机器学习/深度学习书单推荐及学习方法

写在前面

本人是个对数学和人工智能极其感兴趣的人。平时，我也在线上线下经常与国内外的朋友讨论人工智能的各种方面，无论是技术方面还是哲学方面。我帮助过很多实习生和网上的学生，带领他们从入门一步步过渡到足够从事数据挖掘工作。在此期间，我发现了一件很有趣的事情，所有技术好的数据分析/挖掘工作者，都是喜欢”主动学习”的人。

这次在GitChat上发起Chat，就是想帮助更多喜欢数据科学、喜欢主动学习的人能够少走弯路。这个Chat中主要讨论的是如何入门学习机器学习/深度学习的理论知识、如何着手数据挖掘项目以及从事数据挖掘相关岗位所需要的能力。

正文

不论你是学生、想转行数据岗位的在职程序员，都需要自学达成目标，我本身就是一个苦逼自学者例子。

我的本硕都是计算机专业，由于本科搞的是算法编程，硕士开始搞机器学习方向本来也比较适合，但机器学习算法真的让我头疼了很久。在慢慢攻克了机器学习算法，并学到一定程度后，我发现真正的数据挖掘绝不仅仅是会机器学习算法就够了，还要学习很多东西，比如数据清洗等等技巧。作为一个”过来人”，我明白大家在学习中会遇到的困惑。我此次，就是来帮助大家解除困惑。

关于机器学习，非数学/统计专业的人都会有这些疑问：

1. 到底要把数学学到什么程度才能够无障碍地推导机器学习算法？实变、复变、泛

从最基础的来说，微积分、线性代数与概率论是学习机器学习的必会内容。相信大家看到这个答案会很失望，但是放心，这绝对不像你想象的那么难。在本科期间，我们学习数学的目的其实是为了期末考试，需要做很多习题。而在机器学习中，使用到的仅仅是这些他们的特性，而不需要用他们来解题，所以只要知道他们的定义就可以。

比如线性代数，我们仅仅需要了解向量、矩阵和逆矩阵等等的定义，而不需要去真正的计算逆矩阵。

这说明，我们不需要花费大量时间去学习数学，而只需要了解并记住他们的定义。

那么，看什么书好呢？我个人给大家推荐的是《金榜图书》的考研数学讲义系列的《高等数学辅导讲义》、《线性代数辅导讲义》和《概率论与数理统计辅导讲义》。这三本书算是考研数学入门级，他们的优点是简洁、全面，把教科书中几百页的内容压缩到几十页，如果不看其中的例题，几天就能看完。

这三本书外，还有很多数学内容需要学习，比如Jacobian矩阵、张量、特征分解、奇异值分解（SVD）和Moore-Penrose伪逆等等……这些是实分析、复分析、矩阵论等书中的内容，也是机器学习的必会内容。

看到这里你可能会开骂了，我要是会这些，干嘛还看你写的东西！

放心，我就是来给你解决这些问题的。之前我说过，推导机器学习算法的过程中，需要的是了解数学定义，而非系统地学习每门课。然而想要通过看书学习某一个知识点是很难的，因为需要很多先修知识，否则根本看不懂。这里，给大家推荐一个非常良心的免费在线课程可汗学院（www.khanacademy.org，需要科学上网），这个网站中包含很多学科，其中的数学部分几乎把所有机器学习中所需要的数学知识都涵盖到了，而且每个知识点的讲解都是独立的视频，每段视频大约只有几分钟，还有配套的在真实应用中的例子。比如向量微积分中的Jacobian矩阵这个知识点，可汗学院将其分为5个短视频来讲解（1. Jacobian矩阵的先修知识;2. 多变量函数的局部线性法;3. Jacobian矩阵;4. Jacobian矩阵的计算;Jacobian矩阵的决定子应用），这5个短视频从最基础的先修知识慢慢过渡到最难的部分，每个短视频只有3-8分钟，已经足够让我们理解Jacobian矩阵了。

入门机器学习到底要看什么书？

容，比如Break Point、VC维、误差衡量、线性回归、非线性转换和梯度下降，哪一个是不重要的？

当你看完”机器学习基石”，我建议你继续学习他的”机器学习技法”。我相信你刚开始学习”技法”的时候，会发现他比上部更加无聊，于是忍不住去看对应的书籍，想要避免看这种无聊的视频。然而，你最后一定还会回来继续看他的视频，因为你会发现他的视频虽然巨无聊，但讲的真的很细致，每一步推导都讲的很明白！

在学习”技法”课程的时候，你可以同步的看书了，因为光看视频，很多东西会忘记，要不断的看书复习。这里，我推荐给你三本书：《统计学习方法》by李航、《机器学习》by周志华、《机器学习》by Mitchell。这三本书都是机器学习界入门的经典书籍，我之所以同时推荐三本，不是让你做三选一的选择题，而是把这三本对照着看：

- 《统计学习方法》对公式的推理深入；
- Mitchell的书重在算法思路的讲解，对公式的推理很浅显，但易懂；
- 周志华的书内容更加广泛且包含很多两本书中没有的内容。

建议学习的时候以李航的书为基础，与视频对照着看；使用周志华的书进行补充；当无法理解某处的时候看Mitchell的书。

除了机器学习，真正的工作中还哪些必要技巧？

真正的工作和项目中，只会机器学习是不够的。最早的就是对数据进行清洗，数据清洗工作和机器学习算法的选择同样重要。数据清洗注意需要掌握两点：数据处理与正则表达式。这里推荐三本书：《Python数据处理》、《数据科学实战手册（R+Python）》与《正则表达式经典实例》，前两本讲的是数据清洗处理，最后一本讲的是正则表达式的语法。这三本书讲的是通用技巧，在具体工作中所需要的方法是不同的，不要把自己禁锢与书本中，要在具体的工作中自己思考。

如何入门深度学习？

目前，深度学习并非工作的必备技能，仅仅是加分项。所以我建议大家在掌握了一定的机器学习知识后，再开始学习深度学习的内容。

如何着手开始进行数据挖掘项目？

着手数据挖掘项目，首先要选择一门合适的语言。数据挖掘可以用很多语言完成，R、Python、Java等等都可以，但我个人建议大家使用Python，因为使用Python相对简单且大多数公司都要求使用Python进行工作。 刚开始进行数据挖掘项目的时候，很多程序员出身的人会陷入一个误区，认为既然是做项目就要把算法的每一个细节都自己实现，而是不使用现成的工具包，担心自己会变成调包侠。其实，每个高手都是从调包侠开始的，而且自己写的代码都是使用Python完成的，效率与工具包中直接调用C语言的代码相比要差很多。而且，在入门阶段，不应该花费极多的时间只为了对算法造轮子。在后期，当你的能力足够而进行一些非常复杂的项目的时候，才会觉得工具包满足不了你，那时候再顺其自然的造轮子岂不甚好。

说到Python的机器学习工具包，不能不提到scikit-learn。scikit-learn的算法非常齐全(几乎把所有你能想到的机器学习算法都包括在内，甚至数据预处理、特征提取等等都有现成的工具)，调用简单(两条语句就能训练出一个模型)，API非常友善(可以到官网)。学习scikit-learn最好的一本书是《机器学习系统设计》by里彻特。这本书共12章，除去最后一章外，每章都带着我们从最基础一步步地做出一个数据挖掘项目，其实把这本书看完，其实已经有最基础的数据挖掘项目能力了。

除了scikit-learn外，还有一个深度学习框架也非常好，叫做Keras。Keras的底层可以用TensorFlow或者Theano，在理解神经网络的情况下，学习Keras只需要花费极短的时间就可以上手，这里建议大家学习一个免费视频课程：莫烦Python(https://morvanzhou.github.io/tutorials/machine-learning/keras/)，这个网站是一个免费的机器学习视频课程网站，站主莫烦大神用最简洁的方式会使你在不到一个小时的时间内就掌握Keras的语法。

以上是所用到的工具，学完《机器学习系统设计》已经有基础的实战能力，但其中的一些流程该书并没有系统地介绍，如果想要更加系统的学习，推荐大家阅读《数据挖掘导论》，这本书包括分类、关联分析、聚类和异常检测的项目流程实例，还提供数据集和幻灯片，非常适合系统的学习。

在以上的事情都做完后，大家可以上Kaggle进行小的比赛，只要足够努力，拿到一些小比赛的Top5%还是有机会的。

3. 对算法的学习要掌握所谓的”分寸”，相对重要的部分多学，不太主流的部分稍稍了解即可。
4. 英语很重要，你会越来越发觉英语的重要性，请坚持每天学习一点英语。
5. 学习机器学习算法要真正学懂，某些小的地方学不懂便跳过，最终会造成很多漏洞，而这漏洞，填补起来要花费很大的精力。

祝您学习的过程是愉快的，前路虽艰，行则心安。