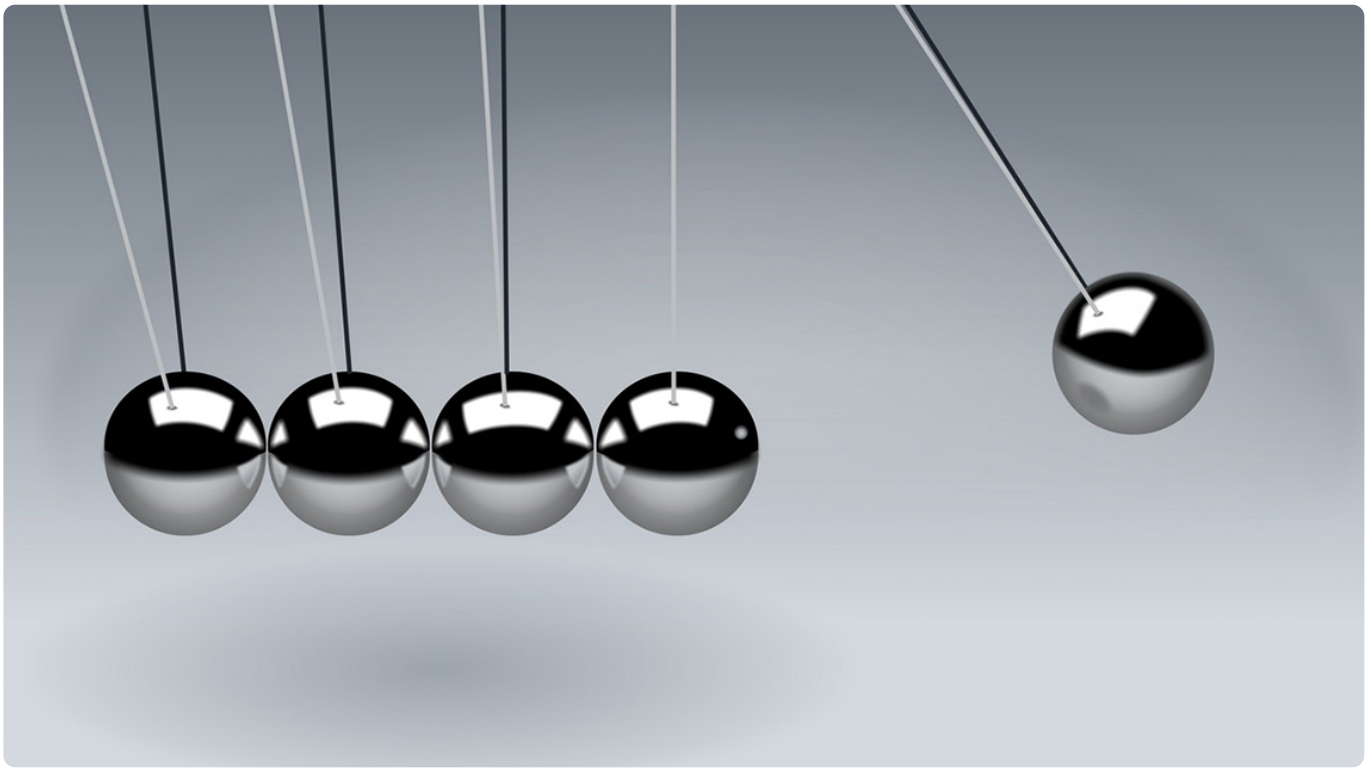
005 | 数据科学家基础能力之系统

2017-10-13 洪亮劼

AI技术内参 进入课程 >



讲述：初明明
时长 09:34 大小 4.39M

对于初学人工智能的工程师或者数据科学家来说，在知识积累的过程中，“系统”往往是一个很容易被忽视的环节。特别是非计算机科学专业出身的朋友，一般都没有真正地建立过“系统”的概念，在今后从事人工智能的相关工作时，很可能会遇到一些障碍。

今天我想给你分享一下，作为人工智能工程师和数据科学家，需要建立的关于“系统”的最基本认知。这些认知能够帮助你把书本的理论知识和现实的应用场景快速结合起来。

理解管道（Pipeline）

在很多人工智能初学者的认知中，机器学习的流程是这样的。有一个已经准备好的数据集，这个数据集里面已经有了各种特征以及所对应的标签或者响应变量。这个时候，你需要做的

就是利用这个数据集和一些现成的机器学习工具包，来训练一些机器学习模型。模型训练好以后，就可以计算一些已知的评估指标了，比如准确度、精度等等。

这是一般教科书和课程上介绍的标准机器学习流程，也是很多机器学习论文中的实验环境。遗憾的是，这个静态的流程并不适用于工业级的数据产品。

要支持工业级的人工智能产品，一个最基本的概念就是，**你需要搭建一个管道让你的环境是动态的、闭环的**。在英文的语言背景里，“管道”这个词很形象地说明了这个环境的特点。我们把数据想象成“管道”里的水，这里面的一个核心思想，就是数据从一个环节到下一个环节，源源不断。我们再把最终的产品，也就是这个管道的末端，和最开始的数据采集部分，也就是这个管道的开始端，结合起来思考，这就是一个闭合的环路。

理解数据产品的核心，就要理解它是一个闭合环路。几乎关于数据产品的一切难点、问题以及解决方案都可以从这个闭合环路中产生。从一个静态的机器学习流程到一个动态的管道似的闭合环路，这是一个质变，对整个环节上的所有步骤都有全新的要求。

我这里就拿数据集来举个例子。静态的流程中，我们不需要太过关注这个数据集的来源。甚至采集数据集的代码或者脚本都可以是一次性的，可以不具备重复使用的价值。但是这种情况在管道的环境中是不可能的。

在管道中，**采集数据的可靠性和可重复性是非常重要的步骤**，这就对采集数据所采用的代码有不一样的要求。这部分代码需要被反复检验，每一步都需要人工智能工程师和数据科学家进行检验。如果我们把这个例子扩展到数据管道的其他部分，就可以很清楚地看到，数据管道对于构建一个机器学习流程所带来的根本变化。

管道的另外一个重要特性是**自动化**，一个不能自动化的管道是不能被称为管道的。这里的自动化有两层意思，一层意思是指数据本身可以被自动采集、整理、分析，然后自动流入机器学习部分，有结果后自动输出并能被线上的系统使用；另一层意思是指，每一个环节本身都不需要人工干预，或者仅需极少数的人工，自身可以高可靠地运行。由此可见，管道的自动化对每个环节的技术选择和实现都有非常高的要求。

现代互联网公司中，每个团队，甚至成立专门的团队，一般都会针对机器学习管道开发工具平台，使管道的灵活度、自动化、可靠性有足够保障。对于初学者而言，尝试从管道的角度去理解问题，从整个系统的角度来认识产品开发过程，认识机器学习流程，才有可能设计出能真正满足线上需求的技术方案。

理解线上和线下的区别

了解了一个数据系统的闭合回路以后，很自然地，就会出现下一个问题，这也是一个核心的系统级问题，在这个管道中，哪些部分是在“线上”，哪些部分又在“线下”呢？

这里我们先来理清“线上”这个概念。“线上”往往是说，对于交互性很强的互联网产品（包括电商、搜索引擎、社交媒体等），从用户来到某一个页面，到我们为这个页面准备好所需内容（例如推荐的商品或者搜索的结果），这中间的响应时间对应的就是“线上”，这部分时间非常短暂，往往只有几百毫秒。如何在这几百毫秒的时间内进行复杂的运算就非常有讲究了。

“线下”的概念是相对于“线上”而言的。通常情况下，不能在这几百毫秒之内完成的运算，都是某种程度的“线下”运算。

理解线上和线下的区别是初学者迈向工业级应用的又一个重要的步骤。哪些计算可以放到线上，哪些可以放到线下，就成了种种机器学习架构的核心区别。

初学者还需要注意的一个问题是，线上和线下都是相对的概念。今天放在线下计算的某些部分，明天可能就会放到线上进行计算。所以，慢慢学会把握两者之间的转换之道，对于初学者进阶至关重要。

我这里举一个简单的线上和线下分割的例子。比方说，我们要构建一个检测垃圾邮件的系统。对于这样一个系统而言，哪些部分是线上，哪些部分是线下呢？

初看，我们在这里讨论的是一个比较容易的架构，但并不代表实现这个架构的难度也很小。在最简单的情况下，检测垃圾邮件需要一个二分分类器。如何训练这个分类器的参数就是一个关键。

假设我们训练一个逻辑回归二分分类器。那么，逻辑回归的参数，也就是一组线性系数究竟应该在什么环境中得到呢？很明显，训练一个逻辑回归肯定需要大量的训练数据。在有一定量（大于几千的垃圾邮件和非垃圾邮件）的训练数据时，训练逻辑回归的参数就不可能在几百毫秒内完成。在这样的思路下，训练逻辑回归就不得不放到线下来计算。一旦这个决定做出以后，一系列的模块就都必须放在线下计算了。

另外，数据的收集肯定也得放到线下，这样才能保证可以把训练数据传输到后面的管道模块中。还有特征的生成，至少是训练数据特征的生成，很自然地也就需要放在线下。

训练逻辑回归本身，刚才我们提到了，需要放在线下。而放在线下这个决定（从某种意义上来说，无所谓时间多了一点还是少了一点，总之无法满足几百毫秒的线上计算就需要放在线下），又可以让训练逻辑回归本身，采用更加复杂的二阶算法，使参数能够得到更好的收敛。

你可以看到，因为一个决定，带来了关于整个管道的一系列决定。而这些决定又影响了模型算法的选择，比如可以选用相对耗时的更复杂的算法。

那么在这个架构下，线上的部分是什么呢？首先，训练完一个模型之后，要想使用这个模型，我们必须把模型的参数存放到某个地方（也许是一个数据库或者是一个存储系统），线上的系统可以在几百毫秒的时间内马上得到这些参数。仅仅得到参数是不够的，还需要对当前的邮件进行判断。

这一步就有一些问题了。一种选择是，线上的部分拿到模型参数，然后实时动态产生这个邮件的特征，再实时计算出一个分数，并且判断是否是垃圾邮件。整个过程的这三个步骤需要在几百毫秒内进行完毕。

实际上，这里面的第二步往往比较耗时，甚至有的特征并不能在线上进行计算。比如，也许有一个特征需要查询这个邮件的来源是否可靠，这里就可能需要数据库操作，这一步也许就会非常耗时（在几百毫秒的场景中而言）。因此，动态产生特征，除非特征都非常简单，很有可能并不能完全在线上完成。

我们可以对框架进行简单的修改。所有的邮件首先输送到一个特征产生的模块中，这里并不是一个完全线上的环境，运算的需求可能超过几百毫秒，但总体只是几秒，最多十多秒。所有的特征产生以后，对邮件的判断也在这里完成，最终将邮件是否是垃圾邮件这个简单的选项保存下来。在线上的系统，也就是用户来到这个邮件系统界面的时候，我们只是从保存的结果中，直接读出一个标签，速度非常快。

如上，我们通过检测垃圾邮件系统的例子，分析了线上和线下的分割情况。现在来做一个思考，刚才描述的这个架构有什么问题吗？问题就是，线上的结果是一个事先计算好的结果，模型本身也是事先计算好的。因此，当有大量突发数据（比如一大批新的垃圾邮件）来临的时候，这个架构可能无法很快反应，更新模型。可见，如何理解线上和线下是一个需要慢慢琢磨的学习过程。

小结

今天我为你讲了数据科学家和人工智能工程师需要掌握的关于系统基础的两个核心概念。一起来回顾下要点：第一，现代数据流程不是一个静态的数据集，而是一个动态的闭环管道。第二，理解什么计算可以放到线上，什么计算可以放到线下至关重要。

最后，给你留一个思考题，如果让你设计一个商品的推荐系统，哪些部分放到线下，哪些部分放到线上呢？

欢迎你给我留言，和我一起讨论。

上一篇　004 | 数据科学家基础能力之机器学习

下一篇　006 | Google的点击率系统模型

精选留言 (8)

写留言

RZ_diversi... 置顶

2017-10-19

4

线上：对于每一个用户的搜索历史记录，商品历史点击率等数据保存在线上数据库中。这些数据要定期输入到线下已经训练好的模型中来对参数进行更新。线下：推荐模型的训练，评估过程。这种思路其实和文中提到的垃圾邮件系统解决方案是类似的。

这里面存在的问题是，这种线上线下分割方法能够确保这个系统是一个pipeline吗？是否...

展开

---

udisyue 置顶

2017-10-21

3

我们先考虑一个商品推荐模型所需要的数据有哪些。对于商品推荐来说，它的数据来源应该有两种，一是用户的搜索记录，二是用户的购买记录。搜索记录的数据表示的是用户想买什么，而购买记录表示的是用户的购买结果。推荐商品的时候，如果用户已经购买过了商品，那么即使我们的模型非常完美预测了用户的购买意向，也并不能产生任何价值。而只有那些进行了搜索并且没有购买的用户才更值得作为数据进行采集。而用户购买了商...

展开

Momo

2017-11-03

29

恰好是做推荐系统的，来回答一下课后题，线上部分只有实时的召回（比如用户触发了某个关键词）和排序的预测过程（用一些简单的实时特征，再加一些提前准备好的只需要查询的离线特征），而其他部分，比如召回集合的规则、离线召回、排序模型的训练等等都是线下。顺便回应一下「小凯」的提问，这么接地气的内容，不是工业界的人根本写不出来，恐怕在书本里是找不到的。

展开

作者回复: 非常好的回答。

吴文敏

2017-10-19

2

最简单的方式线上从Aerospike这类数据库中读取用户的推荐结果，其余全部放到线下

小凯

2017-10-19

1

关于机器学习的“系统” “管道”概念，理论，应用有没有相关参考书，或者参考文献？

杯莫停

2018-08-09

能用时间解，就少引入状态，时间解不了，再考虑保存状态，状态的维护和迭代成本都很高。一些伪状态：各种模型参数，可以看作内容效果向用户体验的妥协，能小则小。

展开

suke

2017-12-02

请问 您说的管道的最终产品 和管道的源头 是如何联系在一起形成闭环的 能再详细解说一下么

作者回复: 主要是需要依靠实验和测量不断对产品的创新进行推进。

范深

2017-10-21

在线和离线特征与打分预测模块放到线上；
数据收集、清洗与模型训练和评估放到线下。