扎实入门机器学习的路子

原创 包包闭关修炼 包包算法笔记 2021-12-30 08:58

收录于话题
#算法工程师 24 #机器学习 11

继续出差酒店原创日更，今天是一点经验分享，包大人当初用这个路子入门机器学习的。

前言

回顾了下我当初入门机器学习方式，发现还是有些特点的。**因为是科班出身，所以是那种稳扎稳打，自顶向下，理论结合实践的方式，比较注重用代码实现去理解原理。**代码有个好处就是不会骗人，光看理论有种我明白了，但是很模糊，如果结合代码就非常清晰了。**用两个字说就是“扎实”。**

主要分为原理入门，编程理解，实战应用，三个步骤。其中非常强调过程中的正反馈，和优质的资源。正反馈是长久坚持的动力来源。优质资料是高效正确的保证。

比较反对下面这样囫囵吞枣，良莠不齐，就想着21天从入门到精通的方式。

第一、反对上来就给你推荐python，sklearn，pandas之类的。这些东西往往把细节都藏起来了，容易变成调包侠。

第二、 不推荐任何国内的非知名大学的视频教程。大部分的东西不是抄就是质量太差，有吴恩达ɑ的公开课不看去买网课，留着钱买个大鸡腿吃不好吗。

第三、不推荐任何上来就是各种花里胡哨从开发到部署的实战项目，抓住人就想心急吃热豆腐的心理贪图你的钱包。

第一阶段：原理入门

目标是搞清楚机器学习的基本概念和基本的算法原理。这个阶段的**正反馈**来自于新知识的获取，原理的理解。不要好高骛远。下面介绍几种入门方法，分别对应看书入门党，看视频入门党，有一些精选的优质资源推荐给初学者。**推荐的资源**，周志华ɑ《机器学习》，李航《统计学习方法ɑ》 Peter Harrington 《机器学习实战ɑ》吴恩达 Coursra 机器学习公开课

首先看书入门党，周志华和李航老师的西瓜书和统计学习方法都可以，可以快速地看完前几章，不要具体到算法，如果你愿意，看完逻辑回归就可以了，首先明白机器学习问题的定义，其次明白几个关键的名词，训练验证测试，偏差方差ɑ，样本，特征，标签。然后去看什么是监督学习什么是无监督学习，大概了解了这些之后，再到具体的算法。再推荐一本书《机器学习实战》绿皮书ɑ，这些书的特点就是原理讲的很明白，《机器学习实战》所有的算法都用代码实现了一遍，逻辑清晰很好理解，比那些用sklaern的书强一万倍。

看视频入门党，推荐吴恩达 Coursera 上的《机器学习》，吴恩达老师设计的课程已经非常适合入门了，侧重原理，逻辑清楚，机器学习的细节也面面俱到。

经过上述阶段，你大概对机器学习要解决的问题，使用的方法和适用场景都有所了解了，这时候，你大概对算法的原理也都八九不离十，但是学习原理总是枯燥的，不过一定要坚持下来，千万不要在这个阶段满足于调包。

第二阶段：在编程中理解

目标是能够自己动手实现算法的细节而不是用sklearn去调包。**正反馈**是自己动手从头正确实现机器学习算法。 **推荐资源** Peter Harrington 《机器学习实战》吴恩达 Coursra 机器学习公开课编程作业。

这里推荐吴恩达老师机器学习课程的作业，不需要把每个算法都实现一遍，但是要在实践中去理解机器学习的基本算法套路，比如梯度下降是怎么做的，链式法则ɑ怎么用程序表达。还有就是《机器学习实战》的配套代码，这本书的最大好处是让你能够用最基本的python语法，从底层上让你构建代码，实现我们常说的比如邮件过滤，数据分类的应用。

很多时候你要写最基本的代码和结构去做这些工作，而不是像sklearn去调用fit 和predit，你能实现算法的底层原理，知道决策树ɑ的分割增益计算如何写代码，梯度下降如何写代码，知道机器学习是如何从0到1实现的。

不过这本书比较老旧了，重点也不是讲解理论方面的东西，可以当成第二个阶段的教材，和第一阶段互补。

另外。如果你是NLP方向的同学，可以看一下词向量GloVe的代码实现，为什么推荐GloVe的代码，他是一个用纯c语言写的机器学习做矩阵分解来求解词向量的程序，包含实现随机梯度下降，损失函数定义，数据并行处理等基本的要素，是麻雀虽小，五脏俱全，代码逻辑清晰，涉及到机器学习的方方面面，而且，毫无调包，代码量不大，很容易看懂。

第三阶段：实战应用

目标是把机器学习应用到实际问题中，加深对算法的理解。**正反馈**来自于使用机器学习工具来解决实际问题。**推荐资源**Kaggle。

这时候，你对机器学习的原理，实现都有了解了，但是机器学习毕竟是一门应用的科学，我们通过在实战中学习机器学习。所以这个阶段非常适合打比赛。这里比较推荐Kaggle平台，不推荐国内的竞赛平台，除非你想给自己添堵，被排行榜上各种骚操作吓呆。至于怎么玩kaggle，推荐kaggle kernel上的开源讨论，以及一些比较好的Grand Master的分享。

后话

上面三个阶段，在具体的知识点上可以互相交叉。比如，你看完了逻辑回归，动手实现了一下，然后上kaggle做了一个数据集的任务。不是说非得把所有的长篇大论看完了，这样能更有利于你学习。

如果你有什么特别好的资源推荐，或者入门方法，欢迎放到评论区 ~

历史精彩文章：
【段子】如何激怒一位算法工程师
【段子】让算法工程师破防的瞬间

【技术】可能是全网写特征工程最通透的...
【技术】一文串起从NLP到CV 预训练技术和范式演进
【技术】工业界文本分类避坑指南
【技术】Kaggle进阶：显著提分trick之指标优化
【技术】从一道数学题面试题到GBDT原理的推导
【技术】所有数据集上给神经网络刷分的通用方法

【闲谈】工作后顶会重要吗?投入精力,结果...
【闲谈】回看互联网十年校招薪资变化，我发现...
【经验】面试官带你破解算法岗诸神黄昏，神挡杀神！
【经验】在读和转行进大厂做算法工程师的捷径
【闲谈】从Zillow用AI指导买房投资血亏说起
【闲谈】如何看顶会论文上关于泄露的乌龙
【经验】算法工程师的术与道：从特征工程谈数据敏感性

对机器学习进行深入理解的一本好书

ChallengeHub

LunchBox 机器学习 K-MEANS聚类算法图像分割案例

Rhino建筑

致初学者的深度学习入门系列（二）—— 卷积神经网络CNN基础

技术开发小圈