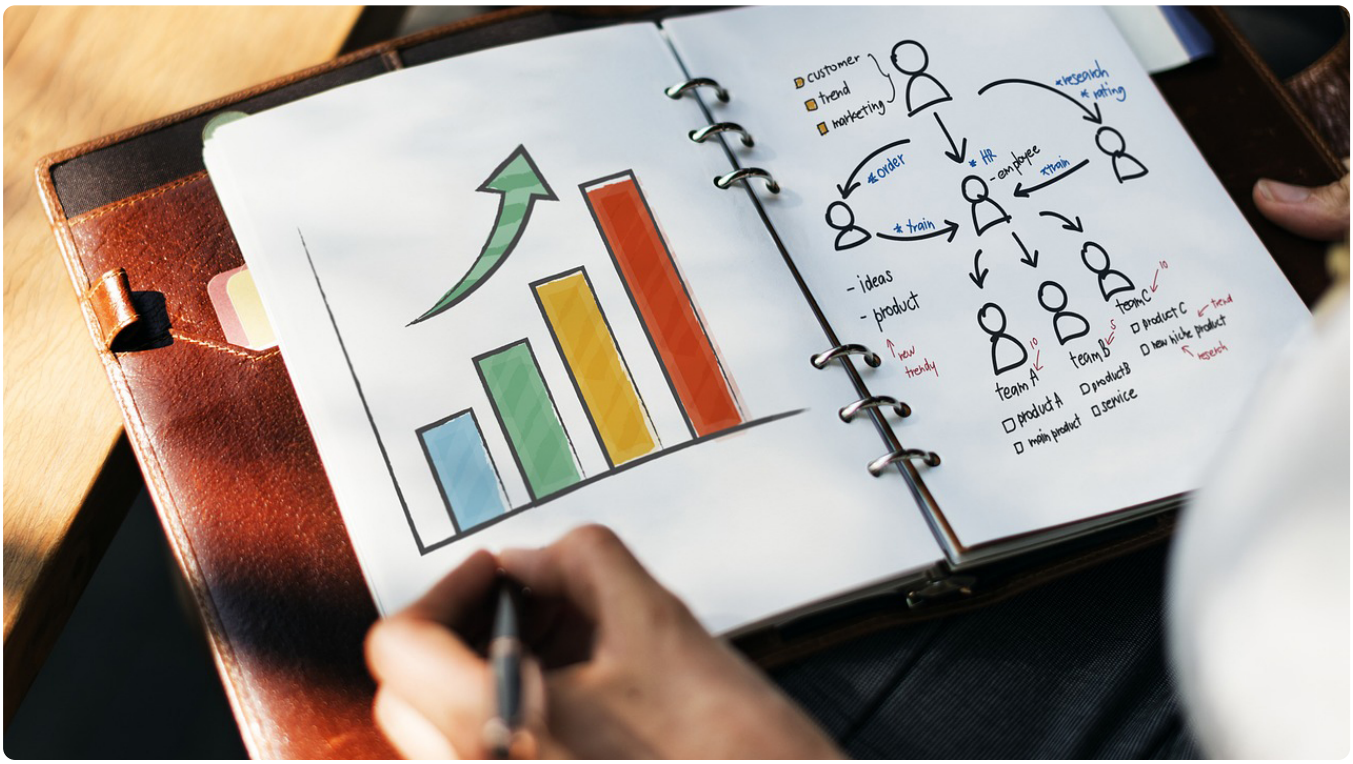
开篇词 | 你为什么需要数据分析能力？

2018-12-17 陈旸



讲述：陈旸
时长 08:01　大小 18.39M

你好，我是陈旸，清华大学计算机系博士毕业。清华有一门课，叫数据挖掘，正是通过这门课，我学会了如何从海量的数据中找到关联关系，以及如何进行价值挖掘。那时候感觉自己掌握了一门利器，就特别想找到一个钉子，来试试自己手里的这把锤子。

当时恰好赶上 2009 年微博的热潮。我用 3 个月的时间就积累了 4 万粉丝，一年的时间积累了上百万粉丝。这是怎么做到的呢？

通过数据采集，我收集了每天的微博热点，然后对热点进行抓取、去广告，再让机器定时自动进行发布。同时我让账号每天都去关注明星的粉丝列表，这样可以获得 15% 的回粉概率。久而久之，就会有源源不断的粉丝。

你看，其实就是数据分析帮我做到了微博的自动化运营。这还只是一个小例子，数据分析的影响已经渗透到了我们工作生活的方方面面。

通过数据分析，我们可以更好地了解用户画像，为企业做留存率、流失率等指标分析，进而精细化产品运营。

如果你关注比特币，数据分析可以帮助你预测比特币的走势。

面对生活中遇到的种种麻烦，数据分析也可以提供解决方案，比如信用卡反欺诈，自动屏蔽垃圾邮件等。

可以说，我们生活在数据驱动一切的时代，数据挖掘和数据分析就是这个时代的“淘金”，从国家、企业、组织到个人，都一定会关注各种数据，从这些数据中得到价值。

也正是这个原因，数据分析人才成了香饽饽，不管是数据分析师，数据分析工程师，还是数据产品经理，有数据思维的运营人员，都变得越来越抢手。你是不是也已经摩拳擦掌，做好了了解这一领域的准备呢？

我想在接下来的 15 周时间里，把自己在清华学习数据挖掘的体会和工作实践中对数据分析的理解，重新梳理整合呈现给你，和你一起在数据分析这个领域来一场急行军。

说了这么多数据分析的重要性，你是不是有这样的疑问：我也知道数据分析能力很重要，但是数据分析是不是很难？到底该怎么学呢？

其实这里有一些误区，数据分析并非遥不可及，它不难，掌握高效的学习方法很重要；但是它也不简单，需要你耐下性子，跟我一起来慢慢掌握数据分析的核心知识点和工具操作。

我招聘过一个实习生，很普通的本科学校。最开始他只会简单的 PHP 语法，实习期间薪水也就只有 3000 元，但到后来他不仅可以做爬虫抓取，还可以做数据分析，薪水就涨到了税后 1.3 万，这个进步用了不到一年的时间。

他的成长速度非常快，这是怎么做到的呢？

总结一下，就是他找到了高效的学习方法，我把它称为**MAS 方法**。

Multi-Dimension：想要掌握一个事物，就要从多个角度去认识它。

Ask：不懂就问，程序员大多都很羞涩，突破这一点，不懂就问最重要。

Sharing：最好的学习就是分享。用自己的语言讲出来，是对知识的进一步梳理。

所以学习这个专栏我们也用 MAS 方法，我来负责你和数据分析建立起多维度连接，你来负责提问和分享。

怎么和数据分析建立多维度连接呢？我特意把内容分成了三个大类。

第一类是基础概念。这是我们学习的基础，一定不能落下。

第二类是工具。这个部分可以很好地锻炼你的实操能力。

第三类是题库。题库的作用是帮你查漏补缺，在这个过程中，你会情不自禁地进行思考。

这个连接的过程，也是我们从“思维”到“工具”再到“实践”的一个突破过程。如果说重要性，一定是“思维”最重要，因为思维是底层逻辑和框架，可以让我们一通百通，举一反三，但是思维修炼也是最难的。所以，我强调把学习重心放在工具和实践上，即学即用，不断积累成就感，思维也就慢慢养成了。

说到底，**学习数据分析的核心就是培养数据思维，掌握挖掘工具，熟练实践并积累经验**。为了能带给你更好的学习效果，我在专栏里设计了五大模块。

1. 预习篇

我会给你介绍数据分析的全景图，和你进一步探讨最佳的学习路径。我还专门准备了 3 篇 Python 入门内容，如果你还没有 Python 基础，希望能帮你快速上手，如果你已掌握了 Python，可以当作一个复习。这么安排是因为 Python 是数据科学领域当之无愧的王牌语言，很多数据分析利器也是基于 Python 的（再或者，你也可以购买极客时间上的“零基础学 Python”视频课程）。

2. 基础篇

我会带你修炼数据思维，从数据分析的基础概念，到数据采集、数据处理以及数据可视化。我们一起从数据准备的整个流程上了解数据的方方面面。

3. 算法篇

算法是数据挖掘的精华所在，也是我们专栏的重点内容。我精选了 10 大算法，包括分类、聚类和预测三大类型。每个算法我们都从原理和案例两个维度来理解，达到即学即用的目的。

4. 实战篇

项目实战是我们学习的一个重要关卡。我准备了 5 个项目带你真实体验。比如在金融行业中，如何使用数据分析算法对信用卡违约率进行分析？现在的互联网产品都进入到千人千面的人工智能阶段，如何针对一个视频网站搭建视频推荐算法？

5. 工作篇

我选择了几个大家最关心的职场问题，比如面试时注意什么，职位晋升路径是怎样的等等，助你一臂之力。

我希望，通过这个专栏，你将有如下收获。

1. 数据和算法思维

这不仅是在技术上的思维模式，更是我们平时看待问题解决问题的思维方式。如果你将数据视为财富，将数据分析视为获得财富的工具，那么在大数据时代，你将获得更宽广的视野。

2. 工具

用好工具，你将拥有收集数据、处理数据、得到结果的能力，它会让你在工作中游刃有余。

3. 更好的工作机会和价值

无论是当前火爆的人工智能，还是数据算法工程师的市场，都很看重数据分析和数据处理的能力。从“思维”到“工具”再到“实践”，沿着这个路径拓展自己的能力边界，拥有更强的竞争力。

在你面前，即将开始一场数据科学之旅。我们一起用 15 周的时间，从算法原理、分析工具和实战案例三个维度体会数据科学之美。

在专栏学习的过程中，如果你遇到问题，不论是概念不懂，还是工具使用遇到 error，你都可以来找我。也希望你可以把自己的学习笔记分享出来，它不仅是最好的自我学习方法，也是最好的交流语言。

我愿意跟你一起，将这些看似“高大上”的内容琢磨得通俗易懂。当你完成这段旅程，你将会发现这个世界从来不缺少“石油”，而它们，正在等着你的勘探。

正式启程之前，我想邀请你聊聊自己对课程的期待，你如何看待数据挖掘和数据分析？你的工作和生活中有什么事情用到过数据思维吗？

下一篇　01丨数据分析全景图及修炼指南

精选留言 (157)

写留言

Hank_Yan
2018-12-17

61

业务洞察是分析数据的前提，分析数据是理解数据的前提，理解数据是数据挖掘的前提。从业务到数据再到挖掘，每一步环环相扣，相辅相成。业务千变万化，规律亘古不变。期待老师提纲挈领，从整体思路点拨，用经典案例教学，让每一位学生学到真本事，共勉。

展开

作者回复: 这位老师总结的也很到位👍

Alex王伟健
2018-12-17

23

期待，当然是以找到相关工作为目的啦。

展开

作者回复: 如果你能15周都坚持下来，每次课都能整理笔记，认真做练习，我也可以给你推荐工作的😊

别问
2018-12-17

22

求推荐一些数据分析的书，谢谢。

展开

作者回复: 思维：
《思维简史：从丛林到宇宙》
数据处理
《数据挖掘：概念与技术》
《Pentaho Kettle解决方案》
《精益数据分析》
《Small Data》
《利用Python进行数据分析》

汪汪汪
2018-12-17

15

本人是转行学习数据分析，想通过两个月时间自学，顺利拿到offer进入岗位进行实操。目前看了《深入浅出数据分析》那本书，然后学了python基础知识，想请问老师，接下来该如何开展学习计划。我想学python常用的几个库，从爬虫开始获取外部数据，熟悉常用的数据挖掘算法，最后花两个星期学习基础的SQL和excel操作。您的建议是什么?我手上的学习资源比较多，所以得重点筛选。期待老师的回信

展开

作者回复: 多谢关注，1）首先从爬虫开始是不错的，这样你能感受到成长的过程。

2）数据挖掘算法，如果你想了解十大算法的话，理论部分你需要花一些功夫。当然这些在Python中都有类库可以使用。做练习的话，你也可以把这些算法都用一遍，然后看下哪个算法模型的结果更好

3）网上这方面的资源确实比较多，他们大多讲的是理论原理。我认为你更注重的在于实战，因为做项目不仅更有成就感，还能更好的让你理解这些算法、爬虫的原理。

我会在专栏里给你做个“专属题库”，对应爬虫、数据挖掘这些的题目，你可以做个评测，不明白的地方，我也会给你做讲解。

4）资料比较多，但其实不用每个都看一遍。尤其是理论的部分，看一遍就可以了。关键是把它抽出来做个思维导图，方便查询，这样下次看导图就能回忆起来讲的是什么。省时又高效！

五岳寻仙

2018-12-17

9

老师好！看到这个专栏很兴奋！对数据挖掘/机器学习很感兴趣，自学有段时间了，也接触了不少工具，但遇到具体问题还是很盲目，有下面几个方面的困惑：

1. 如何做好“特征工程”，没有思路，也没有思考方向，看了不少博客，所谓的技巧也都知道了，但遇到问题还是用不好；...

展开

作者回复: 感谢你的热情和关注，我认为非常有必要自己使用这些机器学习算法来解决实际问题。

当然原理可以采用伪代码的方式，把流程画出来即可。项目中，很多时候都是直接使用类库，所以你更应该关注的机器学习的效率和结果。

很多时候，我们在选择模型的时候，都要试，一次会用多种模型，然后看训练结果的好坏，再决定采用哪个模型。

特征工程，以及调试的过程其实就是经验积累的过程，很多时候调参数的时间，比你写程序的时间还要长。但是这个积累过程还是挺重要的，当你有了更多经验之后，这个“试”的效率就会提升！

Aggi
2018-12-17

9

希望多讲一些分析的思维，以及和实际业务关联的案例的整个流程

展开

作者回复: 这个没问题，专栏中重点就是告诉你如何使用这些工具，以及案例实战训练。当然你也会在案例和工具中，训练你的数据思维，以及对他们的认知

任欣
2018-12-17

6

老师讲的数据就是这个时代的石油，确实是这样子的，在读研的时候深有体会，实验室的很多科研，项目都需要用到数据分析的思维和能力，工作之后也在为现在的公司处理数据帮助运营人员进行精准营销，无论是传统行业还是互联网行业，这都是一门重要的能力，希望以后能够在课上和老师有更多的交流。

作者回复: 好啊，欢迎交流。同意你说的，传统行业和互联网行业，不论是运营岗，还是营销岗，都需要数据分析能力和思维。

Fergus
2018-12-17

4

自己在从事这方面的工作，更多的时候是拿着钉子找锤子，同时朝着“自动化”的方向去改善自己的工作方式。
随着工作的开展，发现自己的基础不扎实、知识过于分散，同时缺乏数据的解读能力，目之所及即是思想的极限。
读完文章和留言已非常有收获，感谢。

展开

作者回复: 感谢关注，你说的我也很有同感。我们处于知识爆炸的时代，参考资料很多，但其实会出现另一个问题：就是知识过于分散。
所以这里，我建议大家要学会整理，每次课程做笔记，总结思维导图。当然课程里，我也会给出思维导图。方便你做知识梳理

孟令湛

4

2018-12-17

数据是无价的，希望通过学习，了解掌握数据分析挖掘的方法，并应用于工作生活里

展开

作者回复: 有我在，你一定可以的！

姜戈 4
2018-12-17

之前一直在看推荐系统的内容，还没入门，就被各种算法搞得头大，浏览了课程安排，希望数据分析45讲让我对推荐系统的学习打下坚实基础.

作者回复: 其实实战是最好的学习，你可以在项目实战中体会这些算法，当然我也会给你讲解这些算法的原理。所以我安排了从“认知”=>“工具”=>“实战”的过程，并且会给你总结“思维导图”和“专属题库”帮你来巩固学习

Yezhiwei 3
2018-12-18

希望能学到老师训练思维的方法

展开

作者回复: 非常认同你说的，我们从小习惯“知识性”的教育，以考试为例。而国外更注重“思维性”的训练，会让你进行主动探索。
所以思维培养，一个很好的方法：就是主动分享，有一颗好奇心！

□汤铭丰... 2
2018-12-25

你好~ 我从事数据分析也有一段时间了~ 现在的主要分析手段还是以hive sql为主，不知道如果当用python处理大体量的数据的时候一般是怎么操作呢？怎么把算法实现和落地到大数据里呢

Conan 2
2018-12-19

Multi-Dimension:
1. 理解每节课中讲到的概念
2. 重复文章中的代码示例
3. 自己根据已经学到的内容再进行拓展学习
Ask:...
展开

作者回复: 加油 总结的不错 你也可以找身边朋友或者同事进行提问。

Louie Zha...
2018-12-19

2

学生党一枚，之前自学python以及相关的第三方库，也了解一些机器学习算法。现在想跟着老师系统地学习一遍，也是自己查漏补缺、完善的过程，希望明年春招能够找个好工作。

作者回复: 加油 多做练习 整理笔记 到时候可以放到简历中

reverse
2018-12-19

2

去找吧 ，我已经把我的极客时间数据分析实战45讲的笔记放在github上了，地址: https://github.com/xiaomiwujiecao/geekTimeDataAnalysisInAction 欢迎大家加入一起维护

猴哥
2018-12-19

2

大一新生，刚好是大数据专业，希望接下来的15周里面可以不掉队，多跟着老师学些有用的知识。感觉老师讲的蛮生动的，一定能物超所值，我非常非常非常满意

作者回复: 加油 大一新生都开始学数据分析了 我当年还没这个觉悟 一定会比你在学校里上课有收获的

微光lu
2018-12-17

2

老师您好，跨专业的研一同学，选了数据挖掘这门课，老师上课主要讲了一些算法 有时候算法原理可以听明白，但是让自己用实际数据编程实践就很困难完成不了，目前学期末了Python才刚简单学了一遍，想问老师算法，编程需要掌握到什么程度，才可以达到能用实际数据分析。谢谢老师

作者回复: 算法原理和使用是两个维度，你们课上老师给你讲的算法肯定是从原理出发，到讲解论证的过程。

这个对你加深理解算法有帮助，但实际使用的时候，你就不用再关心这个论证的过程了，而需要关心：如何使用，结果如何

我建议你：

1）从实战项目出发，我会给专栏的读者制定一个“专属题库”，提升你的上手能力和成就感

2）在实战过程中，你也可以加深对Python使用和算法的理解。

草莓味冰糕
2018-12-17

2

我是一个想转商业数据分析与挖掘的生物学（生物信息方向）硕士研究生，很需要有一门课大概能告诉我一个算法或者数学模型适用于哪些商业或者运营的情景，这是我现在急需的，也是对课程的期望，哪些东西可以解决哪些问题，也希望作者能推荐一些类似的书，期望自己能在这么课收获很多，找到自己的路

作者回复: 我上大学的时候，也了解一些生物信息学的情况，非常能理解你的心情和想转到商业数据分析的决心。

我觉得需要从两个方面来下手：

1）工具角度：课程里讲的算法，你可以帮他当做是个工具。他的诞生是从数学原理开始，形成的理论模型。

这些模型都有自己的特点和适用范围。但总的来说，还是工具

2）商业角度：工作或应用中，首先都是从商业角度出发的，尤其是哪些是高频使用的，或者离“钱”更近的地方，也就是决策价值更大的地方。

当然从工具使用到商业价值的转换，还需要你有自己的思维和建模能力

商业相关书籍推荐：

《洛克菲勒留给儿子的38封信》

《商业冒险：华尔街的12个经典故事》

《从0到1：开启商业与未来的秘密》

《商业的本质》

数据分析相关书籍：

《数据挖掘：概念与技术》

《Pentaho Kettle解决方案》
《精益数据分析》
《Small Data》
《利用Python进行数据分析》

Robin™ 2

2018-12-17

老师好，对于多维数据的可视化的方法论本课程是否可以有涉及？

展开

作者回复: 有的！课程里都会介绍实际工作中最常用的部分，如果哪些地方没有介绍到，只要你们留言了，我也会在答疑篇或者加餐篇整理进去的

upup 2

2018-12-17

思维和业务能画等号吗？我认为不懂业务只会工具和算法的不叫数据分析师，因为他没办法解释业务。有了数据思维能通用于任何行业吗？

作者回复: 同意你说的，我在后面也会讲到，想要用数据挖掘，第一步是对商业的理解，只有确定好了商业目标，数据挖掘才有目标。
数据思维是一种思考方式，世界本身有很多维度，我们从哪个维度看待它，就会从哪个维度收获它