> 本文是斯坦福大学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 备注：请关注github的更新，线性代数已经更新完毕，近期将翻译概率论。

CS 229 机器学习课程复习材料

一、线性代数复习和参考

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^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。

$$xy^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}$$

举一个外积如何使用的一个例子：让$1 \in R^n$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：，

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

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：

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

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:

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

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：

- 矩阵乘法交换律: $(AB)C = A(BC)$
- 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。

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

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：$\|x\|_2^2 = x^T x$

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

1. 对于所有的 $x \in \mathbb{R}^n$, $f(x) \geq 0$(非负).
2. 当且仅当$x = 0$ 时，$f(x) = 0$ (明确性).
3. 对于所有$x \in \mathbb{R}^n$,$t \in \mathbb{R}$，则 $f(tx) = |t| f(x)$ (正齐次性).
4. 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:

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

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$\text{rank}(A)$表示。 以下是秩的一些基本属性：

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

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:

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

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$\{x_1, \ldots x_n\}$是可以表示为$\{x_1, \ldots x_n\}$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:

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

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：

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

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$\text{span}(\{a_1, \dots a_n\})$受到系数$a_1, \dots a_n$的限制的线性组合，$\alpha_1, \cdots, \alpha_n$满足$0 \leq \alpha_i \leq 1, i = 1, \dots, n$。从形式上看，

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

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。

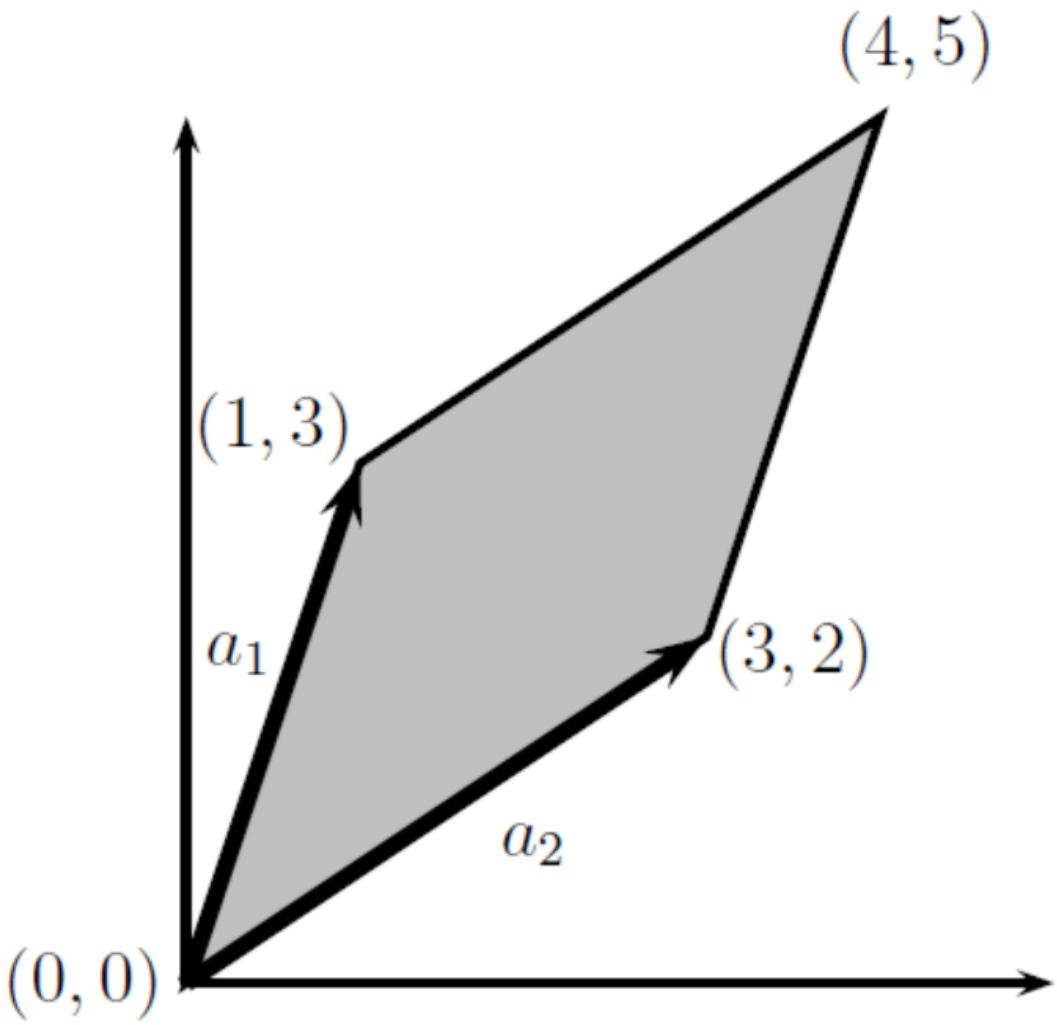

图1：（4）中给出的2×2矩阵A的行列式的图示。 这里，a_1和a_2是对应于A行的向量，并且集合S对应于阴影区域（即，平行四边形）。 这个行列式的绝对值，$|\det A| = 7$，即平行四边形的面积。

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$t\,|A|$

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

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。

3. 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：

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

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

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

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：

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

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

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

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：

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

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$x \in \mathbb{R}^n$，$x^T Ax > 0$，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$x^T Ax \geq 0$，对称矩阵$A \in \mathbb{S}^n$是**半正定**(**positive semidefinite ,PSD**)。 这写为（或$A \succeq 0$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$x^T Ax < 0$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$x^T Ax \leq 0$表示为$A \preceq 0$（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$x_1^T Ax_1 > 0$且$x_2^T Ax_2 < 0$。

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，$x^T Ax = 0$，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$G = A^T A$（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

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

- 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：

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

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$x_1, \cdots x_n$：

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

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：

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

“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$\lambda_i > 0$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：

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

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：

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

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$t \in \mathbb{R}$，$\nabla_x (t f(x)) = t \nabla_x f(x)$

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$f(z) = z^T z$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。

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），因为关于矩阵的黑森方程就必须对矩阵所有元素求偏导数$\partial^2 f(A) / (\partial A_{ij} \partial A_{k\ell})$，将其表示为矩阵相当麻烦。

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：

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

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$\left(\nabla_x f(x)\right)_i = \partial f(x)/\partial x_i$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：

$$\nabla_x^2 f(x) = \begin{bmatrix} \nabla_x (\nabla_x f(x))_1 & \nabla_x (\nabla_x f(x))_2 & \cdots & \nabla_x (\nabla_x f(x))_n \end{bmatrix}$$

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：

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

根据x的梯度，并利用上一节中推导的性质：

$$\begin{aligned}\nabla_x\left(x^TA^TAx - 2b^TAx + b^Tb\right) &= \nabla_x x^TA^TAx - \nabla_x 2b^TAx + \nabla_x b^Tb\\ &= 2A^TAx - 2A^Tb\end{aligned}$$

将最后一个表达式设置为零，然后解出x，得到了正规方程：

$$x = (A^TA)^{-1}A^Tb$$

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：

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

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：

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

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，

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

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）$x^T Ax$的唯一点是A的特征向量。

线性代数已经翻译完毕，后面的概率论部分还在翻译中，请关注github的更新，近期将更新完。

欢迎大家提交PR，对语言进行润色。

翻译：黄海广