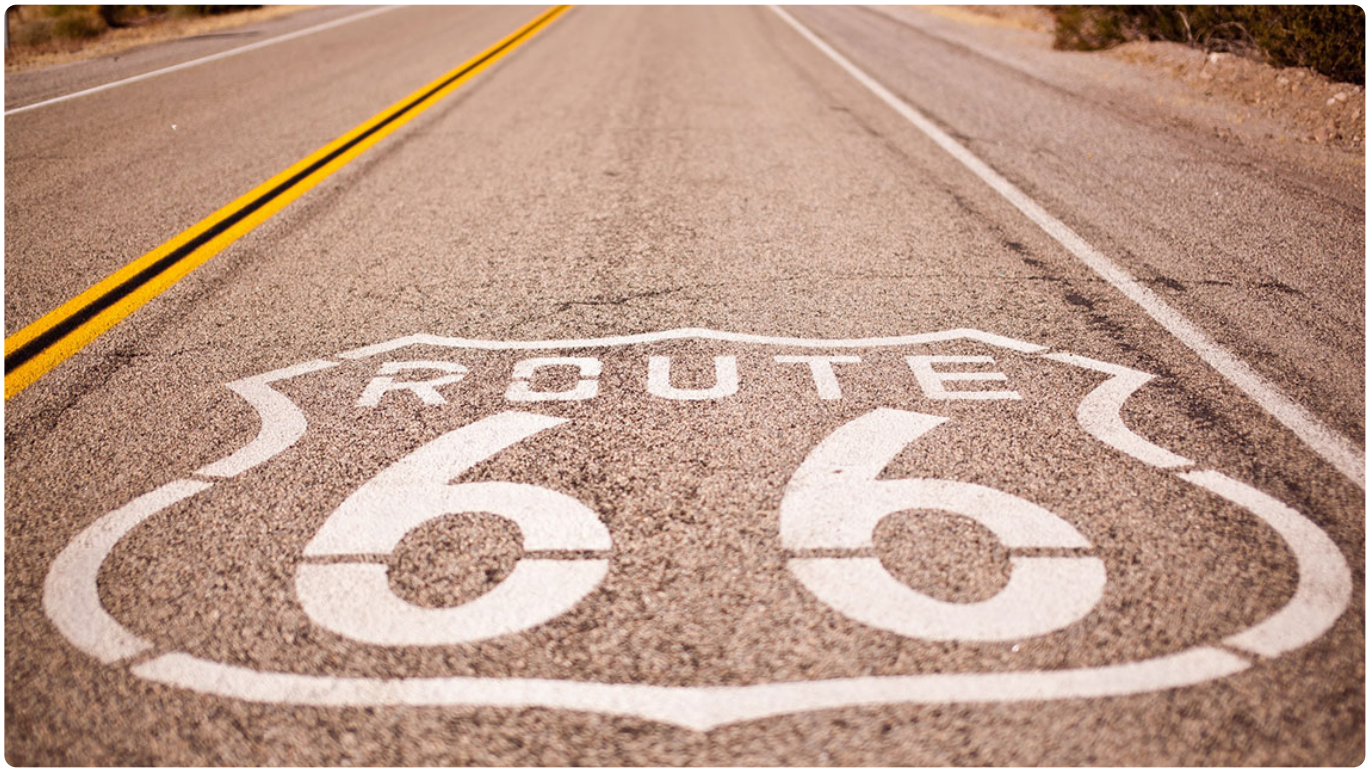
046 | 职场话题：数据科学家应聘要具备哪些能力?

2018-01-17 洪亮劼

AI技术内参 进入课程 >



讲述：初明明
时长 09:13 大小 4.22M

周一，我们探讨了在公司内部，数据科学家和产品团队的其他职能人员在协作中都会遇到哪些问题，以及如何看待数据科学家或者人工智能工程师所做的算法性工作在一个产品发展中的位置。

那么，今天我们稍微换一个方向，来讨论数据科学家和算法工程师在应聘方面的问题。一起来看看，作为数据科学家，在面试一家公司时，究竟应该怎么准备，有哪些信息是需要了解的。

希望今天的内容对正在思考进入这个行业的年轻学者、工程师有所帮助，从大的方向上为你的应聘提供一些可借鉴的内容。

数据科学家应聘的“硬”实力

对于数据科学家或者人工智能工程师来说，最核心的竞争力无疑是他们对人工智能、机器学习等技术的知识积累以及融会贯通的能力。

我们之前的一系列分享中已经提到了这些“硬”实力的大范畴，这里我做一个简单的归纳。

首先，我们需要理解和掌握一些机器学习的基本概念和理论。

第一个重点无疑就是监督学习。

什么是监督学习呢？监督学习就是指我们通过外部的响应变量（Response Variable）来指导模型学习我们关心的任务从而达到我们需要的目的这一过程。**监督学习中需要彻底掌握三个最基础的模型**，包括**线性回归**（Linear Regression）、**对数几率回归**（Logistic Regression）和**决策树**（Decision Trees）。

怎么理解我说的“彻底掌握”呢？这里的彻底掌握有三层含义。

第一，需要了解这些模型的数学含义，能够理解这些模型的假设和解法。比如，线性回归或者对数几率回归的目标函数是什么；写好了目标函数之后，如何求解最优解的过程。对于这些核心模型，必须能够做到完全没有差错地理解。

第二，需要了解什么场景下使用这些模型是最合适的，以及怎样把一个实际问题转化成为这些模型的应用，如果不能直接转换还有什么差距。

第三，能不能写实际的代码或者伪代码来描述这些模型的算法，真正达到对这些算法的掌握。

监督学习当然不限于这三个算法，但是这三个算法是绝大多数机器学习任务在工业界应用的起点，也是学习其他算法模型的支点，可以按照这个思路去了解更多的算法。在面试中，能够对这些基本算法的理解有扎实的基本功，这一点很重要。

了解机器学习的第二个重点就是无监督学习。

无监督学习并没有明显的响应变量，其核心往往是希望发现数据内部潜在的结构和规律，从而为我们进行下一步决断提供参考。

从面试角度来说，**“K 均值算法”往往是考察数据科学家整个无监督学习能力的一个核心点。**因此，对于这个算法有必要认真学习，做到真正的、彻底的理解。

怎么学习呢？和前面我们提到的监督学习一样，也需要从编程实现和算法本身两个方面入手对 K 均值进行把握。在掌握了 K 均值之后，还可以进一步去了解一些**基于概率模型的聚类方法**，扩宽视野，比如“**高斯混合模型**”（Gaussian Mixture Model）。

其次，虽然机器学习和统计学习有不少的重合部分，但是对于合格的数据科学家和人工智能工程师来说，一些机器学习方向不太容易覆盖到的统计题目也是需要掌握的。

第一，我们必须去理解和掌握一些核心的概率分布，包括离散分布和连续分布。这里的重点不仅仅是能够理解概念，而且是能够使用这些概率分布去描述一个真实的场景，并且能够去对这个场景进行抽象建模。

第二，那就是要理解假设检验。这往往是被数据科学家和算法工程师彻底遗忘的一个内容。我们要熟悉假设检验的基本设定和它们背后的假设，清楚这些假设在什么情况下可以使用，如果假设被违背了的话，又需要做哪些工作去弥补。

第三，那就是去学习和理解因果推断（Casual Inference）。这虽然不是经典的统计内容，但是近年来受到越来越多的关注。很多学者和工程师正在利用因果推断来研究机器学习模型所得结果的原因。

再次，还有一个很重要的“硬”技能，就是要对系统有一个基本了解。

第一，就是具备最基本的编程能力，对数据结构和基础算法有一定的掌握。编程语言上，近年来，Python 可以说受到了诸多数据相关从业人员的青睐。因为其语言的自身特点，相对于其他语言而言，比如 C++ 或者 Java，Python 对于从业人员来说是降低了学习和掌握的难度。但另一方面，我们也要意识到，大多数人工智能产品是一个复杂的产品链路。整个链路上通常是需要对多个语言环境都有所了解的。因此，掌握 Python，再学习一两个其他的语言，这时候选择 Java 或者 C++，是十分必要的。另外，很多公司都采用大数据环境，比如 Hadoop、Spark 等来对数据进行整合和挖掘，了解这些技术对于应聘者来常常说是一个让用人单位觉得不错的“加分项”。

第二，就是对于搭建一个人工智能系统（比如搜索系统、人脸识别系统、图像检索系统、推荐系统等）有最基本的认识。机器学习算法能够真正应用到现实的产品中去，必须要依靠一

个完整的系统链路，这里面有数据链路的设计、整体系统的架构、甚至前后端的衔接等多方面的知识。考察候选人这方面的能力是查看候选人能否把算法落地的一个最简单的方式。因此，从我们准备面试的角度来说，这部分的内容往往就是初学者需要花更多时间了解和进阶的地方。

数据科学家应聘的“软”实力

前面我们聊了数据科学家应聘的“硬”技能，下面，我们再来看看候选人还需要注意和培养哪些“软”技能。

数据科学家的第一“软”技能就是如何把一个业务需求转化成机器学习设置的“翻译”能力。

什么意思呢？和纯理论学习的情况有所不同，大多数真实的业务场景都是非常复杂的。当产品经理提到一个产品构思的时候，当设计人员想到一个业务创新的时候，没有人能够告诉你，作为一个数据科学家而言，这个问题是监督学习的问题还是无监督学习问题，这个问题是可以转换成一个分类问题还是一个回归问题。有时候，你会发现好像几条路都走得通。因此，如何能够从逻辑上，从这些不同的设置所依赖的假设上来对业务场景进行分析，就成了数据科学家必不可少的一个核心能力。

分析业务场景这个“软”技能的确非常依赖工作经验。这里不仅仅是一个机器学习问题的“翻译”，还需要对整个系统搭建有所了解，因为真正合适的场景“翻译”往往是机器学习的问题设置和系统局限性的一个平衡和结合。举一个例子，一个推荐系统需要在百毫秒级给一个用户进行推荐，那么相应的方案就必然有一个计算复杂度的限制。

因此，场景的“翻译”其实是考察数据科学家和人工智能工程师的一个非常重要的步骤，也是看候选人是否真正能够学以致用的有效手段。

说到这里，你是不是会有疑问：如果我没有相关的从业经验，那如何来锻炼这种“翻译”能力呢？

其实，现在丰富的互联网产品已经为我们提供了一个无形的平台。当你在现实中看到一个真实产品的时候，比如京东的产品搜索、科大讯飞的语音识别系统等等，你设想一下，如果你是设计者，如果你是需要实现这个产品功能的数据科学家，你会怎么做？

实际上，很多面试问题，都是面试官直接询问你对某一个现成产品的设计思路，比如谷歌的面试官可能会询问你如何设计一个搜索查询关键字拼写检查组件。这个方法一方面是帮助你“开脑洞”，另一方面也是一种非常好的思维锻炼。

另外一个很重要的“软”技能就是数据科学家的沟通表达能力。

这可能会让有一些人感到意外，因为大家也许认为数据科学家和人工智能工程师完全是技术岗位，并不需要与人打交道。其实，这个理解是片面的。就像刚才提到的，数据科学家的一个重要职责就是把现实的业务场景“翻译”成机器学习的设置，那么在这个过程中，会和业务人员、其他工程师、科学家进行高频的沟通和交流。如何把你的思路、方案清晰地表达给同事和团队成员是非常重要的职责。

实际上，数据科学家不仅在公司内部承载着的这样的沟通任务，我们往往还需要在社区中做演讲、参与讲座等活动，成为社区中的一份子，都离不开沟通表达能力的磨练。

如何锻炼沟通表达能力呢？这里，我给初学者一个简单而实用的方法，那就是用一两句话来总结你的方案。你尝试用一小段话，但是不夹带任何专业术语，把你的方案说给不懂机器学习的人听。这个训练方法可以让你反复思考，直到找到一个最简洁有力的表达。

小结

今天我为你讲了人工智能工程师和数据科学家的一个重要的职场话题，那就是作为数据科学家应聘时需具备的“硬”实力和“软”实力。

一起来回顾下要点：第一，我们讨论了机器学习、统计知识和系统这三大“硬”实力。第二，我们分析了场景翻译和沟通能力这两个“软”实力。

最后，给你留一个思考题，当下深度学习框架大行其道，那么对于应聘来说，你觉得了解和掌握各种深度学习框架会让你更有优势吗？

欢迎你给我留言，和我一起讨论。

上一篇　045 | 职场话题：当数据科学家遇见产品团队

下一篇　047 | 职场话题：聊聊数据科学家的职场规划

精选留言 (1)

写留言

吴文敏
2018-02-28

可定会有优势的，现在很多公司都在业务中使用深度学习模型，对这些框架有所了解可以很快选择适合业务场景的框架并将思路转成实际可运行的代码。但是，不可本末倒置，数据科学家的内功仍然是老师在本文所提出的硬实力与软技能。