

# 面向机器学习（深度学习、人工智能）的数学基础

## 技术专题介绍

淡蓝小点@Wechat:bluedotdot\_cn

《面向机器学习（深度学习、人工智能）的数学基础》是淡蓝小点准备上线的一项技术专题，顾名思义它主要介绍机器学习（包括深度学习、人工智能）专业所涉及到的数学基础知识。

其实最开始我并没有打算做这样一项专题，原因有二：

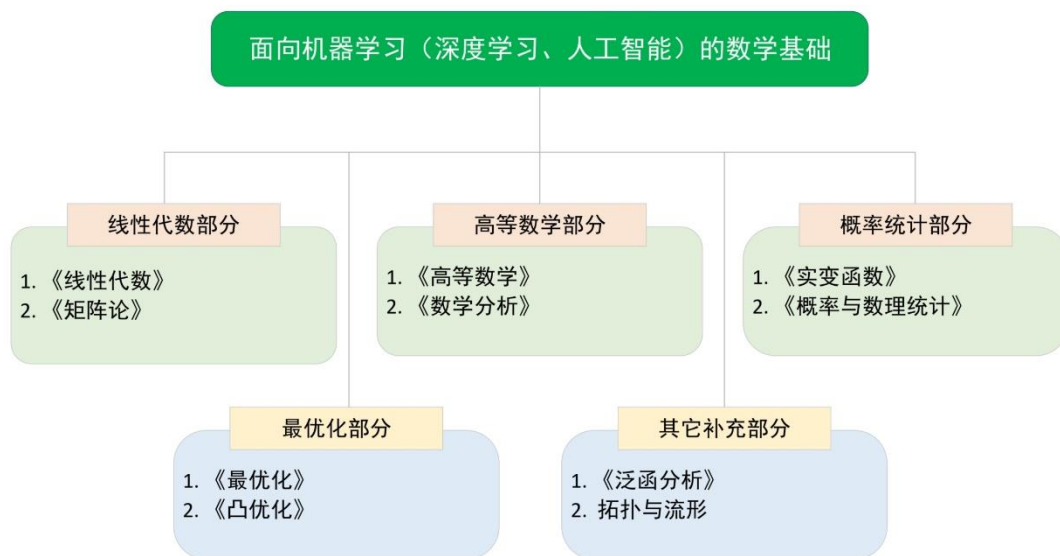
1. PRML Page-by-page 项目已经是一个自治的项目了，我认为 PRML 所需的数学知识已经包含在该项目中，都做了补充介绍，似乎不再需要单独的技术专题来配合它了。
2. PRML Page-by-page 项目的内容量非常大，制作起来需要耗费大量的时间精力，我也担心自己是否能将数学基础这么重要、这么宽泛的专题做好。

然而，道由心起本无意，因缘果报自成章。在跟关注 PRML 的朋友交流过程中，有很多人都提到在做机器学习、人工智能相关工作时，总感觉自己数学基础不够。很多人都认为至少读完 PRML 需要很好的数学基础，经常有人问我是不是数学系毕业，不然怎么能把推导过程讲的那么明白了？其实我不是数学系，我的数学水平也相当有限。在读大学时数学成绩主要徘徊在 80~90 分左右，低的时候 70+，高的时候 90+。我现在展现的一点微末道行，也是后面自学的。但交流的人多了，我就越发觉得似乎有不少人都需要一个专门针对机器学习的数学基础专题，帮助大家打牢数学基础，以更好更快的在人工智能的大道上前行。闲言少叙，下面讲点有关此技术专题的内容。

### 一、此技术专题包含哪些内容？

首先声明，包含的内容还在不断更新、整理、安排中，所以今天的答案不一定就是最终所呈现的内容。就目前而言，共分成五个部分：线性代数部分、高等数学部分、概率统计部分、最优化部分、其它补充部分。每个部分一般又会融合多门学科，将其中跟机器学习密切相关的内容提取出来，经过梳理后以恰当的形式呈现出来。例如线性代数部分，涉及到大学中的《线性代数》课程，并有相当一部分更专业的《矩阵论》内容，最后还有一点关于矩阵向量求导的内容。对机器学习有一定了解的同学应该有切身体会：仅仅依靠《线性代数》一门课程在机器学中是不够的，还有不多不少的内容需要额外补充。

在这几个部分中，前三个部分（淡绿色）的内容已经准备大半，尤其线性代数和概率统计部分已几乎准备好。高等数学部分是难度最大、概念最多、涉及面最广的，因此才准备了约三分之一。后两部分（淡蓝色）只是计划，至于如何实施、安排什么内容，到目前还没有具体想法（还没忙到这儿来）。



为了让大家有更具体的感受，我把线性代数部分的知识点在这儿罗列一下：  
(如果图片效果不好，请点击[此处浏览 PDF 版](#))

## 线性代数部分

### 矩阵

定义

常见矩阵：方阵、对称阵、单位阵、三角阵、对角阵

矩阵运算：加减、数乘、乘、转置、逆、左右逆、伪逆、内外积、Hadamard积、Kronecker积、卷积

### 矩阵的重要数值量

行列式

秩

迹

范数

条件数

其它数值量：特征值、奇异值、谱半径、惯性指数

### 线性空间及线性映射

线性空间定义

子空间：直和、仿射子集、商空间

向量空间：张成、线性无关、维数

线性映射：同构、矩阵表示、基变换、坐标变换、图像缩放平移旋转

### 特征值及特征向量

定义：左右特征向量

几何解释

重根与重数：代数重数、几何重数

重要性质

广义特征值特征向量

在PCA上的应用

### 矩阵分解

QR分解

LU分解

Cholesky分解

特征分解

谱分解

SVD

SVD在PCA上的应用

### 二次型及正定阵

定义：标准形、规范形

协方差是正定的

### 向量及矩阵求导

分子布局和分母布局

常用公式

### 参考文献

很明显，关于此技术专题的内容有两个重要问题：

- 1) 哪些内容应该被选中，哪些内容可以被丢弃
- 2) 选中内容如何呈现出来

对于这两个问题的答案：

- 1) 很明显，如何识别内容是否跟机器学习相关、是否重要既跟个人的经验水平有关，也跟个人的主观感受有关。总的而言我是根据我的经验进行选择的，这个经验的很大一部分又来源于我阅读 PRML 时的实际需求。所以很难说它一定适合所有人，但我个人觉得此专题应该能涵盖机器学习所需数学基础的绝大部分。
- 2) 如何呈现内容是另一个要费点脑筋的问题。如果只是把现有的教材拿出来照本宣科，那我觉得这个项目似乎就失去意义了。因为如果这样就可以的话，那大家完全可以去 B 站上搜索各个学科的公开课，跟着学一遍就行了。但问题是这样做就不叫“面向机器学习”的数学基础了，而且很多人也没有那么多时间去从头到尾都学一遍。也许线性代数部分还好，那么泛函、实变、拓扑这些内容怎么办？也像数学系的同学一样把这些课程都学一遍么？很显然不是这样的。所以经过考虑，我决定以专题的形式介绍相关内容。例如线性代数部分，有一个矩阵分解的专题专门介绍各类分解方法；又比如概率部分的贝叶斯专题，就专门介绍有关贝叶斯的思想和方法；更典型的是高等数学（数学分析）部分，连续、可导、级数都是一个个独立的专题。我觉得这样既能兼顾大家已经修过这些课程具有一定的基础，又能集中篇幅把一个概念讲透。比如连续，在介绍了它的定义和概念之后，在专题中会一口气介绍利普希茨连续、赫尔德连续、绝对连续、一致连续、一般连续、几乎处处连续，几乎将所有的连续一网打尽。

## 二、面向机器学习的数学和常规的数学课程有什么区别？

这是一个必须要回答的问题。如果面向机器学习的数学和数学课程没有区别，那只需要一个有经验的人把机器学习所涉及到的数学课程列下来，然后一个个去学就好了。理论上这样做也没问题，毕竟多学一点比少学一点要好。但在实际上这样却很难行通：

1. 没有那么多时间去一门门课程学习。  
这是一个很重要的现实状况，大家都很忙啊，即使是数学专业一个学期也只能学两到三门专业课。大家都有自己的主业，哪有那么多时间是一门门学习了？怎么来得及了？大家需要的是在较短、较为集中的时间内，快速突破。
2. 人工智能专业毕竟不是数学专业，很多课程不需要学那么多、那么深。  
这也是一个非常关键的问题。如果我们所需的跟数学专业一样深奥一样严谨，那就直接去数学系学习吧。但是机器学习毕竟是一个应用专业，它对数学的需求其实比数学系要少很多很多。即使是高等数学这门课，所需也仅是大家大一所学的  $N$  分之一。因此，把重要的部分选出来、形象的理解选出来的部分而非应付考试式的死记硬背，是制作这个专题时要时刻牢记的宗旨。

一个例子就是概率统计部分，第一个专题是有关概率的定义。什么是概率，这是一个常常被大家忽略的问题。在本专题中是按这样的逻辑介绍的：现代概率是公理化定义，所以先介绍概率的公理化定义；其次，满足这个公理化的一种定义就是用频率除以试验次数并趋于极限。但为什么这个定义能够满足公理化了？这就需要证明了。只有证明了才能认为它是一个合理有效的概率定义；那如何证明了？这就要用到概率中

的极限定理了，具体而言包括两个不等式（马尔可夫不等式、切比雪夫不等式）、三弱一强大数定律（切比雪夫、辛钦、伯努利）、两条中心极限定理（林德伯格-列维中心极限定理、棣莫弗-拉普拉斯中心极限定理）。只有这样，才算逻辑完整的理解了到底什么是概率，为什么概率会定义成这个样子。这就和一般教材呈现较大差异了。教材一般会先介绍古典概型、几何概型，然后介绍随机变量、期望、方差的概念，差不多到最后才介绍极限定理。这是可以理解的，因为前面的部分具体易理解，极限定义抽象不太好接受。但是作为有数学基础的从业者，我们不应该再按初学者的路径重复行走，而应找到更高、更好、更深入的视角去学习。

再比如机器学习很多地方都要用到流形，流形本身属于非欧几何，想要系统性的学习是需要很多数学基础而且难度很大的。但是有经验的人能体会到，很多时候我们更需要的是理解到底什么是流形？至于那些深奥的关于非欧几何的研究成果，绝大部分在绝大多数时候是不需要的（除非你是很深入的、专门做流形学习的学者）。因此，能够感性的理解什么是流形，可能在很多时候对我们而言就够了。这一点也适用于拓扑。拓扑也是非常抽象但又经常用到的，是否有必要那么深入的学习它了？我觉得没必要。能理解拓扑的定义，很多时候也就够了。这样既节省了时间也降低了难度。我觉得这可能才是大家真正需要的。

### 三、这个专题什么发布？以什么形式发布？什么时候发布？

我考虑应该会以直播+视频的形式发布。直播是为了跟大家拉近距离，大家学习的效率会更高更投入，而且直播时大家可以随时提问交流，我感觉这样比较好。但是由于内容太多，好像又不太可能将全部内容直播（几百页的PPT好像讲不完一样）。所以应该会直播一部分、视频一部分。直播的时间、方式到时候在群里告诉大家，也许和之前一样仍然就用腾讯会议。

还有一个可能大家比较敏感的问题，收费么？君子役物，小人役于物。完全免费为大家劳动，无论是经济上还是动力上似乎都很难长久和持续。但漫天要价、以次充好、以卖代骗也不是我所欲也。怎么办了？我觉得可行的办法是：“**收费合理尽量低价、充分了解完全自愿**”。大家骂我水平不够我可以接受，但是如果有人骂我坑蒙拐骗我会内心难过。毕竟前者是能力问题，后者是态度问题人品问题。不过万幸，大家对我比较包容。到目前为止无论是《PRML Page-by-page》项目还是《学习理论精炼介绍》都还没人骂过我，至少没当面骂我，我心甚慰，谢谢大家的善良。倒是有不少人跟我说价格公道非常值。微信群里就有朋友私信跟我说过，《PRML Page-by-page》的1699是他今年花的最值的一笔钱。他马上就要继续深造博士了，我也因为我的内容能对他有切实的帮助而感到高兴。OK，总结起来应该还是会收费。收费标准会参考《PRML Page-by-page》和《学习理论精炼介绍》，我估计应该在每小时15-25元之间吧。具体多少到时候再定。我个人觉得这个价格不算太贵太离谱吧，毕竟我听说小学生的补习课一个小时就要好几百，与之相比我简直是奴仆般的价格啊！当然为了做到“充分了解”，会向以往一样把学习材料免费送给大家，我想这样能最大程度的让大家避免“被我坑”。

就目前来看，这个专题的内容量也是比较大的。比如概率统计部分（含实变函数）的PPT大概有230页左右，如果每次直播两小时、每次讲40页（我感觉有可能讲不了这么多），也需要5~6天。而这仅仅只是整个专题的一部分而已。


我想在八月份之内就把概率统计部分发布出来，因为已经拖的太久了，大家在群里已经催过我很多遍了。

#### 四、其它问题

暂时没想到更多其它问题，大家有问题的话请给我留言，或者微信来跟我沟通。由于材料还没完全制作好，现在还不能直接给大家。需要的话可以先找我拿个半成品。

最后：本文应该会多次更新，感谢大家的支持和认可。感兴趣的朋友可以加我微信，我拉你进群（群现在不能扫码进入了，只能靠人拉了）。不是我自吹啊，群里有很多藏龙卧虎的高手，很多专业问题都有更专业的人回答，我觉得我这个群主都要当不下去了。谢谢大家！（下面是半成品材料的贴图）

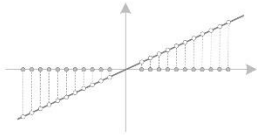
高等数学—连续—各类不同连续



利普希茨连续(Lipschitz)>赫尔德连续(Hölder)>绝对连续>一致连续>一般连续>几乎处处连续


几乎处处连续：一个函数 $f(x)$ 在定义域 $D$ 上被认为是几乎处处连续（almost everywhere continuous），如果存在一个测度为零的集合 $S \subset D$ ，使得 $f$ 在 $S$ 上不连续但在 $D \setminus S$ 上连续

$$g(x) = \begin{cases} 0, & x \in \mathbb{Q} \\ x, & x \notin \mathbb{Q} \end{cases}$$



淡蓝小点Bluedotdot (微信: bluedotdot.cn) 淡蓝小点技术系列：面向机器学习（深度学习、人工智能）的数字基 2024年8月12日 181 / 415

线性代数—矩阵分解：Cholesky分解



Cholesky分解是针对正定、半正定阵的，而协方差阵一定是正定（或半正定的），因此它的应用也非常广泛。在PRML的5.6节273页及11.1.1节528页都涉及到了Cholesky分解。

Cholesky分解是说：任意一个正定阵 $A_{n \times n}$ 都可以被分解为 $A = R^T R$ ，其中 $R$ 是一个 $n \times n$ 型的上三角阵。或者有时也写作 $A = L L^T$ ，其中 $L$ 是一个下三角阵。如果要求 $L$ （或 $R$ ）的对角元必须是正的，那么Cholesky分解是唯一的。

注意：虽然正定阵的定义不仅要求对任意 $\mathbf{x}$ 都有 $\mathbf{x}^T A \mathbf{x} > 0$ ，还要求 $A$ 是对称的。这是因为对称阵有很多很好的性质，我们想要重点研究。下面这个矩阵虽然满足 $\mathbf{x}^T A \mathbf{x} > 0$ 但它不是对称的，因而它不包含在通常所指的的正定阵范围内。

$$M = \begin{pmatrix} 2 & 0 \\ 2 & 2 \end{pmatrix}, \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} 2 & 0 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (x_1 + x_2)^2 + x_1^2 x_2^2$$

淡蓝小点Bluedotdot (微信: bluedotdot.cn) 淡蓝小点技术系列：面向机器学习（深度学习、人工智能）的数字基 2024年8月12日 75 / 415



特征函数（characteristic function）在形式上类似于矩母函数，但它是在复指数上定义的。设随机变量 $X$ 的特征函数为 $\phi_X(t)$ 为

$$\phi_X(t) = E[e^{itX}] = \sum_X e^{itX} p(X) \quad (\text{离散}), \quad \phi_X(t) = E[e^{itX}] = \int_{-\infty}^{+\infty} e^{itX} p(X) dX \quad (\text{连续})$$

特征函数和矩母函数主要异同点

- 独立变量之和的特征函数等于特征函数之积： $\phi_z(t) = E[e^{itZ}] = E[e^{it(x_1+x_2)}] = E[e^{itx_1}]E[e^{itx_2}] = \phi_{x_1}(t)\phi_{x_2}(t)$
- 任意一个概率分布特征函数总是存在，但矩母函数不一定总是存在<sup>209</sup>
- 概率分布的特征函数总是唯一的<sup>210</sup>
- 都可以通过求 $n$ 阶导得到分布的 $n$ 阶矩
- 矩母函数类似于密度函数的拉普拉斯变换，特征函数类似于密度函数的傅里叶变换<sup>211</sup>
- 特征函数在处理周期性、振荡性问题时更有效

<sup>209</sup> 因为 $e^{itX}$ 总是有界的而 $e^{tX}$ 可能是无界的

<sup>210</sup> 一般情况下我们可以认为若矩母函数存在也是唯一的，但详细讨论时还要分情况

<sup>211</sup> 离散型分布一样

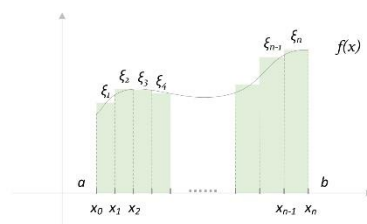
## 实变函数



实变函数：讨论以实数为变量的函数，但重点是将数学分析中的黎曼积分改造拓展为勒贝格积分<sup>126</sup>

- 测度：集合、测度、Lebesgue可测集等
- 积分：可测函数、Lebesgue积分等

机器学习主要用到测度的概念，它是现代概率论的基础



<sup>126</sup> Lebesgue integral, 凡是黎曼可积的一定勒贝格可积



现在还有一个很大的问题：基于概率公理化如何证明或推导出 $P(A) = \lim_{n \rightarrow \infty} \frac{N(A)}{N}$ 。这需要用到极限定理，主要包括大数定律和中心极限定理。

本系列极限定理主要介绍以下内容

- ① 随机变量序列的两种收敛方式：依概率收敛和依分布收敛
- ② 两个不等式：马尔可夫不等式、切比雪夫不等式
- ③ 三条弱大数定律：切比雪夫大数定律、辛钦大数定律、伯努利大数定律，一条强大数定律
- ④ 两条中心极限定理：林德伯格-列维中心极限定理、棣莫弗-拉普拉斯中心极限定理<sup>156</sup>

<sup>156</sup>Lindberg-Levi也称独立同分布中心极限定理，就是我们一般所说的中心极限定理；De Moivre-Laplace是局部中心极限定理