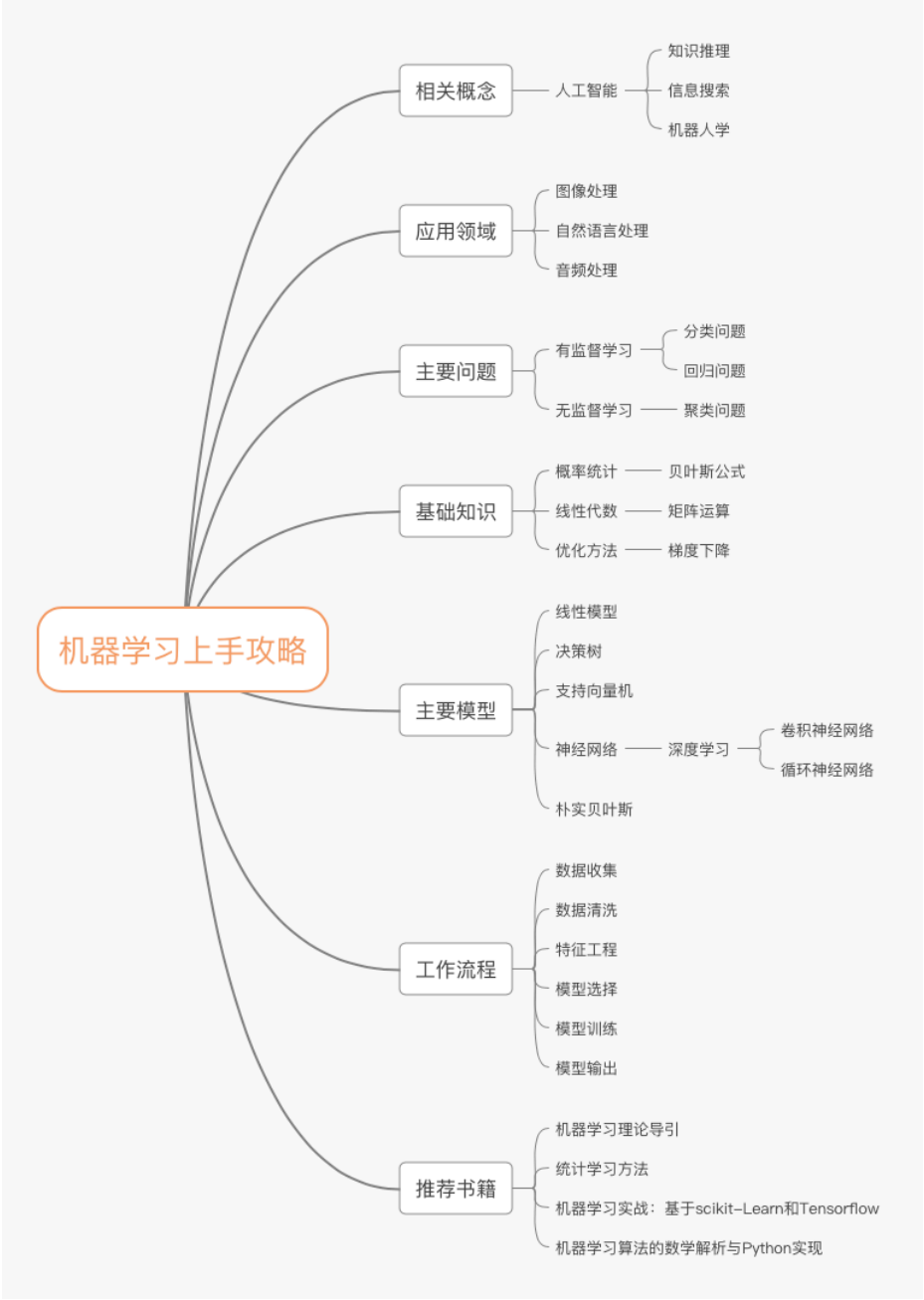
机器学习上手攻略



机器学习近几年大热，大家都想要了解，但机器学习已经形成一套枝叶繁茂的知识体系，而且往往建筑在复杂的数学基础之上，又容易让人无从下手。初学者最常问的，不是某个具体的重点难点知识，反而是机器学习究竟该怎样学。下面我会为大家梳理出机器学习的总体脉络，对于一些较为复杂抽象的概念，我习惯采用更形象容易理解的例子来进行解释，这也是我一贯“娱乐向解说”的风格，相信大家看完以后都能够对机器学习以及如何学习机器学

习，有一个较为整体和清晰的了解，而且过程也不会因为各种公式概念而看得太痛苦。

计划开始学习机器学习，建议首先搞清楚两个和机器学习经常一起出现的概念，分别是人工智能和深度学习。这三个词近几年曝光度非常高，有时候还出现混用的情况，三者究竟是什么关系呢？

是包含关系，人工智能>机器学习>深度学习。我们经常可以看到一幅图，上面画着三个大圆，最外面的圆是人工智能，中间小一点的是机器学习，里面最小的是深度学习。这张图比较直观第说明三者是包含关系，不过也容易产生一些误导，让人觉得机器学习是人工智能的核心，而深度学习又是机器学习的核心。

人工智能范围最大，机器学习是人工智能下面的一个热门研究方向，其它包括了知识推理、信息搜索和机器人学等等。要注意的是，如何实现“智能”是人工智能研究的目标，通往这个目标有很多条道路，到底那条路才是正道，现在还不知道，而机器学习只是其实一种很有希望能走通的道路，而其它的研究方向对于人工智能研究来说同样也很重要。

不同的研究方向，有自己侧重和擅长的应用领域。机器学习主要的应用领域有三大块，分别为图像处理、自然语言处理和音频处理。图像处理，包括了对图片以及视频的处理，而自然语言处理主要研究对文本的处理，而音频处理顾名思义，是研究对音频，譬如各种波形的处理。每一块子领域，又通常包含了识别和生成两大研究方向。可以看出，机器学习关注的都是我们人类日常接触最多的信息形式。

那机器学习怎样完成这些工作呢？通过问题定义。无论是图像处理、自然语言处理还是音频处理，机器学习都将这些具体问题根据不同的学习方法分为两个大类，分别是有监督学习和无监督学习。这里的监督，指的是模型训练时有没有给出参考答案。给出参考答案的，通常分为分类问题和回归问题。

分类问题顾名思义，该类问题要解决的就是如何将目标对象正确分类，也就是预测类别。这个过程和我们进行垃圾分类异曲同工。我们要做垃圾分类，首先得先培训，有一位老师告诉我们什么垃圾属于哪一类，譬如告诉我们说树叶属于厨余垃圾。机器学习同样有这个过程，在机器学习里面，把培训称为模型训练，培训的案例称为样本，而有监督学习的模型训练，要求每一个样本需要有一个“参考答案”，也就是你把树叶拿出来，还得告诉我这种属于厨余垃圾。培训结束之后，我们就可以自己出门去倒垃圾了，知道家里的垃圾都该扔哪个桶。模型也同样，训练完成之后，就可以输出分类结果，告诉你刚才输入的应该属于哪个类别。

回归问题是另一种有监督学习问题，不过它预测的是连续的数值，譬如气温、股价等等。它训练过程和分类问题差不多，也需要在模型训练的时候带上参考答案，训练完成后就能输

出预测结果，唯一不同的是，预测的结果是连续的数值，而不是类别。

除了有监督学习，另一种大类是无监督学习，也就是训练过程不再有参考答案了。这就容易让初学者，特别是刚刚学完有监督学习的初学者感到疑惑：没有参考答案那要怎么训练模型呢，就算瞎蒙也得给说一句蒙对了没呀？确实是这样，所以无监督学习一般用来解决没有参考答案的问题，最典型的就是聚类问题。

聚类问题要做的事分两步，首先把样本数据聚类成几个类，有一点搞小圈子的意思，这里的小圈子称为“簇”，然后就是不断判断新来的样本点应该属于哪个簇。聚类问题初学容易和分类问题搞混，其实很好分辨。前一段很喜欢谈论前浪后浪，那么如果问眼前这位满脸故事的老哥是属于前浪还是后浪，那这属于分类问题，但如果老哥就像那首老歌唱的是“我们不一样，每个人都有不同的际遇”，要你给他找一群意趣相投的老哥，那就是聚类问题。

说完了问题说模型。前面我们好几个地方提到“模型”，这里所说的模型不是拼装玩具，指的是数学模型。别急，我们大都有个共同的地方，就是听到“数学”就容易产生头晕、目眩、恶心反胃等生理症状。先忍着，我有良方。

首先，我可以很负责任地说，机器学习需要学习数学。那些告诉你机器学习可以不需要学数学而且点击就送的教程，不好说都是骗人的，至少可以肯定无法让你真正读懂机器学习的灵魂。同时，我可以很负责任地说，学机器学习不需要需很多数学。学机器学习，目的很重要，机器学习如果是一座金字塔，它的塔基就是数学，如果你要为金字塔添砖加瓦，譬如提出新的算法，或者改进现有算法，那你肯定得对数学十分了解。不过，如果更想要的是了解金字塔的结构，以及怎么解决你手头上让你睡得不香的问题，那你只需要了解数学这门“语言”。是的，平时我们描述一件事，会用汉语，或者用英语，而在机器学习这里，描述一件事用的是“数学语”，你要做的只是听懂它，或者找一个人来把它翻译成你能听懂的话。

那机器学习的数学基础该怎么学呢？经常有人问的一个问题是，要不要先学完数学再来学机器学习。这种想法的出发点我很赞赏，夯实基础才好筑高塔，不过，数学毕竟家大业大，子领域非常多，就算是职业数学家也只能了解自己领域的数学知识，而且，像机器学习这种大家一起开脑洞构建起来的学科，自然会横跨多个数学领域，如果真的要都学完然后才开始，恐怕就没有然后了。好在数学知识如果是大海汪洋，那机器学习用到的数学知识那只是弱水一瓢，我推荐按需学习法，大概了解机器学习都要用到哪些数学知识，有个整体印象，然后再根据需要深入学习。

知道了机器学习该怎么学，那又该学什么呢？大概分为三大块，概率统计、线性代数和优化方法。概率统计是机器学习的灵魂，机器学习的很多模型方法，追根溯源都是从统计学

中发展而来，可谓是思想的源泉，其中频繁用到贝叶斯公式的相关概念，可以重点学习。线性代数是机器学习的骨架，具体涉及到运算的，都离不开线性代数的知识。不过，线性代数是数学中很重要的一条学科分支，内容还是非常多，怎么办呢？那就重点学习矩阵运算相关的内容，线性代数在机器学习方面发挥作用的，主要就是矩阵运算。

以上两块都是数学大类下的具体子类，机器学习还有一块需要数学基础的内容，通常称为优化方法，这个领域属于机器学习私人定制的数学子领域，重点是研究一类数学问题，凸优化问题。凸优化名字听起挺吓人，让人觉得是好像是什么高深的研究放弃，其实要做的事非常简单，就是花样求极值，最终目的是实现 max 函数和 min 函数类似的效果。既然是花样求极值，花样自然挺不少，内容足够写一本满是公式很厚的书，不过，重点是梯度下降，搞懂什么是梯度，为什么下降就能达到求极值的效果，大的方向也就了然于胸了。

大家了解的数学背景知识，就可以着手学习机器学习的核心内容，也就是各种机器学习的模型。很多初学者首先会问机器学习究竟是什么，希望有一个简单直接的回答。最简单的回答是机器学习是算法，但不是一种算法，而是一群算法的总称。这些算法解决的问题是一样的，也就是前面说所的分类问题、回归问题等等，但用的思想不同，有一点八仙过海的意思。不同的思想催生了不同的模型，机器学习中主要的模型有线性模型、决策树、支持向量机、神经网络和朴素贝叶斯等等，从数学的角度来说，这些模型有的很简单，有的很复杂，但根据机器学习中知名的 NFL 定理，并非越简单的模型效果就一定差，越复杂的模型效果就一定好，也就是只有合适，没有好坏。

大家可能也注意到了，这些模型中有一款名叫“神经网络”的模型，在学习机器学习时，可以对这款模型加以特别关注。为什么呢，这款模型难道有什么特别？前面我们说，机器学习常和人工智能、深度学习一起出现，深度学习现在是个爆款名词，有人干脆把现在称为“深度学习”时代，如果此前不了解机器学习的发展，很容易让人觉得深度学习是近几年突然发展起来的新技术。其实不然，深度学习就是从机器学习下的一个分支，神经网络分支发展起来的。神经网络名字起的不错，不过身世比较坎坷，历经几起几落，早几年被支持向量机等模型按在地上摩擦，一直在坐冷板凳。这几年随着硬件技术大发展，神经网络也迎来了一波利好，在几个领域都取得了突破，声名大噪之后，不但知识体系在不断扩展丰富，譬如发展出擅长处理图像的卷积神经网络，和擅长处理音频文本等时序类型数据的循环神经网络，索性名字都换了，现在大家都管它叫深度学习。

接下来我们聊最后一个问题，也是非常多的人在学之前和学之后都爱问的一个问题：应该怎样用机器学习解决手上的任务呢？很简单，共分六步。首先是数据，想用机器学习解决

问题，数据处理很重要，有数据才能训练模型，有好的数据才能训练好的模型。在进行一次机器学习任务，最耗时间的往往不是处理模型，而是处理数据。数据要怎么处理呢？分三步，收集数据，清洗数据，和特征工程，也就是先把原始数据收集起来，然后把各种干扰项、杂质去掉，最后更有价值的特征提取出来。数据准备好以后，接下来的工作就是把数据喂给模型，这个过程也分三步，首先选择最合适的模型，然后使用数据进行训练，想法设法不断提升模型预测效果。随着模型训练的结束，机器学习的前期准备工作也就告一段落了，最后的工作就是把模型部署到生产环境，让模型发挥作用。

上面我们介绍了机器学习的学习路径，也把相关的主要知识快速串了一遍，让大家在脑海里勾勒出机器学习的整体印象。当然我也知道，机器学习所包含的内容十分丰富，大家的疑问也一定不止这些，只用一篇文章是很难把全部内容都讲清楚的。想要进一步深入了解，我推荐四本书。

第一本是《机器学习理论导引》，这是西瓜书的作者周志华的最新作，可以作为机器学习的入门导引。

第二本是《统计学习方法》，深入介绍了机器学习模型背后的数学原理，翻开每一页满满的都是数学符号，覆盖全面量又足，适合进阶阅读。第三本是《机器学习实战：基于 scikit-learn 和 Tensorflow》，机器学习最重要的毕竟还是解决实际问题，这本书就是介绍如何通过 Python 来编程使用机器学习，值得一提的是这本书里都是各种代码，有程序员那味儿了。

最后一本厚着脸皮推荐自己写的这本《机器学习算法的数学解析与 Python 实现》。我最开始学机器学习时，发现机器学习的书都有一个问题，我想看懂一个概念，往往就会又牵扯三四个陌生的概念，而这三四个陌生概念背后又有概念，这样层层嵌套下去的结果就是，我为了搞懂一个很简单的问题，却非得需要花费好多额外的时间学懂这些概念都是什么意思。我只是想了解机器学习算法的基本原理，其实不太关心背后究竟还有几串数学名词。那有没有一本干脆利索，用几句简洁直接的话就能把意思说清楚的书呢？我找了很久，都没有找到，而且我发现，学机器学习是既要学数学理论，又要学编程应用，但市面上机器学习的书要么太偏数学，要么太偏编程，看完以后总感觉理论和应用是脱节的。所以，我自己写了一本，希望和大家一起俯览机器学习的世界，而且阅读起来心情还比较愉悦。我也开了一个公众号，分享自己对大家关注的一些机器学习问题的心得，叫“睡前机器学习”，欢迎大家来玩。