

(2) 设A,B为n阶方阵，则$|AB|=|A||B|=|B||A|=|BA|$，但$|A\pm B|=|A|\pm|B|$不一定成立。

(3) $|kA|=k^n|A|,A$为n阶方阵。

(4) 设A为n阶方阵，$|A^T|=|A|;|A^{-1}|=|A|^{-1}$（若$A$可逆），$|A^*|=|A|^{n-1}$

$n\geq 2$

(5) $\begin{vmatrix}A&O\\O&B\end{vmatrix}=\begin{vmatrix}A&C\\O&B\end{vmatrix}=\begin{vmatrix}A&O\\C&B\end{vmatrix}=|A||B|$，$A,B$为方阵，但

$\begin{vmatrix}O&A_{m\times m}\\B_{n\times n}&O\end{vmatrix}=(-1)^{mn}\cdot|A||B|$。

(6) 范德蒙行列式$D_n=\begin{vmatrix}1&1&\dots&1\\x_1&x_2&\dots&x_n\\\dots&\dots&\dots&\dots\\x_1^{n-1}&x_2^{n-1}&\dots&x_n^{n-1}\end{vmatrix}==\prod_{1\le j<i\le n}(x_i-x_j)$

设A是n阶方阵，$\lambda_i(i=1,2\cdots,n)$是A的n个特征值，则 $|A|=\prod_{i=1}^{n}\lambda_i$

矩阵

矩阵：$m \times n$个数a_{ij}排成m行n列的表格$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$称为矩阵，简记为$A$，

或者$(a_{ij})_{m\times n}$ 。若$m = n$，则称A是n阶矩阵或n阶方阵。

矩阵的线性运算

1. 矩阵的加法

设$A = (a_{ij}), B = (b_{ij})$是两个$m \times n$矩阵，则$m \times n$ 矩阵$C = (c_{ij}) = a_{ij} + b_{ij}$称为矩阵$A$与$B$的和，记为$A + B = C$ 。

2. 矩阵的数乘

设$A = (a_{ij})$是$m \times n$矩阵，k是一个常数，则$m \times n$矩阵(ka_{ij})称为数k与矩阵A的数乘，记为kA。

3. 矩阵的乘法

设$A = (a_{ij})$是$m \times n$矩阵，$B = (b_{ij})$是$n \times s$矩阵，那么$m \times s$矩阵$C = (c_{ij})$，其中 $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj}$称为$AB$的乘积，记为$C = AB$ 。

4. A^T、A^{-1}、A^*三者之间的关系

(1) $(A^T)^T = A, (AB)^T = B^T A^T, (kA)^T = kA^T, (A \pm B)^T = A^T \pm B^T$

(2) $(A^{-1})^{-1} = A, (AB)^{-1} = B^{-1}A^{-1}, (kA)^{-1} = \frac{1}{k}A^{-1},$

但 $(A \pm B)^{-1} = A^{-1} \pm B^{-1}$不一定成立。

(3) $(A^*)^* = |A|^{n-2}\ A \quad (n \geq 3)$，$(AB)^* = B^*A^*, \quad (kA)^* = k^{n-1}A^* \quad (n \geq 2)$

但$(A \pm B)^* = A^* \pm B^*$不一定成立。

(4) $(A^{-1})^T = (A^T)^{-1}, (A^{-1})^* = (AA^*)^{-1}, (A^*)^T = (A^T)^*$

5. 有关A^*的结论

(1) $AA^* = A^*A = |A|E$

(2) $|A^*| = |A|^{n-1}\ (n \geq 2), \quad (kA)^* = k^{n-1}A^*, \quad (A^*)^* = |A|^{n-2}A(n \geq 3)$

(3) 若A可逆，则$A^* = |A|A^{-1}, (A^*)^* = \frac{1}{|A|}A$

(4) 若A为n阶方阵，则：

$$r(A^*) = \begin{cases} n, & r(A) = n \\ 1, & r(A) = n-1 \\ 0, & r(A) < n-1 \end{cases}$$

6. 有关A^{-1}的结论

A可逆$\Leftrightarrow AB = E; \Leftrightarrow |A| \neq 0; \Leftrightarrow r(A) = n;$

$\Leftrightarrow A$可以表示为初等矩阵的乘积；$\Leftrightarrow A$无零特征值； $\Leftrightarrow \mathrm{Ax} = 0$ 只有零解。

7. 有关矩阵秩的结论

(1) 秩$r(A)$=行秩=列秩；

(2) $r(A_{m\times n}) \leq \min(m, n);$

(3) $A \neq 0 \Rightarrow r(A) \geq 1;$

(4) $r(A \pm B) \leq r(A) + r(B);$

(5) 初等变换不改变矩阵的秩

(6) $r(A) + r(B) - n \leq r(AB) \leq \min(r(A), r(B))$,特别若$AB = O$

则：$r(A) + r(B) \leq n$

(7) 若A^{-1}存在$\Rightarrow r(AB) = r(B)$; 若B^{-1}存在 $\Rightarrow r(AB) = r(A);$

若$r(A_{m\times n}) = n \Rightarrow r(AB) = r(B)$; 若$r(A_{m\times s}) = n \Rightarrow r(AB) = r(A)$。

(8) $r(A_{m\times s}) = n \Leftrightarrow Ax = 0$只有零解

8.分块求逆公式

$$\begin{pmatrix} A & O \\ O & B \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & O \\ O & B^{-1} \end{pmatrix};\qquad \begin{pmatrix} A & C \\ O & B \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & -A^{-1}CB^{-1} \\ O & B^{-1} \end{pmatrix};$$

$$\begin{pmatrix} A & O \\ C & B \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & O \\ -B^{-1}CA^{-1} & B^{-1} \end{pmatrix};\qquad \begin{pmatrix} O & A \\ B & O \end{pmatrix}^{-1} = \begin{pmatrix} O & B^{-1} \\ A^{-1} & O \end{pmatrix}$$

这里A，B均为可逆方阵。

向量

1.有关向量组的线性表示

(1) $\alpha_1,\alpha_2,\cdots,\alpha_s$线性相关⇔至少有一个向量可以用其余向量线性表示。

(2) $\alpha_1,\alpha_2,\cdots,\alpha_s$线性无关，$\alpha_1,\alpha_2,\cdots,\alpha_s$，$\beta$线性相关⇔$\beta$可以由$\alpha_1,\alpha_2,\cdots,\alpha_s$唯一线性表示。

(3) β可以由$\alpha_1,\alpha_2,\cdots,\alpha_s$线性表示 $\Leftrightarrow r(\alpha_1,\alpha_2,\cdots,\alpha_s)=r(\alpha_1,\alpha_2,\cdots,\alpha_s,\beta)$ 。

2.有关向量组的线性相关性

(1)部分相关，整体相关；整体无关，部分无关.

(2) ① n个n维向量 $\alpha_1,\alpha_2\cdots\alpha_n$线性无关⇔$|[\alpha_1\alpha_2\cdots\alpha_n]|\neq 0$， n个n维向量$\alpha_1,\alpha_2\cdots\alpha_n$线性相关 $\Leftrightarrow |[\alpha_1,\alpha_2,\cdots,\alpha_n]|=0$ 。

② $n+1$个n维向量线性相关。

③ 若$\alpha_1,\alpha_2\cdots\alpha_s$线性无关，则添加分量后仍线性无关；或一组向量线性相关，去掉某些分量后仍线性相关。

3.有关向量组的线性表示

(1) $\alpha_1,\alpha_2,\cdots,\alpha_s$线性相关⇔至少有一个向量可以用其余向量线性表示。

(2) $\alpha_1,\alpha_2,\cdots,\alpha_s$线性无关，$\alpha_1,\alpha_2,\cdots,\alpha_s$，$\beta$线性相关⇔$\beta$ 可以由$\alpha_1,\alpha_2,\cdots,\alpha_s$唯一线性表示。

(3) β可以由$\alpha_1,\alpha_2,\cdots,\alpha_s$线性表示 $\Leftrightarrow r(\alpha_1,\alpha_2,\cdots,\alpha_s)=r(\alpha_1,\alpha_2,\cdots,\alpha_s,\beta)$

4.向量组的秩与矩阵的秩之间的关系

设$r(A_{m\times n})=r$，则A的秩$r(A)$与A的行列向量组的线性相关性关系为：

(1) 若$r(A_{m\times n})=r=m$，则A的行向量组线性无关。

(2) 若$r(A_{m\times n})=r<m$，则A的行向量组线性相关。

(3) 若$r(A_{m\times n})=r=n$，则A的列向量组线性无关。

(4) 若$r(A_{m\times n})=r<n$，则A的列向量组线性相关。

5.n维向量空间的基变换公式及过渡矩阵

若$\alpha_1,\alpha_2,\cdots,\alpha_n$与$\beta_1,\beta_2,\cdots,\beta_n$是向量空间$V$的两组基，则基变换公式为：

$$(\beta_1,\beta_2,\cdots,\beta_n)=(\alpha_1,\alpha_2,\cdots,\alpha_n)\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}=(\alpha_1,\alpha_2,\cdots,\alpha_n)C$$

其中C是可逆矩阵，称为由基$\alpha_1,\alpha_2,\cdots,\alpha_n$到基$\beta_1,\beta_2,\cdots,\beta_n$的过渡矩阵。

6.坐标变换公式

若向量γ在基$\alpha_1,\alpha_2,\cdots,\alpha_n$与基$\beta_1,\beta_2,\cdots,\beta_n$的坐标分别是 $X=(x_1,x_2,\cdots,x_n)^T$，

$Y=(y_1,y_2,\cdots,y_n)^T$ 即： $\gamma=x_1\alpha_1+x_2\alpha_2+\cdots+x_n\alpha_n=y_1\beta_1+y_2\beta_2+\cdots+y_n\beta_n$，则向量坐标变换公式为$X=CY$ 或 $Y=C^{-1}X$ ，其中C是从基$\alpha_1,\alpha_2,\cdots,\alpha_n$到基$\beta_1,\beta_2,\cdots,\beta_n$的过渡矩阵。

7.向量的内积

$(\alpha,\beta)=a_1b_1+a_2b_2+\cdots+a_nb_n=\alpha^T\beta=\beta^T\alpha$

8.Schmidt 正交化

若$\alpha_1,\alpha_2,\cdots,\alpha_s$线性无关，则可构造$\beta_1,\beta_2,\cdots,\beta_s$使其两两正交，且$\beta_i$仅是$\alpha_1,\alpha_2,\cdots,\alpha_i$的线性组合$(i=1,2,\cdots,n)$，再把$\beta_i$单位化，记$\gamma_i=\frac{\beta_i}{|\beta_i|}$，则$\gamma_1,\gamma_2,\cdots,\gamma_i$是规范正交向量组。其中

$\beta_1=\alpha_1$， $\beta_2=\alpha_2-\frac{(\alpha_2,\beta_1)}{(\beta_1,\beta_1)}\beta_1$ ， $\beta_3=\alpha_3-\frac{(\alpha_3,\beta_1)}{(\beta_1,\beta_1)}\beta_1-\frac{(\alpha_3,\beta_2)}{(\beta_2,\beta_2)}\beta_2$ ，

. . . . . .

$$\beta_s = \alpha_s - \frac{(\alpha_s,\beta_1)}{(\beta_1,\beta_1)}\beta_1 - \frac{(\alpha_s,\beta_2)}{(\beta_2,\beta_2)}\beta_2 - \cdots - \frac{(\alpha_s,\beta_{s-1})}{(\beta_{s-1},\beta_{s-1})}\beta_{s-1}$$

9. 正交基及规范正交基

向量空间一组基中的向量如果两两正交，就称为正交基；若正交基中每个向量都是单位向量，就称其为规范正交基。

线性方程组

1．克莱姆法则

线性方程组$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases}$，如果系数行列式$D = |A| \neq 0$，则方程组有唯一解，$x_1 = \frac{D_1}{D}, x_2 = \frac{D_2}{D}, \cdots, x_n = \frac{D_n}{D}$，其中$D_j$是把$D$中第$j$列元素换成方程组右端的常数列所得的行列式。

2. n阶矩阵A可逆⇔ $Ax = 0$只有零解。⇔ $\forall b, Ax = b$总有唯一解，一般地，$r(A_{m\times n}) = n \Leftrightarrow Ax = 0$只有零解。

3. 非奇次线性方程组有解的充分必要条件，线性方程组解的性质和解的结构

(1) 设A为$m \times n$矩阵，若$r(A_{m\times n}) = m$，则对$Ax = b$而言必有$r(A) = r(A:b) = m$，从而$Ax = b$有解。

(2) 设$x_1, x_2, \cdots x_s$为$Ax = b$的解，则$k_1x_1 + k_2x_2 + \cdots + k_sx_s$当$k_1 + k_2 + \cdots + k_s = 1$时仍为$Ax = b$的解；但当$k_1 + k_2 + \cdots + k_s = 0$时，则为$Ax = 0$的解。特别$\frac{x_1+x_2}{2}$为$Ax = b$的解；$2x_3 - (x_1 + x_2)$为$Ax = 0$的解。

(3) 非齐次线性方程组$Ax = b$无解⇔ $r(A) + 1 = r(\overline{A}) \Leftrightarrow b$不能由$A$的列向量$\alpha_1, \alpha_2, \cdots, \alpha_n$线性表示。

4. 奇次线性方程组的基础解系和通解，解空间，非奇次线性方程组的通解

(1) 齐次方程组$Ax=0$恒有解(必有零解)。当有非零解时，由于解向量的任意线性组合仍是该齐次方程组的解向量，因此$Ax=0$的全体解向量构成一个向量空间，称为该方程组的解空间，解空间的维数是$n-r(A)$，解空间的一组基称为齐次方程组的基础解系。

(2) $\eta_1,\eta_2,\cdots,\eta_t$是$Ax=0$的基础解系，即：

1) $\eta_1,\eta_2,\cdots,\eta_t$是$Ax=0$的解；

2) $\eta_1,\eta_2,\cdots,\eta_t$线性无关；

3) $Ax=0$的任一解都可以由$\eta_1,\eta_2,\cdots,\eta_t$线性表出. $k_1\eta_1+k_2\eta_2+\cdots+k_t\eta_t$是$Ax=0$的通解，其中$k_1,k_2,\cdots,k_t$是任意常数。

矩阵的特征值和特征向量

1.矩阵的特征值和特征向量的概念及性质

(1) 设λ是A的一个特征值，则 $kA,aA+bE,A^2,A^m,f(A),A^T,A^{-1},A^*$有一个特征值分别为 $k\lambda,a\lambda+b,\lambda^2,\lambda^m,f(\lambda),\lambda,\lambda^{-1},\frac{|A|}{\lambda}$,且对应特征向量相同（$A^T$ 例外）。

(2) 若$\lambda_1,\lambda_2,\cdots,\lambda_n$为$A$的$n$个特征值，则$\sum_{i=1}^n\lambda_i=\sum_{i=1}^n a_{ii}$,$\prod_{i=1}^n\lambda_i=|A|$,从而$|A|\neq 0\Leftrightarrow A$没有特征值。

(3) 设$\lambda_1,\lambda_2,\cdots,\lambda_s$为$A$的$s$个特征值，对应特征向量为 $\alpha_1,\alpha_2,\cdots,\alpha_s$，

若：$\alpha=k_1\alpha_1+k_2\alpha_2+\cdots+k_s\alpha_s$ ，

则：$A^n\alpha=k_1A^n\alpha_1+k_2A^n\alpha_2+\cdots+k_sA^n\alpha_s=k_1\lambda_1^n\alpha_1+k_2\lambda_2^n\alpha_2+\cdots k_s\lambda_s^n\alpha_s$ 。

2.相似变换、相似矩阵的概念及性质

(1) 若$A\sim B$，则

1) $A^T\sim B^T,A^{-1}\sim B^{-1},,A^*\sim B^*$

2) $|A|=|B|,\sum_{i=1}^n A_{ii}=\sum_{i=1}^n b_{ii},r(A)=r(B)$

3) $|\lambda E-A|=|\lambda E-B|$，对$\forall\lambda$成立

3.矩阵可相似对角化的充分必要条件

(1) 设A为n阶方阵，则A可对角化⇔对每个k_i重根特征值λ_i，有$n-r(\lambda_i E-A)=k_i$

(2) 设A可对角化，则由$P^{-1}AP=\Lambda$,有$A=P\Lambda P^{-1}$，从而$A^n=P\Lambda^n P^{-1}$

(3) 重要结论

1) 若$A\sim B,C\sim D$，则$\begin{bmatrix} A & O \\ O & C \end{bmatrix}\sim\begin{bmatrix} B & O \\ O & D \end{bmatrix}$.

2) 若$A\sim B$，则$f(A)\sim f(B),|f(A)|\sim|f(B)|$，其中$f(A)$为关于$n$阶方阵$A$的多项式。

3) 若A为可对角化矩阵，则其非零特征值的个数(重根重复计算)＝秩(A)

4.实对称矩阵的特征值、特征向量及相似对角阵

(1)相似矩阵：设A,B为两个n阶方阵，如果存在一个可逆矩阵P，使得$B=P^{-1}AP$成立，则称矩阵A与B相似，记为$A\sim B$。

(2)相似矩阵的性质：如果$A\sim B$则有：

1) $A^T\sim B^T$

2) $A^{-1}\sim B^{-1}$ （若A，B均可逆）

3) $A^k\sim B^k$ （k为正整数）

4) $|\lambda E-A|=|\lambda E-B|$，从而$A,B$ 有相同的特征值

5) $|A|=|B|$，从而A,B同时可逆或者不可逆

6) 秩(A) =秩$(B),|\lambda E-A|=|\lambda E-B|$，$A,B$不一定相似

二次型

1.n个变量$x_1,x_2,\cdots,x_n$的二次齐次函数

$f(x_1,x_2,\cdots,x_n)=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_iy_j$，其中$a_{ij}=a_{ji}(i,j=1,2,\cdots,n)$，称为$n$元二次型，简称二次型．若令$x=\begin{bmatrix}x_1\\x_1\\\vdots\\x_n\end{bmatrix},A=\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1n}\\a_{21}&a_{22}&\cdots&a_{2n}\\\cdots&\cdots&\cdots&\cdots\\a_{n1}&a_{n2}&\cdots&a_{nn}\end{bmatrix}$,这二次型$f$可改写成矩阵向量形式$f=x^TAx$。其中$A$称为二次型矩阵，因为$a_{ij}=a_{ji}(i,j=1,2,\cdots,n)$，所以二次型矩阵均为对称矩阵，且二次型与对称矩阵一一对应，并把矩阵A的秩称为二次型的秩。

2.惯性定理，二次型的标准形和规范形

(1) 惯性定理

对于任一二次型，不论选取怎样的合同变换使它化为仅含平方项的标准型，其正负惯性指数与所选变换无关，这就是所谓的惯性定理。

(2) 标准形

二次型$f=(x_1,x_2,\cdots,x_n)=x^TAx$经过合同变换$x=Cy$化为$f=x^TAx=y^TC^TAC$

$y=\sum_{i=1}^{r}d_iy_i^2$称为 $f(r\le n)$的标准形。在一般的数域内，二次型的标准形不是唯一的，与所作的合同变换有关，但系数不为零的平方项的个数由$r(A$的秩$)$唯一确定。

(3) 规范形

任一实二次型f都可经过合同变换化为规范形$f=z_1^2+z_2^2+\cdots+z_p^2-z_{p+1}^2-\cdots-z_r^2$，其中$r$为$A$的秩，$p$为正惯性指数，$r-p$为负惯性指数，且规范型唯一。

3.用正交变换和配方法化二次型为标准形，二次型及其矩阵的正定性

设A正定$\Rightarrow kA(k>0),A^T,A^{-1},A^*$正定；$|A|>0,A$可逆；$a_{ii}>0$，且$|A_{ii}|>0$

A，B正定$\Rightarrow A+B$正定，但AB，BA不一定正定

A正定$\Leftrightarrow f(x)=x^TAx>0,\forall x\neq 0$

$\Leftrightarrow A$的各阶顺序主子式全大于零

$\Leftrightarrow A$的所有特征值大于零

$\Leftrightarrow A$的正惯性指数为n

⇔存在可逆阵P使$A = P^T P$

⇔存在正交矩阵Q，使$Q^T AQ = Q^{-1} AQ = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$,

其中$\lambda_i > 0, i = 1,2,\cdots,n$.正定⇒ $kA(k > 0), A^T, A^{-1}, A^*$正定； $|A| > 0, A$可逆；$a_{ii} > 0$，且$|A_{ii}| > 0$ 。

概率论和数理统计

随机事件和概率

1.事件的关系与运算

(1) 子事件：$A \subset B$，若A发生，则B发生。

(2) 相等事件：$A = B$，即$A \subset B$，且$B \subset A$ 。

(3) 和事件：$A \cup B$（或$A + B$），A与B中至少有一个发生。

(4) 差事件：$A - B$，A发生但B不发生。

(5) 积事件：$A \cap B$（或AB），A与B同时发生。

(6) 互斥事件（互不相容）：$A \cap B = \varnothing$。

(7) 互逆事件（对立事件）： $A \cap B = \varnothing, A \cup B = \Omega, A = \overline{B}, B = \overline{A}$ 。

2.运算律

(1) 交换律：$A \cup B = B \cup A, A \cap B = B \cap A$

(2) 结合律：$(A \cup B) \cup C = A \cup (B \cup C)$； $(A \cap B) \cap C = A \cap (B \cap C)$

(3) 分配律：$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

3.德.摩根律

$\overline{A \cup B} = \overline{A} \cap \overline{B}$　　$\overline{A \cap B} = \overline{A} \cup \overline{B}$

4.完全事件组

$A_1 A_2 \cdots A_n$两两互斥，且和事件为必然事件，即$A_i \cap A_j = \varnothing, i \neq j, \bigcup_{i=1}^{n} = \Omega$

5.概率的基本概念

(1) 概率：事件发生的可能性大小的度量，其严格定义如下：

概率$P(g)$为定义在事件集合上的满足下面 3 个条件的函数：

1)对任何事件A，$P(A) \geq 0$

2)对必然事件Ω，$P(\Omega) = 1$

3)对$A_1A_2 \cdots A_n, \cdots$，若$A_iA_j = \varnothing (i \neq j)$，则：$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A)$.

(2) 概率的基本性质

1) $P(\overline{A}) = 1 - P(A)$;

2) $P(A - B) = P(A) - P(AB)$;

3) $P(A \cup B) = P(A) + P(B) - P(AB)$ 特别，当$B \subset A$时，$P(A - B) = P(A) - P(B)$且$P(B) \leq P(A)$； $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(BC) - P(AC) + P(ABC)$

4) 若$A_1, A_2, \cdots, A_n$两两互斥，则$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} (P(A_i)$

(3) 古典型概率：实验的所有结果只有有限个， 且每个结果发生的可能性相同，其概率计算公式： $P(A) = \frac{\text{事件}A\text{发生的基本事件数}}{\text{基本事件总数}}$

(4) 几何型概率：样本空间Ω为欧氏空间中的一个区域， 且每个样本点的出现具有等可能性，其概率计算公式：$P(A) = \frac{A\text{的度量(长度、面积、体积)}}{\Omega\text{的度量(长度、面积、体积)}}$

6.概率的基本公式

(1) 条件概率： $P(B|A) = \frac{P(AB)}{P(A)}$,表示A发生的条件下，B发生的概率

(2) 全概率公式： $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i), B_iB_j = \varnothing, i \neq j, \bigcup_{i=1}^{n} B_i = \Omega.$

(3) Bayes 公式：

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}, j = 1,2, \cdots, n$$

注：上述公式中事件B_i的个数可为可列个.

(4)乘法公式： $P(A_1A_2)=P(A_1)P(A_2|A_1)=P(A_2)P(A_1|A_2)$ $P(A_1A_2\cdots A_n)=P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1})$

7.事件的独立性

(1) A 与 B 相互独立⇔ $P(AB)=P(A)P(B)$

(2) A，B，C 两两独立 ⇔ $P(AB)=P(A)P(B);P(BC)=P(B)P(C);\ P(AC)=P(A)P(C);$

(3) A，B，C 相互独立 ⇔ $P(AB)=P(A)P(B);\ P(BC)=P(B)P(C);\ P(AC)=P(A)P(C);$ $P(ABC)=P(A)P(B)P(C).$

8.独立重复试验

将某试验独立重复 n 次，若每次实验中事件 A 发生的概率为 p，则 n 次试验中 A 发生 k 次的概率为： $P(X=k)=C_n^kp^k(1-p)^{n-k}$ 。

9.重要公式与结论

(1) $P(\overline{A})=1-P(A)$

(2) $P(A\cup B)=P(A)+P(B)-P(AB)$

$P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(AB)-P(BC)-P(AC)+P(ABC)$

(3) $P(A-B)=P(A)-P(AB)$

(4) $P(A\overline{B})=P(A)-P(AB),P(A)=P(AB)+P(A\overline{B}),\quad P(A\cup B)=P(A)+P(\overline{A}B)=P(AB)+P(A\overline{B})+P(\overline{A}B)$

(5) 条件概率$P(\cdot|B)$满足概率的所有性质，

例如：. $P(\overline{A}_1|B)=1-P(A_1|B)\quad P(A_1\cup A_2|B)=P(A_1|B)+P(A_2|B)-P(A_1A_2|B)$ $P(A_1A_2|B)=P(A_1|B)P(A_2|A_1B)$

(6) 若$A_1,A_2,\cdots,A_n$相互独立，则$P(\cap_{i=1}^nA_i)=\prod_{i=1}^nP(A_i),\quad P(\cup_{i=1}^nA_i)=\prod_{i=1}^n(1-P(A_i))$

(7) 互斥、互逆与独立性之间的关系： A 与 B 互逆⇒A 与 B 互斥，但反之不成立，A 与 B 互 斥（或互逆）且均非零概率事件⇒A 与 B 不独立.

(8) 若$A_1,A_2,\cdots,A_m,B_1,B_2,\cdots,B_n$相互独立，则$f(A_1,A_2,\cdots,A_m)$与 $g(B_1,B_2,\cdots,B_n)$也相互独立，其中$f(\cdot),g(\cdot)$分别表示对相应事件做任意事件运算后所得的事件，另外，概率为 1（或 0）的事件与任何事件相互独立.

随机变量及其概率分布

1.随机变量及概率分布

取值带有随机性的变量，严格地说是定义在样本空间上，取值于实数的函数称为随机变量，概率分布通常指分布函数或分布律

2.分布函数的概念与性质

定义： $F(x)=P(X\le x),-\infty<x<+\infty$

性质：(1)$0\le F(x)\le 1$　　(2)$F(x)$单调不减

(3)右连续$F(x+0)=F(x)$　　(4)$F(-\infty)=0,F(+\infty)=1$

3.离散型随机变量的概率分布

$P(X=x_i)=p_i,i=1,2,\cdots,n,\cdots$　　$p_i\ge 0,\sum_{i=1}^{\infty}p_i=1$

4.连续型随机变量的概率密度

概率密度$f(x)$;非负可积，且:(1)$f(x)\ge 0$, (2)$\int_{-\infty}^{+\infty}f(x)dx=1$ (3)x为$f(x)$的连续点，则:

$f(x)=F'(x)$分布函数$F(x)=\int_{-\infty}^{x}f(t)dt$

5.常见分布

(1) 0-1 分布:$P(X=k)=p^k(1-p)^{1-k},k=0,1$

(2) 二项分布:$B(n,p)$:　$P(X=k)=C_n^kp^k(1-p)^{n-k},k=0,1,\cdots,n$

(3) Poisson 分布:$p(\lambda)$:　$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},\lambda>0,k=0,1,2\cdots$

(4) 均匀分布$U(a,b)$：$f(x)=\begin{cases}\frac{1}{b-a},a<x<b\\0,\end{cases}$

(5) 正态分布:$N(\mu,\sigma^2)$: $\varphi(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\sigma>0,-\infty<x<+\infty$

(6)指数分布:$E(\lambda):f(x)=\begin{cases}\lambda e^{-\lambda x},x>0,\lambda>0\\0,\end{cases}$

(7)几何分布:$G(p):P(X=k)=(1-p)^{k-1}p,0<p<1,k=1,2,\cdots.$

(8)超几何分布: $H(N,M,n):P(X=k)=\frac{C_M^kC_{N-M}^{n-k}}{C_N^n},k=0,1,\cdots,\min(n,M)$

6. 随机变量函数的概率分布

(1)离散型：$P(X=x_1)=p_i,Y=g(X)$

则: $P(Y=y_j)=\sum_{g(x_i)=y_i}P(X=x_i)$

(2)连续型：$X\tilde{}\ f_X(x),Y=g(x)$

则:$F_y(y)=P(Y\le y)=P(g(X)\le y)=\int_{g(x)\le y}f_x(x)dx,\quad f_Y(y)=F'_Y(y)$

7. 重要公式与结论

(1) $X\sim N(0,1)\Rightarrow\varphi(0)=\frac{1}{\sqrt{2\pi}},\Phi(0)=\frac{1}{2},\ \Phi(-a)=P(X\le-a)=1-\Phi(a)$

(2) $X\sim N(\mu,\sigma^2)\Rightarrow\frac{X-\mu}{\sigma}\sim N(0,1),P(X\le a)=\Phi(\frac{a-\mu}{\sigma})$

(3) $X\sim E(\lambda)\Rightarrow P(X>s+t|X>s)=P(X>t)$

(4) $X\sim G(p)\Rightarrow P(X=m+k|X>m)=P(X=k)$

(5) 离散型随机变量的分布函数为阶梯间断函数；连续型随机变量的分布函数为连续函数，但不一定为处处可导函数。

(6) 存在既非离散也非连续型随机变量。

多维随机变量及其分布

1. 二维随机变量及其联合分布

由两个随机变量构成的随机向量(X,Y)， 联合分布为$F(x,y)=P(X\le x,Y\le y)$

2. 二维离散型随机变量的分布

(1) 联合概率分布律 $P\{X=x_i,Y=y_j\}=p_{ij};i,j=1,2,\cdots$

(2) 边缘分布律 $p_{i\cdot}=\sum_{j=1}^{\infty}p_{ij},i=1,2,\cdots\ p_{\cdot j}=\sum_{i}^{\infty}p_{ij},j=1,2,\cdots$

(3) 条件分布律 $P\{X=x_i|Y=y_j\}=\frac{p_{ij}}{p_{\cdot j}}$ $P\{Y=y_j|X=x_i\}=\frac{p_{ij}}{p_{i\cdot}}$

3. 二维连续性随机变量的密度

(1) 联合概率密度$f(x,y)$:

1) $f(x,y)\geq 0$ 2) $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f(x,y)dxdy=1$

(2) 分布函数：$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)dudv$

(3) 边缘概率密度： $f_X(x)=\int_{-\infty}^{+\infty}f(x,y)dy$ $f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)dx$

(4) 条件概率密度：$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}$ $f_{Y|X}(y|x)=\frac{f(x,y)}{f_X(x)}$

4. 常见二维随机变量的联合分布

(1) 二维均匀分布：$(x,y)\sim U(D)$ ，$f(x,y)=\begin{cases}\frac{1}{S(D)},(x,y)\in D\\ 0,\text{ 其他}\end{cases}$

(2) 二维正态分布：$(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}.\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

5. 随机变量的独立性和相关性

X和Y的相互独立:$\Leftrightarrow F(x,y)=F_X(x)F_Y(y)$:

$\Leftrightarrow p_{ij}=p_{i\cdot}\cdot p_{\cdot j}$（离散型） $\Leftrightarrow f(x,y)=f_X(x)f_Y(y)$（连续型）

X和Y的相关性：

相关系数$\rho_{XY}=0$时，称X和Y不相关， 否则称X和Y相关

6.两个随机变量简单函数的概率分布

离散型： $P(X=x_i,Y=y_i)=p_{ij},Z=g(X,Y)$ 则：

$P(Z=z_k)=P\{g(X,Y)=z_k\}=\sum_{g(x_i,y_i)=z_k}P\left(X=x_i,Y=y_j\right)$

连续型： $(X,Y)\sim f(x,y),Z=g(X,Y)$ 则：

$F_z(z)=P\{g(X,Y)\le z\}=\iint_{g(x,y)\le z}f(x,y)dxdy$， $f_z(z)=F'_z(z)$

7.重要公式与结论

(1) 边缘密度公式： $f_X(x)=\int_{-\infty}^{+\infty}f(x,y)dy$, $f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)dx$

(2) $P\{(X,Y)\in D\}=\iint_D f(x,y)dxdy$

(3) 若(X,Y)服从二维正态分布$N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$ 则有：

1) $X\sim N(\mu_1,\sigma_1^2),Y\sim N(\mu_2,\sigma_2^2)$.

2) X与Y相互独立$\Leftrightarrow\rho=0$，即X与Y不相关。

3) $C_1X+C_2Y\sim N(C_1\mu_1+C_2\mu_2,C_1^2\sigma_1^2+C_2^2\sigma_2^2+2C_1C_2\sigma_1\sigma_2\rho)$

4) X关于 Y=y 的条件分布为： $N(\mu_1+\rho\frac{\sigma_1}{\sigma_2}(y-\mu_2),\sigma_1^2(1-\rho^2))$

5) Y关于$X=x$的条件分布为： $N(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),\sigma_2^2(1-\rho^2))$

(4) 若X与Y独立，且分别服从$N(\mu_1,\sigma_1^2),N(\mu_1,\sigma_2^2)$, 则：

$(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,0)$, $C_1X+C_2Y\tilde{}\ N(C_1\mu_1+C_2\mu_2,C_1^2\sigma_1^2+C_2^2\sigma_2^2)$.

(5) 若X与Y相互独立，$f(x)$和$g(x)$为连续函数， 则$f(X)$和$g(Y)$也相互独立。

随机变量的数字特征

1.数学期望

离散型：$P\{X=x_i\}=p_i,E(X)=\sum_i x_ip_i$；

连续型： $X\sim f(x), E(X)=\int_{-\infty}^{+\infty} xf(x)dx$

性质：

(1) $E(C)=C, E[E(X)]=E(X)$

(2) $E(C_1X+C_2Y)=C_1E(X)+C_2E(Y)$

(3) 若 X 和 Y 独立，则$E(XY)=E(X)E(Y)$ (4)$[E(XY)]^2\le E(X^2)E(Y^2)$

2. 方差：$D(X)=E[X-E(X)]^2=E(X^2)-[E(X)]^2$

3. 标准差：$\sqrt{D(X)}$，

4. 离散型：$D(X)=\sum_i[x_i-E(X)]^2p_i$

5. 连续型：$D(X)=\int_{-\infty}^{+\infty}[x-E(X)]^2 f(x)dx$

性质：

(1) $D(C)=0, D[E(X)]=0, D[D(X)]=0$

(2) X与Y相互独立，则$D(X\pm Y)=D(X)+D(Y)$

(3) $D(C_1X+C_2)=C_1^2D(X)$

(4) 一般有 $D(X\pm Y)=D(X)+D(Y)\pm 2Cov(X,Y)=D(X)+D(Y)\pm 2\rho\sqrt{D(X)}\sqrt{D(Y)}$

(5) $D(X)<E(X-C)^2, C\ne E(X)$

(6) $D(X)=0\Leftrightarrow P\{X=C\}=1$

6. 随机变量函数的数学期望

(1) 对于函数$Y=g(x)$

X为离散型：$P\{X=x_i\}=p_i, E(Y)=\sum_i g(x_i)p_i$；

X为连续型：$X\sim f(x), E(Y)=\int_{-\infty}^{+\infty} g(x)f(x)dx$

(2) $Z=g(X,Y);(X,Y)\sim P\{X=x_i,Y=y_j\}=p_{ij}$； $E(Z)=\sum_i\sum_j g(x_i,y_j)p_{ij}$

$(X,Y)\sim f(x,y);E(Z)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}g(x,y)f(x,y)dxdy$

7. 协方差 $Cov(X,Y)=E[(X-E(X)(Y-E(Y))]$

8. 相关系数 $\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$, k阶原点矩 $E(X^k)$； k阶中心矩 $E\{[X-E(X)]^k\}$

性质：

(1) $Cov(X,Y)=Cov(Y,X)$

(2) $Cov(aX,bY)=abCov(Y,X)$

(3) $Cov(X_1+X_2,Y)=Cov(X_1,Y)+Cov(X_2,Y)$

(4) $|\rho(X,Y)|\le 1$

(5) $\rho(X,Y)=1\Leftrightarrow P(Y=aX+b)=1$ ，其中$a>0$

$\rho(X,Y)=-1\Leftrightarrow P(Y=aX+b)=1$ ，其中$a<0$

9. 重要公式与结论

(1) $D(X)=E(X^2)-E^2(X)$

(2) $Cov(X,Y)=E(XY)-E(X)E(Y)$

(3) $|\rho(X,Y)|\le 1$,且 $\rho(X,Y)=1\Leftrightarrow P(Y=aX+b)=1$，其中$a>0$

$\rho(X,Y)=-1\Leftrightarrow P(Y=aX+b)=1$，其中$a<0$

(4) 下面 5 个条件互为充要条件：

$\rho(X,Y)=0\Leftrightarrow Cov(X,Y)=0\Leftrightarrow E(X,Y)=E(X)E(Y)\Leftrightarrow D(X+Y)=D(X)+D(Y)\Leftrightarrow D(X-Y)=D(X)+D(Y)$

注：X与Y独立为上述 5 个条件中任何一个成立的充分条件，但非必要条件。

数理统计的基本概念

1. 基本概念

总体：研究对象的全体，它是一个随机变量，用X表示。

个体：组成总体的每个基本元素。

简单随机样本：来自总体X的n个相互独立且与总体同分布的随机变量$X_1, X_2 \cdots, X_n$，称为容量为n的简单随机样本，简称样本。

统计量：设$X_1, X_2 \cdots, X_n$,是来自总体X的一个样本，$g(X_1, X_2 \cdots, X_n)$）是样本的连续函数，且$g(\cdot)$中不含任何未知参数，则称$g(X_1, X_2 \cdots, X_n)$为统计量

样本均值：$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

样本方差：$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

样本矩：样本k阶原点矩：$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, k = 1,2, \cdots$

样本k阶中心矩：$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k, k = 1,2, \cdots$

2. 分布

χ^2分布：$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi^2(n)$，其中$X_1, X_2 \cdots, X_n$,相互独立，且同服从$N(0,1)$

t分布：$T = \frac{X}{\sqrt{Y/n}} \sim t(n)$ ，其中$X \sim N(0,1), Y \sim \chi^2(n)$,且$X$，$Y$ 相互独立。

F 分布：$F = \frac{X/n_1}{Y/n_2} \sim F(n_1, n_2)$，其中$X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$,且$X$，$Y$相互独立。

分位数：若$P(X \leq x_\alpha) = \alpha$,则称$x_\alpha$为$X$的$\alpha$分位数

3. 正态总体的常用样本分布

(1) 设$X_1, X_2 \cdots, X_n$为来自正态总体$N(\mu, \sigma^2)$的样本，

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$,则：

1) $\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$ 或者$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\sim N(0,1)$

2) $\frac{(n-1)S^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\sim\chi^2(n-1)$

3) $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\sim\chi^2(n)$

4) $\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$

4. 重要公式与结论

(1) 对于$\chi^2\sim\chi^2(n)$，有$E(\chi^2(n))=n,D(\chi^2(n))=2n;$

(2) 对于$T\sim t(n)$，有$E(T)=0,D(T)=\frac{n}{n-2}(n>2);$

(3) 对于$F^{\sim}F(m,n)$，有 $\frac{1}{F}\sim F(n,m),F_{a/2}(m,n)=\frac{1}{F_{1-a/2}(n,m)};$

(4) 对于任意总体X，有 $E(\overline{X})=E(X),E(S^2)=D(X),D(\overline{X})=\frac{D(X)}{n}$