Mathematics for Computer Science Researchers

Hai-Liang Zhao*
hliangzhao97@gmail.com

2018 年 9 月 30 日

目录

1 About **3**

2 What is Linear Algebra? **3**

2.1 Vector spaces . . . 3
2.1.1 Euclidean space . . . 5
2.1.2 Subspaces . . . 6
2.2 Linear maps . . . 6
2.2.1 What is matrix? . . . 7
2.2.2 The matrix of a linear map . . . 7
2.2.3 Nullspace, range . . . 9
2.3 Metric spaces . . . 10
2.4 Normed spaces . . . 10
2.5 More details about matrices . . . 11
2.5.1 Basic properties of matrices . . . 11
2.5.2 Deternimant . . . 12
2.5.3 Inverse of matrices . . . 14
2.5.4 Useful matrix identities . . . 14
2.6 Inner product spaces . . . 15
2.6.1 Pythagorean Theorem . . . 15
2.6.2 Cauchy-Schwarz inequality . . . 15
2.6.3 Orthogonal complements and projections . . . 15
2.7 Solving the problem: $\mathbf{Ax} = \mathbf{b}$ . . . 15
2.7.1 Cramer's Rule . . . 15
2.7.2 Elementary transformation of matrices . . . 15
2.7.3 Rank of matrices . . . 15
2.7.4 Solution of the problem . . . 16
2.8 In-depth understanding of Linear Dependence . . . 16
2.8.1 Rank of vector groups . . . 16
2.8.2 Structure of the solution of $\mathbf{Ax} = \mathbf{b}$ . . . 16

*Hai-Liang Zhao is a pre-graduate student in Zhejiang University. Currently he is a fourth year undergraduate in School of Computer Science and Technology, WUT.

2.9 Eigenthings . . . 16
2.10 Trace . . . 16
2.11 Orthogonal matrices . . . 16
2.12 Symmetric matrices . . . 16
2.13 Singular value decomposition . . . 16
2.14 Fundamental Theorem of Linear Algebra . . . 16
2.15 Low-rank Approximation . . . 16
2.16 Pseudoinverse . . . 16

3 Probability Theory **16**
3.1 Basics . . . 16
3.2 Random variables . . . 16
3.3 Joint distributions . . . 16
3.4 Great Expectations . . . 16
3.5 Variance . . . 16
3.6 Covariance . . . 16
3.7 Random Vectors . . . 16
3.8 Estimation of Parameters . . . 16

4 Calculus **16**
4.1 Extrema . . . 17
4.2 Gradients . . . 17
4.3 The Jacobian . . . 17
4.4 The Hessian . . . 17
4.5 Matrix calculus . . . 17
4.6 Taylor's theorm . . . 17
4.7 Conditions for local minima . . . 17

5 Optimization **17**
5.1 Convexity . . . 17
5.1.1 Convex sets . . . 17
5.1.2 Convex functions . . . 17
5.2 Convex Optimization problems . . . 17
5.2.1 Linear optimization problems . . . 17
5.2.2 Quadratic optimization problems . . . 17
5.2.3 Geometric programming . . . 17
5.3 Duality . . . 17
5.3.1 The Lagrange dual function . . . 17
5.3.2 Optimality conditions . . . 17
5.3.3 Perturbation and sensitivity analysis . . . 17
5.4 Unconstrained minimization . . . 17
5.5 Equality constrained minimization . . . 17
5.6 Interior-point methods . . . 17

6 Optimization in Machine Learning **17**
6.1 Batch gradient methods 18
6.2 Stochastic optimization methods 18
6.3 Second-order methods 18

1 About

很早之前就在脑海中萌生了写作此文档的意图.

此刻距离我第一次接触机器学习中的种种算法背后的数学理论已经超过了一年, 但是在这些算法背后的数学推导部分却几乎没有长进. 追根溯源，原因就在于我没有从本质上理解Linear Algebra, Calculus and Optimization以及Probability Theory 存在的意义. 如果我问'半正定矩阵有哪些性质?', '函数极值的判定条件是什么?', 应当难不倒大家. 但是如果我问'人们为什么要定义向量空间的概念?', '为什么我们偏爱半正定矩阵?'，或许可以难倒一大批人.

前段时间做了第一次尝试. 当时写作了一篇题为'What is Linear Algebra?'的文档, 虽然远远没完成, 至今也没有再续笔, 但是在此文中我就已经开始想解答Linear Algebra 诞生的原因了(当然, 可能我的理解是错误的). 如果仅仅去阅读工科教材《线性代数》、《高等数学》, 恐怕永远也不会明白上述这些问题.

当我开始阅读一些优化理论的英文教材时, 我明显地感受到了何谓捉襟见肘. 中文教材倾向于先把概念放出来, 然后再去解释其性质、运用等等. 至于为什么这个概念被引入对我们却完全透明. 我们可以明白从A到B的推导过程, 却不知道A为何被提出. 这篇文档就试图去解决这个问题.

这篇文档算是我对自己将要研究的方向所需要使用的数学基础作一总结.

此处的基础, 不仅仅是指这些理论本身, 还指研究这些问题所需要的思维方式. 比如主成分分析(Principal Components Analysis, PCA)完全可以使用Linear Algebra的知识得到, 机器学习中常用的KL散度(Kullback–Leibler Divergence)依靠概率论和信息理论即可得到.

此文档的写作参考了许多材料. 这些资料主要有Garrett Thomas的*Mathematics for Machine Learning*, 李航的《统计学习方法》, 大名鼎鼎的*Pattern Recognition and Machine Learning*, Lan GoodFellow等人的*Deep Learning*, 国内几本常用的工科数学教材以及Stephen Boyd的*Convex Optimization*等.

由于长时间浸淫在科技论文的英文写作环境中, 导致这篇文档许多地方'中英文混杂'. 原因有二: 首先, 作为学术工作者, 大家更在意的是你的idea, 只要能够清晰地表达自己的观点, 呈现的形式并不重要; 其次, 许多术语的中文翻译不甚清楚, 为了防止引起歧义, 我将会采用该术语被定义时所使用的语言.

自2018年9月14日启笔, 完成时应当至少有200页. 希望由此文档能够顺利完成.

You are free to distribute this document as you wish. The latest version can be found at my GitHub Page, please report any mistakes to `hliangzhao97@gmail.com`.

2 What is Linear Algebra?

2.1 Vector spaces

我们究竟身处于怎样的空间(space)中? 数据(data)和对数据的操作(operation)定义在何处? 在我们的观念中, 空间中包含着一切事物. 空间可对应着集合(set), 空间中的事物则可看成是集合中的元素(element). 因此, 不妨用集合定义我们所处的这个世界. 想弄明白我们处在怎样一个世界, 首先要挖掘这个世界的特征(feature). 那便是, 这个世界是可以被度量(metric)的, 它至少

具备这些观念(notions): 距离(distance), 长度(length), 角度(angle), etc. 数学家们将我们所身处的这个物理世界称之为欧几里得空间(Euclidean Space), i.e.,

Euclidean Space *is used to mathematically represent physical space, with notions such as distance, length and angles.*

回到我们最初的思路, 即用集合定义空间. 我们希望这个定义能够很好地抓住欧几里得空间的特征, 但是又不想仅仅描述这一种空间(兴许宇宙中存在着人类无法理解的空间结构呢:-D), i.e., 我们应当采用一种更广义的方式(a more general way)来定义空间这个概念. 于是就有了线性空间(or向量空间, or线性向量空间, vector space)这个概念:

线性空间V是一个内部元素被称为向量(vector)的集合. 该集合内的元素上定义了加法(addition)和数乘(scalar multiplication)[1]两项基本操作, 这一集合至少具备以下特性:

1 集合中存在零元($\forall \mathbf{v} \in V, \mathbf{v} + \mathbf{0} = \mathbf{v}$);

2 集合内任意元素均存在逆元、单位元($\forall \mathbf{v} \in V, \mathbf{v} + (-\mathbf{v}) = \mathbf{0}, 1\mathbf{v} = \mathbf{v}$);

3 加法和数乘运算满足交换律($\forall \mathbf{x}, \mathbf{y} \in V, \mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$),

4 结合律($\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V, \alpha, \beta \in \mathbb{R}, (\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z}), \alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$),

5 以及分配律($\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V, \alpha, \beta \in \mathbb{R}, \alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \beta\mathbf{y}, (\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{y}$).

接下来我们将会看到, 如何通过加强条件让线性空间一步步转变成我们所身处的欧几里得空间. 回望整个数学史和物理史, **人类从未停止过用简洁的真理(如公式, 定理)归纳整个宇宙终极密码的尝试**. 1931年狄拉克从理论上用极精美的公式预言: 磁单极子是可以独立存在的. 杨振宁评此方程'性灵出万象, 风骨超常伦'; 霍金一生(至少年轻时)都在追求能够统一广义相对论和量子力学的万物理论(the Theory of Everything). 这股精神在线性空间中亦得到了充分彰显.

首先请思考, 线性空间中的元素(即vector)之间具有怎样的关系? 两个向量之间可以相加, 这会形成新的向量, 每个向量乘上不同的标量(scalars)又会得到新的向量, 仔细观察向量加法和数乘的定义, 如何用一个式子将两种运算组合在一起呢? 那就写到一起好了:

$$\mathbf{v} \leftarrow \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + ... + \alpha_n\mathbf{v}_n, \mathbf{v}_i \in V, i \in \{1, 2, ..., n\}, \tag{2.1}$$

于是就有了线性组合(linear combination)这个概念! 只要$n-1$个标量为0, 便是数乘; 标量均为1, 便是向量加法. 显然通过这两种方式衍生出的向量是无穷无尽的. 因此, 相较于关注这些向量本身, 数学家们更愿意挖掘这些向量通过名为线性组合的运算所组成的全体(set)所具有的特性. 我们称这个集合全体为这组向量的张成(span). 若这组向量$\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\}$满足

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + ... + \alpha_n\mathbf{v}_n = 0 \quad \text{implies} \quad \alpha_1 = \alpha_2 = ... = \alpha_n = 0, \tag{2.2}$$

也就意味着这些向量之间不能通过两项基本运算相互得到, i.e., 这组向量作为一个群体而言, 任意一个元素均是不可或缺的; 同时, 这一组元素又可以通过线性组合表示span$\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\}$中的所有向量. 很自然地, 如果这组元素的span就是向量空间V全体, 那么数学家们不就实现了他们的目的吗! 即**用少数几个元素表达整个全体的特性**!

我们知道, 向量空间其实就是由向量们组成的集合罢了. 因此若集合内元素不同或元素个数不同, 那么就是不同的向量空间了(different vector spaces). 如何区别这些不同的向量空间?

数学家们将想到的方式是: 用那些能够代表整个向量空间全体的少数向量的个数来区分. 其实这很容易理解, 能够表征整个向量空间全体的少数几个元素, 想必各自代表了这个向量空间的一个元信息(meta-information), 我们将这样的一个方向的信息定义为维度(dimension), 这些元素作为这个向量空间的基石, 我们称之为基(basis).

[1]此处的数乘我已经默认为**被实数乘**. 实际上, 数乘运算可以定义在任意域$\mathbb{F}$(field)上, 此处取$\mathbb{F} = \mathbb{R}$是为了避免将运算引入抽象代数的分支中——这就与本文档的目的相悖了.

基中元素的个数称为这个向量空间的维度. 丰富的数学学习经验告诉我们一定要严格区分无穷(infinite)与有穷(finite)的情形, 因为**就算是同一事物, 其某一组成部分从有穷增加到无穷的时候, 性质都会发生显著变化(函数和数列的极限中可找到许多例子)**. 如果一个线性空间是由有穷个向量的张成得到的, 我们就称这个线性空间是**有穷维的(finite-dimensional)**. 否则这个线性空间就是**无穷维的(infinite-dimensional)**. 本文档只关注有穷维线性空间V, 其维度记为$\dim V$.

为了加深理解, 不妨将思绪从星辰大海中拉回这个真实的世界. 在我们所处的三维空间中, 三个方向便可描述整个空间; 同样地, 在平面坐标系上, 我们只需画出一个直角坐标系便可描述这个平面; 对于一条直线而言, 一个向量的数乘即可完美诠释整个直线. 当然, 点线面以及空间均是我们所身处的这个欧几里得空间中的概念, 他们都是很具体的linear spaces. 不过, 在来自不同的空间结构的外星人造访我们所认知的宇宙之前, 我们还是先思考一下如何为向量空间**赋予更多特性**使其成为我们所身处的物理世界吧.

2.1.1 Euclidean space

前文已经说过, 数学家们定义向量空间的概念正是受到了我们所身处其中的**欧几里得空间**的启发, 并且向量空间是一个广义的定义. 很自然地, 在为线性空间引入更多特性之前, 我们可以从向量空间中引出欧几里得空间的**具体表现形式**. 此外, 还需要说明的是, 欧几里得空间虽然诞生于我们所身处的**三维**空间(请注意, 此处所提到的'维'与向量空间的维(dimension)虽然本质一样, 但这只是日常生活中并不正式的称呼), 但维度可不仅仅限于三——数学家们喜欢抽象而普适的定义.

严格地说, **线性空间只是一个抽象的、描述集合的规范——它并未对集合内元素的表现形式做任何限定**. 实际上, 线性空间可以定义在任何**域(field)**$\mathbb{F}$上. 而**欧几里得空间就是线性空间定义在$\mathbb{F}=\mathbb{R}$上的结果**. 因此有:

> **欧几里得空间是被定义在实数域$\mathbb{R}$上且$\dim V = n$的线性空间, 记为$\mathbb{R}^n$. 其内部元素x(也就是vector)由n元实数组(n-tuples of real numbers)构成:**
>
> $$\mathbf{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}=\begin{bmatrix} x_1, ..., x_n \end{bmatrix}^\top.$$
>
> **加法和数乘运算则按元素的分量逐个对应(component-wise)作出:**
>
> $$\mathbf{x}+\mathbf{y}=\begin{bmatrix} x_1+y_1 \\ \vdots \\ x_n+y_n \end{bmatrix}, \alpha\mathbf{x}=\begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}.$$

尽管对于生活在三维空间的我们而言, $n>3$时会变得难以想象(hard to visualize). 但是从数学的角度出发, 这些概念的生成是简洁而优美的. 因此我们有理由相信这个神秘的宇宙存在生活在更高维度空间的生命体:-D.

在欧几里得空间中, **向量是按照列进行排列的阵列**(column vectors). 一个有意思的问题是, 为什么是按列排列呢? 为了减小纸面上空间的浪费, 还需要借助转置符号来描述, 为什么不直接将向量定义为按行排列的阵列? 其实是为了处理方便. 数学家们研究线性代数的一个重要动机是**给出线性方程组$\mathbf{Ax}=\mathbf{b}$的通解和解的结构**, 未知变量$\mathbf{x}$及常数项$\mathbf{b}$就是按照列来排列的, 直接将

这种表现形式挪用过来就形成了列向量.

2.1.2 Subspaces

线性空间仅仅是内部元素具有特定性质的集合. 因此, 集合论中的结论在此处显然也适用. 集合与集合之间可以存在包含关系, 类似地, 一个线性空间也可以包含另一个线性空间. 这便引出了子空间(subspace)的定义.

> **设V和S均为向量空间且$S \subseteq V$, 若S满足**
> **1** $\mathbf{0} \in S$;
> **2** S**对于向量加法运算是一个闭包(closure)**$(\forall \mathbf{x}, \mathbf{y} \in S, \mathbf{x}+\mathbf{y} \in S)$,
> **3 对于向量数乘运算也是一个闭包**$(\forall \mathbf{x} \in S, \alpha \in \mathbb{R}, \alpha\mathbf{x} \in S)$,
> **则称向量空间S是向量空间V的子空间.**

如何理解子空间的概念? 以欧几里得空间为例, 我们在$\mathbb{R}^3$中建立一个坐标系, 那么穿过原点的直线、包含原点的平面均可看成$\mathbb{R}^3$的子空间. 为什么子空间一定要包含零元? 这是因为子空间**首先得是一个线性空间**[2], 然后才能是一个子空间. 子空间S包含零元总是可以实现的——如果$\mathbb{R}^3$中的一条直线没有穿过原点, 总可以**通过平移、旋转或者反射坐标系等操作**让其通过原点, 但这时候$V = \mathbb{R}^3$就不再是原来的线性空间了. 上述这些操作就对应这下文即将阐述的线性映射(or线性变换, linear maps).

显然, 任意线性空间V总是其自身的子空间. 若一个线性空间仅包含$\mathbf{0}$, 则称其为**trivial vector space**.

关于线性空间的求和(sum)、求和所形成的新的线性空间的维度与原空间的关系暂且不做阐述.

2.2 Linear maps

上一节我们已经稍稍涉及到了线性变换(linear map), 这里给出线性变换严格的定义.

> **线性变换T是一个从线性空间V到线性空间W的一个映射(函数), i.e.,** $T : V \to W$, **且满足:**
> **1** $\forall \mathbf{x}, \mathbf{y} \in V, T(\mathbf{x}+\mathbf{y}) = T\mathbf{x} + T\mathbf{y}$;
> **2** $\forall \mathbf{x} \in V, \alpha \in \mathbb{R}, T(\alpha\mathbf{x}) = \alpha T\mathbf{x}$.

从任意向量空间V到其自身的线性变换称之为**线性算子(linear operator)**. 上一节所述的对三维空间(的坐标系)做平移、旋转或反射等变换, 得到的仍然是三维空间, 因此这种线性变换就是线性算子.

仔细观察线性变换的定义, 我们会发现两个需要满足的条件均是为了保证变换后得到的新空间的结构不被破坏——因为这两个条件将线性变换看成了一种运算——一种保留了线性空间加法和数乘特性的运算. **在英文文献中这段话的描述如下: The definition of a linear map is suited to reflect the structure of vector spaces, since it preserves vector spaces' two main operations, addition and scalar multiplication.**

线性空间在线性变换的作用下, 会引出一系列有趣的观念. 在代数领域, 线性变换被称为**向量空间的同态(homomorphism of vector spaces)**. 一个可逆的、向量空间的同态被称

[2] 由定义可知, 线性空间必须包含零元.

为**向量空间的同构(isomorphism of vector spaces)**. 如果存在一个从向量空间V到W的同构(isomorphism), 则称**向量空间V和W是同构的(isomorphic), 记为**$V \cong W$. 顾名思义, 我们说两个事物**同构**就是说二者在某一层面上(通常指较为本质的架构)可以看成是等价的. 类似地, 向量空间V与W同构意味着**二者在代数结构上等价**.

定义在同一域上、维度相同的有限维向量空间总是同构的. 对于欧几里得空间而言, 任意n维线性空间总是与$\mathbb{R}^n$同构. 这一点很容易理解: 对一个三维空间做平移、旋转或反射等变换, 得到的仍然是三维空间, 因此它们均与$\mathbb{R}^3$同构.

2.2.1 What is matrix?

在深入理解线性变换之前, 我们先来引入矩阵(matrix)的概念. 为此, 首先回顾一下那个经典而古老的问题: 对于含有n个未知数, 由m个线性方程所组成的方程组[3]

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2 \\ \cdots \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = b_n \end{cases} \tag{2.3}$$

是否有解? 若有解, 解是否唯一? 若有多解, 如何求出多解?

显然, 线性方程组(2.3)关于上述问题的答案完全取决于它的$m \times n$个系数a_{ij}和右端的常数项所组成的 $(m+1)$行n列矩形数表. 当我们把对线性方程组的处理放到计算机中执行时, 对于解决问题而言冗余的部分能省则省. 很自然地, 我们就可以给出矩阵的定义

$$\mathbf{B} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{bmatrix}. \tag{2.4}$$

这样就可以把线性方程组问题简单描述为$(\mathbf{A})_{m\times n}\mathbf{x} = \mathbf{b}$, 其中$\mathbf{A}$是系数矩阵, $\mathbf{x} \triangleq [x_1, ..., x_n]^\top$是未知数矩阵, $\mathbf{b} \triangleq [b_1, ..., b_m]^\top$是常数项矩阵, 而(2.4)给出的$\mathbf{B}$则是增广矩阵.

简言之, 矩阵以二维阵列的形式归纳出了对原始问题的解起核心作用的参数.

2.2.2 The matrix of a linear map

考虑从n个变量$x_1, x_2, ..., x_n$到m个变量$y_1, y_2, ..., y_m$的一个线性变换:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \\ y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n \\ \cdots \\ y_n = a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n \end{cases} \tag{2.5}$$

此处的**线性变换**从表现形式上看似乎并非上文给出了严格定义的**线性变换**, 但实际上它们是一回事. 为什么? 我们将n个变量$x_1, x_2, ..., x_n$看成一个整体, 也就是来自$\mathbb{R}^n$的一个向量$(\mathbf{x})_{n\times 1}$; 同样地, 将m个变量$y_1, y_2, ..., y_m$看成来自$\mathbb{R}^m$的一个向量$(\mathbf{y})_{m\times 1}$, **这不就意味着系数矩阵$\mathbf{A} = (a_{ij})_{m\times n}$扮演了'协助从线性空间$\mathbb{R}^n$到线性空间$\mathbb{R}^m$的一个线性变换 (即linear map)'这样的一个角色吗!**

也就是说, 每一个矩阵$\mathbf{A} \in \mathbb{R}^{m\times n}$总是可以对应一个线性变换$T : \mathbb{R}^n \to \mathbb{R}^m$:

$$\mathbf{y} = T\mathbf{x} = \mathbf{A}\mathbf{x}. \tag{2.6}$$

[3]显然这个问题是定义在实数域$\mathbb{R}$(本文档默认)上的.

因此可以得出结论：**矩阵是实施线性变换(投影，反射，旋转，拉伸，etc.)的载体.** 例如矩阵 $\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}$ 可看作是将$\mathbb{R}^2$上的任意向量$[x,y]^\top$向x轴进行投影变换的结果(变成了$[x,0]^\top$); 高中数学就学过的矩阵 $\begin{bmatrix}\cos\phi & -\sin\phi\\ \sin\phi & \cos\phi\end{bmatrix}$ 可看作是将$\mathbb{R}^2$上的任意向量$[x,y]^\top$作'以原点为中心逆时针旋转ϕ角'的旋转变换. 线性变换可能会给向量带来**降维打击**[4], 也可能给低维生物赋予在高维空间中的表现形态——这就由矩阵的行数和列数的大小关系来决定了.

从这里我们我们也可以窥见矩阵运算为什么会按照教材上给出的方式来进行——我们都知道矩阵与矩阵的乘法是前者的行乘后者对应的列, 但是大家思考过为什么吗?

考虑两个线性变换

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ y_2 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \end{cases} \tag{2.7}$$

$$\begin{cases} x_1 = b_{11}t_1 + b_{12}t_2 \\ x_2 = b_{21}t_1 + b_{22}t_2 \\ x_3 = b_{31}t_1 + b_{32}t_2 \end{cases} \tag{2.8}$$

若想从中得到$[t_1,t_2]^\top$到$[y_1,y_2]^\top$的线性变换, 可将(2.8)带入(2.7), 得到

$$\begin{cases} y_1 = (a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31})t_1 + (a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32})t_2 \\ y_2 = (a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31})t_1 + (a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32})t_2 \end{cases} \tag{2.9}$$

而线性变换(2.9)的系数矩阵正是(2.7)中的矩阵 $\mathbf{A} = (a_{ij})_{2\times 3}$和(2.8)中的矩阵$\mathbf{B} = (b_{ij})_{3\times 2}$ 相乘的结果. 也就是说, 两个矩阵相乘可以看成是对某个目标向量**累次进行两次线性变换**. 也是**矩阵乘法要求前者的列数等于后者的行数**的原因——否则将(2.8)带入(2.7)一定会出错!

既然每一个矩阵$\mathbf{A} \in \mathbb{R}^{m\times n}$总是可以对应一个线性变换$T:\mathbb{R}^n \to \mathbb{R}^m$, 那么, 已知一个线性变换, 是否总是可以对应一个矩阵? 答案是肯定的[5]. **实话说, 对这一部分的解释一度让我十分头疼. 我自己本人花了不少时间反复思索、联想才对此有了一点浅显的理解. 此外, 从自己弄懂到阐述清晰让大家都弄懂之间可能存在巨大的鸿沟. 此处的解读需要使用后文才引入的内积空间的概念, 暂且先拿来使用.**

请思考: 为什么空间坐标系要被定义为**直角**坐标系? 本节对此暂不作解释. 我们首先运用已有的数学经验, 定义欧几里得空间$\mathbb{R}^n$最常用的一组标准正交基(orthonormal basis): $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n$, 其中

$$\mathbf{v}_i = \begin{bmatrix} 0, ..., 0, 1, 0, ..., 0 \end{bmatrix}^\top, \tag{2.10}$$

即第i个位置的分量为1, 其余均为0. 那么$\mathbb{R}^n$中的任意向量$\mathbf{v}$总可以通过这一组基的线性组合得到, i.e.,

$$\mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + ... + \alpha_n\mathbf{v}_n = \begin{bmatrix} \alpha_1, \alpha_2, ..., \alpha_n \end{bmatrix}^\top.$$

同样地, 定义欧几里得空间$\mathbb{R}^m$最常用的一组标准正交基为$\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n$, 其中$\mathbf{w}_j$第$j$个位置的分量为1, 其余均为0. 若矩阵$(\mathbf{A})_{m\times n}$是线性变换$T:\mathbb{R}^n \to \mathbb{R}^m$所对应的矩阵, 那么对于$\mathbb{R}^m$中的某

[4]此处的'降维打击'和《三体》中提到的概念不是一回事呦:-D.

[5]注意本节的架构. 我首先论证一个矩阵总能推导出一个对应的线性变换; 其次论证一个线性变换也总能推导出一个对应的矩阵.

一向量$\mathbf{w}$, 由(2.6)可得

$$\begin{aligned}\mathbf{w} &= \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} = \alpha_1 \begin{bmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{m1} \end{bmatrix} + ... + \alpha_n \begin{bmatrix} A_{1n} \\ A_{2n} \\ \vdots \\ A_{mn} \end{bmatrix} \\ &= \alpha_1 \mathbf{A}\mathbf{v}_1 + ... + \alpha_n \mathbf{A}\mathbf{v}_n = \alpha_1 T\mathbf{v}_1 + ... + \alpha_n T\mathbf{v}_n.\end{aligned} \tag{2.11}$$

仔细观察(2.11), 将 $\begin{bmatrix} A_{1j}, A_{2j}, ..., A_{mj} \end{bmatrix}^\top$ 按每个分量拆开, 有

$$\begin{aligned}\begin{bmatrix} A_{1j} \\ A_{2j} \\ \vdots \\ A_{mj} \end{bmatrix} &= A_{1j} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + A_{2j} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + ... + A_{mj} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \\ &= A_{1j}\mathbf{w}_1 + A_{2j}\mathbf{w}_2 + ... + A_{mj}\mathbf{w}_m.\end{aligned} \tag{2.12}$$

结合(2.11)与(2.12), 可得

$$\begin{aligned}\mathbf{w} &= \alpha_1 T\mathbf{v}_1 + ... + \alpha_j T\mathbf{v}_j + ... + \alpha_n T\mathbf{v}_n \\ &= \alpha_1(A_{11}\mathbf{w}_1 + ... + A_{m1}\mathbf{w}_m) + ... + \alpha_j(A_{1j}\mathbf{w}_1 + ... + A_{mj}\mathbf{w}_m) + ... \\ &+ \alpha_n(A_{1n}\mathbf{w}_1 + ... + A_{mn}\mathbf{w}_m).\end{aligned} \tag{2.13}$$

(2.13)的结果表明, **线性变换$T : \mathbb{R}^n \to \mathbb{R}^m$所推导出的矩阵$(\mathbf{A})_{m\times n}$的第$j$列正是$T\mathbf{v}_j$在$\mathbb{R}^m$选定的基下的系数坐标.** 至此, 我们已经给出了线性变换T所推导出的矩阵的具体表现形式.

从本节开始, 事情似乎开始变得复杂而有趣了. 不知道大家是否观察到一处细节: 既然$T : V \to W$是一个映射(函数), 为什么不加上括号? 我从头到尾使用的符号都是$T\mathbf{x}$而非$T(\mathbf{x})$. 这是因为**线性变换总是使用矩阵来实现其具体操作的, 去掉括号我们可以十分'接地气地'将线性变换这个操作看成左乘一个矩阵.** 前文说过, 两个矩阵相乘可以看成是对目标向量接连进行两次线性变换, 因此**线性变换的复合(composition of linear maps)**我们写成ST, 而非$S \circ T$这样传统的方式.

2.2.3 Nullspace, range

现在我们来回忆一下**齐次线性方程组**问题, 即$(\mathbf{A})_{m\times n}(\mathbf{x})_{n\times 1} = \mathbf{0}$. 设矩阵$\mathbf{A}$对应的线性变换为$T : V \to W$, 则齐次线性方程组的解集可以表示为

$$\text{null}(T) = \{\mathbf{v} \in V | T\mathbf{v} = \mathbf{0}\}, \tag{2.14}$$

我们称其为**线性变换T的零空间(nullspace)**, 代数学家们称其为**核(kernel)**.

对于**非齐次线性方程组问题**, 即$(\mathbf{A})_{m\times n}(\mathbf{x})_{n\times 1} = (\mathbf{b})_{m\times 1}$, 由(2.11), 我们总可以将$\mathbf{A}\mathbf{x}$看成矩阵$\mathbf{A}$的$n$个列向量的线性组合, 系数分别为向量$\mathbf{x}$的各个分量. 数学家们称这些通过线性组合得到的向量的集合为**矩阵A的列空间(columnspace of matrix A)**. 向量$\mathbf{x}$作为我们想要寻找的未知变量, 意味着可以取任意值, 而原方程组有解等价于**来自$\mathbb{R}^m$的向量$\mathbf{b}$位于矩阵$\mathbf{A}$的n个列向量的线性组合所形成的集合中**. 设$T : V \to W$是矩阵$\mathbf{A}$所对应的线性变换, 这些满足条件的$\mathbf{b}$所组成的集合被称为**线性变换T的值域(range of T)**, i.e.,

$$\text{range}(T) = \{\mathbf{w} \in W | \exists \mathbf{v} \in V, T\mathbf{v} = \mathbf{w}\}. \tag{2.15}$$

显然, **任意矩阵A的列空间就是A所对应的线性变换T的值域, 因此可以采用**range($\mathbf{A}$)**代表A的列空间,** 用range($\mathbf{A}^\top$)代表$\mathbf{A}$的行空间(rowspace).

任意线性变换T的nullspace和range均是其定义域(domain)V和陪域(or上域, codomain)W的子空间. **借助线性空间的定义即可证明.** 此时线性变换作为一种保留了向量空间加法和数乘的运算在证明过程中功不可没.

涉及到秩的概念的, 暂时按下不表.

2.3 Metric spaces

回到一开始我们对于线性空间的介绍. 我们最初的目标是**为线性空间引入更多的特性, 使其一步步地称为我们所身处的这个物理世界**. 为此, 有哪些特性需要引入呢? 实际上我已经在第一节提到了, 即**两个事物之间的距离(distance)、事物本身的长度(length)**以及**两个不同方向的事物之间的角度(angle)**. 这三个特性是层层递进的——两两之间的后者是对前者概念的扩展, i.e., 满足后者特性的事物一定具备前者特性. 这三个特性层层递进式地作用在向量空间, 就分别得到了**度量空间(metric spaces)、赋范空间(normed spaces)**以及**内积空间(inner product spaces)**.

接下来给出严格的定义. 首先给出度量函数的概念:

> **定义在集合S上的度量(metric)是一个函数(映射)**$d: S\times S\to\mathbb{R}$, $\forall x,y,z\in S$**满足**
> **1** $d(x,y)\geq 0$**, 当且仅当**$x=y$**时取等;**
> **2** $d(x,y)=d(y,x)$**;**
> **3** $d(x,z)\leq d(x,y)+d(y,z)$**(三角不等式).**

度量函数使用形式化的数学语言定义了这个物理世界的**距离**这一概念. 距离被用来度量两个事物之间的位置关系, 这一关系与二者顺序的先后无关, 是对二者关系最精简的描述. **在其上定义了度量函数的线性空间就是度量空间**. 度量函数是从两个实数之间的距离扩展出来的概念[6]——**它使得极限这一概念可以定义在来自任意集合S的任意事物上:**

若对于任意的$\epsilon>0$, 总存在$N\in\mathbb{N}$使得$n>N$时满足$d(x_n,x)<\epsilon$, 则称序列$\{x_n\}\subseteq S$收敛于极限x.

2.4 Normed spaces

接下来我们为线性空间引入长度(length)的概念. 和度量函数一样, 我们首先给出范数的概念:

> **定义在线性空间V上的范数(norm)是一个函数(映射)**$\|\cdot\|: V\to\mathbb{R}$, $\forall\mathbf{x},\mathbf{y}\in V,\alpha\in\mathbb{R}$**,** **满足**
> **1** $\|\mathbf{x}\|\geq 0$**, 当且仅当**$\mathbf{x}=\mathbf{0}$**时取等;**
> **2** $\|\alpha\mathbf{x}\|=|\alpha|\|\mathbf{x}\|$**;**
> **3** $\|\mathbf{x}+\mathbf{y}\|\leq\|\mathbf{x}\|+\|\mathbf{y}\|$**(三角不等式).**

赋予了范数概念的线性空间就是赋范空间(normed space). V上的任意范数总是可以推导出V上的一个度量函数, i.e.,

$$d(\mathbf{x},\mathbf{y})=\|\mathbf{x}-\mathbf{y}\|. \tag{2.16}$$

[6] 若$x,y\in\mathbb{R}, d(x,y)=|x-y|$.

One can verify that the axioms for metrics are satisfied under this definition[7] and follow directly from the axioms for norms. 这意味着**任意的赋范空间总是一个度量空间**. 若某一赋范空间对于度量函数的运算满足**完备性**[8], 那么这个赋范空间就称之为**巴拿赫空间(Banach space)**.

在欧几里得空间上[9], 最常用的范数是p-范数, 定义为

$$\|\mathbf{x}\| \triangleq \left(\sum_{i=1}^{n}|x_i|^p\right)^{\frac{1}{p}}, p \geq 1. \tag{2.17}$$

1-范数、2-范数以及无穷范数分别是(2.17)中的p取1、2、∞[10]下的结果:

$$\begin{aligned}\|\mathbf{x}\|_1 &= \sum_{i=1}^{n}|x_i|,\\ \|\mathbf{x}\|_2 &= \sqrt{\sum_{i=1}^{n}x_i^2},\\ \|\mathbf{x}\|_\infty &= \max_{1\leq i\leq n}|x_i|.\end{aligned} \tag{2.18}$$

为什么p-范数一定要满足$p \geq 1$呢? 这是因为**倘若该条件不满足, 范数的定义中第三条——三角不等式便不成立.**

关于范数有一个有趣的结论.

> **对于任意有穷维向量空间V上的任意向量$\mathbf{x}$, 对于任意两类范数$\|\cdot\|_A, \|\cdot\|_B$, 总存在实数$\alpha, \beta > 0$使得**
>
> $$\alpha\|\mathbf{x}\|_A \leq \|\mathbf{x}\|_B \leq \beta\|\mathbf{x}\|_A,$$
>
> **这意味着在有限维线性空间V上, 同一元素的任意两种范数在敛散性上是等价的. 这一优秀的特性将会在后续关于凸优化理论的介绍中大放异彩.**

2.5 More details about matrices

按照进度, 接下来应当介绍将角度(angle)引入线性空间之后所擦出的一系列火花. 不过, 引入角度概念的内积空间的许多优美特性的证明需要大量矩阵运算及变换的技巧, 因此, 不妨先熟悉一下矩阵相关的基本操作.

2.5.1 Basic properties of matrices

本节介绍对角矩阵、恒等矩阵的由来以及转置等运算的特性.

我们知道, 矩阵是实施线性变换(投影, 反射, 旋转, 拉伸, etc.)的载体. 回忆(2.5), (2.6), 不可避免地, 我们会关注一些特殊的线性变换, 它们对应着特殊的矩阵.

线性变换

$$\begin{cases} y_1 = \lambda_1 x_1\\ y_2 = \lambda_2 x_2\\ \cdots\\ y_n = \lambda_n x_n \end{cases} \tag{2.19}$$

[7]'this definition'指的就是(2.16).

[8]完备性是指在某一空间下的极限运算结果仍属于该空间. 就加法而言, 实数集具有完备性, 无理数集则不具备.

[9]之所以强调'在欧几里得空间'上, 是因为到目前为止我们只给出了欧几里得空间内元素的具体表现形式. *按元素分量对应作出(component-wise)*也只是欧几里得空间内的元素的运算特征.

[10]推导无穷范数的表达式需要用到极限相关知识.

对应于n阶**方阵**

$$\mathbf{\Lambda}=\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_n\end{bmatrix}=\mathrm{diag}(\lambda_1,\lambda_2,...,\lambda_n). \tag{2.20}$$

这便是**对角阵(diagonal matrix)**. **我们关心某个特殊矩阵总是因为它对应着一个特殊的线性变换.** 当$\lambda_1=\lambda_2=...=\lambda_n=1$时, (2.19)称之为**恒等变换**, (2.20)称之为**单位矩阵(identity matrix)**.

矩阵作为一种新的数学概念, 同样具备自己的运算规则——加法、数乘以及乘幂等. **关于矩阵的运算规则, 请总是将其与线性变换和线性方程组的求解联想到一起.**

前文我们已经给出了矩阵乘法的由来, 对于别的运算规则下文直接给出, 此处不再做详细解释. 大家可亲自从观念上进行验证.

1 矩阵的加法运算规则:

$$\begin{aligned}&\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A};\\ &(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C});\\ &\mathbf{A}+(-\mathbf{A})=\mathbf{O}\quad\text{induces}\quad\mathbf{B}-\mathbf{A}=\mathbf{B}+(-\mathbf{A}).\end{aligned} \tag{2.21}$$

2 矩阵的数乘运算规则:

$$\begin{aligned}&(\lambda\mu)\mathbf{A}=\lambda(\mu\mathbf{A});\\ &(\lambda+\mu)\mathbf{A}=\lambda\mathbf{A}+\mu\mathbf{A};\\ &\lambda(\mathbf{A}+\mathbf{B})=\lambda\mathbf{A}+\lambda\mathbf{B}.\end{aligned} \tag{2.22}$$

3 矩阵乘法的运算规则(不满足交换律):

$$\begin{aligned}&(\mathbf{AB})\mathbf{C}=\mathbf{A}(\mathbf{BC});\\ &\lambda(\mathbf{AB})=(\lambda\mathbf{A})\mathbf{B}=\mathbf{A}(\lambda\mathbf{B});\\ &\lambda(\mathbf{A}+\mathbf{B})=\lambda\mathbf{A}+\lambda\mathbf{B}.\end{aligned} \tag{2.23}$$

4 引入矩阵的转置之后, 关于转置的运算规则:

$$\begin{aligned}&(\mathbf{A}^\top)^\top=\mathbf{A};\\ &(\mathbf{A}+\mathbf{B})^\top=\mathbf{A}^\top+\mathbf{B}^\top;\\ &(\alpha\mathbf{A})^\top=\alpha\mathbf{A}^\top;\\ &(\mathbf{AB})^\top=\mathbf{B}^\top\mathbf{A}^\top.\end{aligned} \tag{2.24}$$

2.5.2 Deternimant

引入行列式的概念, 定义矩阵的行列式. 最后得到公式$\mathbf{A}\mathbf{A}^\star=|\mathbf{A}|\mathbf{E}$.

什么是行列式(determinant)[11]? 我们在任意一本工科教材上都可以轻易找到其定义, 但这远远不够. 为了给出行列式存在的意义, 接下来首先给出**欧几里得空间**上行列式的定义, 帮助大家回忆一下其计算方法.

[11]本节内容大量参考了维基百科关于行列式的主页, 感兴趣的读者可访问该页面获取更深入的了解.

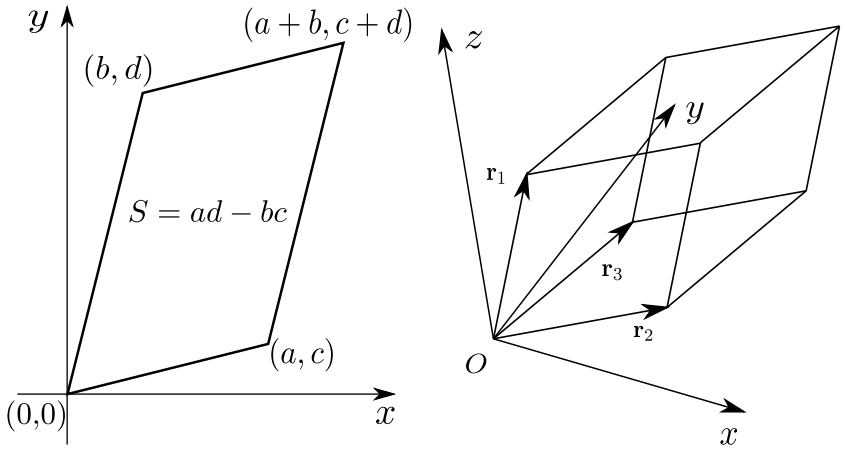
定义在欧几里得空间上的行列式是一个函数(映射)$|\cdot|:\mathbb{R}^{n\times n}\to\mathbb{R}$, 又记作$\det(\cdot)$, 它将来自$\mathbb{R}^{n\times n}$一个矩阵$\mathbf{A}$映射为一个标量:

$$\det(\mathbf{A})=\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\prod_{i=1}^{n}(\mathbf{A})_{i,\sigma(i)}, \tag{2.25}$$

其中S_n是自然数集合$\{1,2,...,n\}$上任意排列(or置换)的全体, $\operatorname{sgn}(\sigma)$是置换$\sigma$的符号差[12].

因为对于任意自然数n, 集合S_n中有$n!$个元素(即全排列的个数), 因此n阶方阵$\mathbf{A}$的行列式是$n!$个求和项相加的结果, 是一个有限次的求和. 对于简单的2阶和3阶的矩阵, 行列式的表达式相对简单, 而且恰好是每条主对角线(左上至右下)元素乘积之和减去每条副对角线(右上至左下)元素乘积之和.

在明悉了定义之后, 我们需要理解的是, **行列式是对物理世界中n维平行体的体积*(volume)*这一概念的扩展**.



(a) 行列式与平行四边形　　(b) 行列式与三维平行体

图 2.1: 行列式是$\mathbb{R}^n$中的n维平行体的有向积

如图(2.1-a)所示, 在$\mathbb{R}^2$上, 由向量$\mathbf{v}\triangleq[a,c]^\top$和$\mathbf{w}\triangleq[b,d]^\top$所组成的矩阵的行列式为

$$\det(\mathbf{v},\mathbf{w})=\begin{vmatrix}a & b\\ c & d\end{vmatrix}=ad-bc. \tag{2.26}$$

如果我们为其在二维直角坐标系赋予具体数值, 可以轻易验证其这个行列式就是该平行四边形的(有向)面积. 何谓'有向'? 其意义是: **平行四边形面积为正当且仅当以原点为不动点将$\mathbf{v}$逆时针旋转到$\mathbf{w}$处时, 扫过的地方在平行四边形里, 否则面积则为负值**. 而行列式任意两列向量共线会导致值为0则是因为平行四边形已然消失.

在$\mathbb{R}^3$上, 行列式的有向体积由右手定则来确定, 此处不再展开阐述.

对于欧几里得空间上的行列式, 我们需要记住以下常用性质. $\forall\mathbf{A},\mathbf{I}\in\mathbb{R}^{n\times n}$,

1 $\det(\mathbf{I})=1$;

2 $\det(\mathbf{A}^\top)=\det(\mathbf{A})$;

[12]满足$1\le i\le j\le n$但$\sigma(i)>\sigma(j)$的有序实数对(i,j)称为σ的一个逆序. 若σ的逆序数有奇数个, 则符号差为-1, 否则符号差为1.

3 $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$;
4 $\det(\mathbf{A}^{-1}) = \det(\mathbf{A})^{-1}$;[13]
5 $\det(\alpha\mathbf{A}) = \alpha^n \det(\mathbf{A})$.

如果大家还有印象的话, 应该还记得关于行列式的知识点是在《线性代数》第一章节介绍的. 教材没有从任何角度给出行列式诞生的原因, 而是直接给出了定义和运算性质. 这些性质有:

(i) 对换行列式的两行(列), 行列式变号;

(ii) 把行列式的某一行(列)的各元素乘同一数然后加到另一行(列)对应的元素上去, 行列式不变.

从这些性质可以构造许多计算题. 在我看来, 这些题目不过是**对技巧的卖弄**, 这与做研究完全无关. 尽管如此, 我们需要掌握计算行列式的思维方式——化归. 即**将高阶行列式用低阶行列式来表示**. 这就引入了**余子式(minor)**和**代数余子式(cofactor)**的观念.

$\forall \mathbf{A} \in \mathbb{R}^{n\times n}$, **把**$(i,j)$**元**$A_{ij}$**所在的第**$i$**行和第**$j$**列划掉后, 留下来的**$n-1$**阶行列式称为元**$a_{ij}$**的余子式(minor), 记作**$M_{ij}$**.** $A_{ij} \triangleq (-1)^{i+j}M_{ij}$**称为元**$a_{ij}$**的代数余子式(cofactor). 此外, 有拉普拉斯展开(Laplace Expansion)**

$$\det(\mathbf{A}) = \sum_{j=1}^{n}(-1)^{i+j}a_{ij}M_{ij} = \sum_{i=1}^{n}(-1)^{i+j}a_{ij}M_{ij}. \tag{2.27}$$

$\det(\mathbf{A})$**的各个元素的代数余子式**A_{ij}**所组成的矩阵的转置称为**$\mathbf{A}$**的伴随矩阵(adjugate matrix), i.e.,**

$$\mathbf{A}^\star = \mathrm{adj}(\mathbf{A}) \triangleq (-1)^{i+j}M_{ji}. \tag{2.28}$$

当$i=j$**时,**

$$(\mathbf{AA}^\star)_{ij} = a_{i1}A_{j1} + a_{i2}A_{j2} + ... + a_{in}A_{jn} = \det(\mathbf{A}); \tag{2.29}$$

当$i\neq j$**时,** $(\mathbf{AA}^\star)_{ij} = 0$**. 因此**

$$\mathbf{AA}^\star = |\mathbf{A}|\mathbf{E}. \tag{2.30}$$

至此, 关于行列式的诸多概念已经阐述完毕.

2.5.3 Inverse of matrices

本节引入逆矩阵的概念, 为后文介绍克拉默法则奠定基础.

2.5.4 Useful matrix identities

第一个矩阵变换的一致性我们前文已经提及过——**将矩阵-向量乘法看成矩阵列向量的线性组合(matrix-vector product as linear combination of matrix columns)**, 具体变换过程见(2.11).

第二个一致性是**将向量外积(outer product)之和看成矩阵-矩阵的乘积**. 对于$\mathbf{a}\in\mathbb{R}^m, \mathbf{b}\in\mathbb{R}^n$, 外积$\mathbf{ab}^\top$满足

$$[\mathbf{ab}^\top]_{ij} = a_i b_j. \tag{2.31}$$

令$\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_k} \in \mathbb{R}^m, \mathbf{b_1}, \mathbf{b_2}, ..., \mathbf{b_k} \in \mathbb{R}^n$, $\mathbf{A}, \mathbf{B}$分别是以这些向量为列的矩阵, 即$(\mathbf{A})_{m\times k} \triangleq [\mathbf{a_1}...\mathbf{a_k}], (\mathbf{B})_{n\times k} \triangleq [\mathbf{b_1}...\mathbf{b_k}]$, 那么对于任意的$(i,j)$, 有

$$\left[\sum_{l=1}^{k}\mathbf{a}_l\mathbf{b}_l^\top\right]_{ij} = \sum_{l=1}^{k}[\mathbf{a}_l\mathbf{b}_l^\top]_{ij}. \tag{2.32}$$

[13]此处涉及到了'矩阵的逆', 将在下一节阐述.

为什么(2.32)成立? 这是因为(i,j)与累加符号无关, 可以放入累加符号内部. 结合(2.31)可得

$$\sum_{l=1}^{k}[\mathbf{a}_l\mathbf{b}_l^\top]_{ij}=\sum_{l=1}^{k}[\mathbf{a}_l]_i[\mathbf{b}_l]_j=\sum_{l=1}^{k}A_{il}B_{jl}=\sum_{l=1}^{k}(\mathbf{A})_{il}(\mathbf{B}^\top)_{lj}. \tag{2.33}$$

所以

$$\sum_{l=1}^{k}\mathbf{a}_l\mathbf{b}_l^\top=\mathbf{A}\mathbf{B}^\top. \tag{2.34}$$

最后一个矩阵一致性与二次型有关, 可能会放到对称阵相关章节阐述.

2.6 Inner product spaces

现在, 我们将角度(angle)的概念引入线性空间. 首先给出**内积(innner product)**的定义.

> **内积是定义在向量空间上的一个函数(映射)**$\langle\cdot,\cdot\rangle:V\times V\to\mathbb{R},\forall\mathbf{x},\mathbf{y},\mathbf{z}\in V,\forall\alpha\in\mathbb{R}$**满足**
>
> **1** $\langle\mathbf{x},\mathbf{x}\rangle\geq 0$**, 当且仅当**$\mathbf{x}=\mathbf{0}$**时取等;**
>
> **2 满足对第一变元的线性性:** $\langle\mathbf{x}+\mathbf{y},\mathbf{z}\rangle=\langle\mathbf{x},\mathbf{z}\rangle+\langle\mathbf{y},\mathbf{z}\rangle,\langle\alpha\mathbf{x},\mathbf{y}\rangle=\alpha\langle\mathbf{x},\mathbf{y}\rangle$**,**
>
> **3 以及正定性**[14]**:** $\langle\mathbf{x},\mathbf{y}\rangle=\langle\mathbf{y},\mathbf{x}\rangle$**.**

赋予了内积概念的线性空间就是**内积空间(inner product space)**. 定义在线性空间V中任意向量$\mathbf{x}$上的内积总是可以推导出V上的一个范数, i.e.,

$$\|\mathbf{x}\|=\sqrt{\langle\mathbf{x},\mathbf{x}\rangle}. \tag{2.35}$$

在这个定义下, 我们可以轻易验证内积运算所具备的公理总是可以推导出范数所需要满足的公理, i.e., **任意内积空间总是一个赋范空间, 因此也总是一个度量空间.** 如果内积空间V对度量函数的运算(induced by its innner product)满足完备性, 那么该内积空间就称之为**希尔伯特空间(Hibert space)**.

内积是在线性空间上对物理世界中的'角度'这一概念的模拟.

2.6.1 Pythagorean Theorem

2.6.2 Cauchy-Schwarz inequality

2.6.3 Orthogonal complements and projections

2.7 Solving the problem: Ax = b

2.7.1 Cramer's Rule

2.7.2 Elementary transformation of matrices

2.7.3 Rank of matrices

在本章节解释矩阵的秩、矩阵的行空间的秩与矩阵的列空间的秩的关系.

[14]结合第二点与第三点可知整个内积运算均满足线性性.

2.7.4 Solution of the problem

2.8 In-depth understanding of Linear Dependence

2.8.1 Rank of vector groups

2.8.2 Structure of the solution of Ax = b

2.9 Eigenthings

2.10 Trace

2.11 Orthogonal matrices

2.12 Symmetric matrices

2.13 Singular value decomposition

2.14 Fundamental Theorem of Linear Algebra

2.15 Low-rank Approximation

2.16 Pseudoinverse

3 Probability Theory

看待概率的观点, 我们从**Bayesian paradigm**和**Frequentist paradigm**的区别说起. 随后依次引入随机变量、概率密度函数以及概率分布等观念.

3.1 Basics

3.2 Random variables

3.3 Joint distributions

3.4 Great Expectations

3.5 Variance

3.6 Covariance

3.7 Random Vectors

3.8 Estimation of Parameters

4 Calculus

微积分的运算几乎是全部机器学习算法的基础. 本章首先给出极限的定义, 随后引入梯度和海森矩阵的概念, 最后给出去的局部极值的条件. **本章节的重点处理的是matrix calculus**.

4.1 Extrema

4.2 Gradients

4.3 The Jacobian

4.4 The Hessian

4.5 Matrix calculus

4.6 Taylor's theorm

4.7 Conditions for local minima

5 Optimization

终于进入了优化领域! Stephen Boyd花了七百多页才浅显涉及到的领域, 我自然无法透彻地阐明(实际上, 距离我理解此书还有不少差距:-D). **本文档的目标是解释观念诞生的原因, 不会过分注重高深的数学推导**.

5.1 Convexity

5.1.1 Convex sets

5.1.2 Convex functions

5.2 Convex Optimization problems

5.2.1 Linear optimization problems

5.2.2 Quadratic optimization problems

5.2.3 Geometric programming

5.3 Duality

5.3.1 The Lagrange dual function

5.3.2 Optimality conditions

5.3.3 Perturbation and sensitivity analysis

5.4 Unconstrained minimization

5.5 Equality constrained minimization

5.6 Interior-point methods

6 Optimization in Machine Learning

本章节参考的主要文档有Bottou等人于2018年2月份写就的*Optimization Methods for Large-Scale Machine Learning*, 来自微软研究院的Bubeck写就的*Convex Optimization: Algorithms and Complexity*, 这些monographs均可免费下载.

6.1 Batch gradient methods

6.2 Stochastic optimization methods

6.3 Second-order methods