DC学院
class.pkbigdata.com

数据分析师成长手册

一份完善的优质数据分析师成长计划

DC 学院　研制

开启数据分析之旅

你的职业生涯可能面临很多选择，每一种选择都预示着一种可能性。相信当你打开这份技能清单时，已经做好了学习数据分析核心技能，甚至成为一名数据分析师的准备了。

数据分析师被誉为是未来最性感的职业之一，他们认识世界基于大数据，因为真实可靠，他们能从看似毫无关联的数据中，提取出别人看不到的信息和知识。

数据分析师的这种能力，基于对统计学知识的了解，也基于对编程语言和数据挖掘方法的认知。我们总结了很多优秀数据分析师的经验和目前主流的招聘需求，整理了这份数据分析师必备技能清单，帮助你了解数据分析的整个技术知识体系，你也可以据此规划你自己的学习路径。

DC 学院也推出了《数据分析师（入门）》体系课程，从数据爬取、数据存取、数据分析/数据挖掘、报告及可视化这四个数据分析的基本流程展开。由此帮你掌握 Python 基础和网络爬虫、SQL 数据库语言与 MySQL 数据库管理软件、概率统计知识、Python 数据分析以及机器学习建模的理论，并能够独立完成商业数据分析项目。

不管怎样，恭喜你，已经迈出了数据分析的第一步。要相信你会从中获得很多思考、分析问题的方法和技术，这些都是可以在不同的工作中迁移的核心技能。

如果你已经了解并确信这就是你需要的技能，你可以看看我们的课程《数据分析师（入门）》，或者继续阅读这份成长手册！

数据分析师技能清单

总的来说，数据分析师需要具备基本的概率统计基础知识，数据库的基本操作，科学计算部分的编程知识（推荐 python）及初级的机器学习知识，以下是整个知识体系的概览，你可以先有一个初步的全面了解。

- Python基础与网络爬虫
  - Python基础语法
  - Python网络爬虫
- SQL数据库知识
  - MySQL数据库
  - SQL操作语句
- 概率统计知识
  - 概率论
  - 统计学
- 利用 Python 进行数据分析和可视化
  - Python数据分析
  - Python数据可视化
- 机器学习基础
  - 监督学习
  - 无监督学习
  - scikit-learn机器学习包

第一部分
Python基础与网络爬虫

Python已经是最受欢迎的动态编程语言之一，也是实现数据分析最主流的语言。加上Python的开源特点和不断更新的库，使得Python跃升为数据分析的一大利器。掌握Python基础语法和爬虫功能是入门数据分析的第一步。

Python 优点

- 高层次的结合了解释性、编译性、互动性和面向对象的脚本语言；
- 易于学习、易于阅读、易于维护；
- 具有丰富的、广泛的库，可以解决各种问题；
- 提供了针对科学计算、结构化数据处理以及数据可视化的功能强大的库；
- 提供了主要的商业数据库接口。

1.1 Python 基础语法

想要使用Python写网络爬虫获取网上数据，或者对数据进行操作、处理，使用Python进行数据可视化等，都需要你首先掌握Python的基础语法。就像游戏中掌握每种道具的基本属性和规律，才可以真正用好这个工具，打起boss来不费劲儿。Python的基础语法很简单，适合快速入门。需要掌握的基本知识点如下（包括但不限于）：

Python 基本术语

- 解释器：要运行代码就需要用Python解释器来运行，主流的解释器有CPython、IPython、PyPy等；
- 数据类型：字符串、布尔型、整数、浮点数、列表、元组、字典、集合等；
- 运算符：Python主要的运算符有数学运算符、逻辑运算符、比较运算符；
- 表达式：由值、变量、运算符组成；
- 控制流：Python有三种控制流，if/for/while来控制表达式执行的顺序；
- 函数、变量作用域（局部和全局）、lambda函数：使得代码变得更简洁和更具有可迁移性；
- 字符串操作：替换、删除、截取、复制、连接、比较、查找、包含、大小写转换、去空格、分割等；
- 数据操作：数据索引、切片、添加、插入、移除、排序等方法；
- 正则表达式：使用正则表达式可以实现模糊匹配、替换和拆分。

1.2 Python 网络爬虫

数据的获取方式有很多，你可以直接使用现成数据集、下载网上公开的数据集、利用Python连接API进行爬取、利用Python进行基于HTML网页爬取，从数据库提取想要的数据等。在这个部分你需要学习如何通过自己编写的代码来从网上获取你想要的数据集，这也是数据获取中最有趣味和技术含量的一种方式。

基于HTML网页的爬取

自己编写代码爬取数据，可以绕过网站的限制，但是对知识储备的要求相对高一些，这类爬虫方式的本质依旧是HTTP请求。

- HTML基础：对网页审查元素的了解是爬虫的基础，你需要学会用浏览器工具来对网页信息进行定位；
- 可供调用的包：Python的Beautifulsoup包从网页中可以抽取定位到的信息；
- 爬虫技巧：在编写代码的时候需要注意不同网页的特征，并构造合适的query；同时越来越多的网站为了维护正常用户的访问需求，而实施了一些反爬虫技巧，使用代理服务器或利用手机端网页可以简化爬虫的难度；
- 常用的字符编码及转换：在爬取中文网页时，结果经常会返回一串看起来无意义的字符，是因为同一个汉字在不同的编码格式下，差异巨大，Python如同很多语言一样，不能智能地识别编码。所以在爬取数据的过程中，你还需要注意字符的编码格式问题。

基于API的网络爬虫

基于API的爬取方式是最简单直接的，调用Python的urllib、urllib.request包连接API接口就可以进行。

第二部分
SQL数据库知识

2.1 MySQL 基础知识

MySQL为世界上最受欢迎、使用最广的数据库管理系统之一，是一个稳定运行、检索效率较高的系统，作为想入门数据分析师的你，数据库管理系统将是你连接数据的重要通道。

- 数据库设计原理：设计为客户机-服务器，用户面对客户机，而关于数据的请求由服务器处理；
- 数据类型和时间格式：数据类型定义每个字段的存储的规则，如存储的长度和精度范围；
- 数据库编码：数据库管理系统是一个非常讲究规范的系统，MySQL也有自己要求的编码格式，可根据指令可以查询、转换、存储成符合MySQL编码要求的数据；
- 文件格式：数据文件的常见格式有CSV,JSON等等。

2.2 SQL 语言基础

SQL语言是结构化的查询语言，是连接使用者和数据库之间的通道，几乎所有的数据库管理系统都可以通用SQL，数据分析师每天都需要和数据打交道，SQL也是你的必修技能。

SQL表格操作

SQL语句操作的对象是表格，学会如何建立、更新表格

- 建立表：SQL语句操作的对象是表格，表格的建立是操作的基础，可以使用CREATE命令进行建立；
- 插入、更新和删除数据：建立了表格之后需要使用INSERT,DELETE,DROP来对表格进行更进一步的插入新值新列、删除部分数据、删除整张表的操作。

SQL查询操作

- 数据检索：用SELECT语句来实现检索功能，它是最经常被使用到的SQL语句；
- 数据排序：ORDER BY语句；
- 数据过滤：WHERE语句限定了搜索的条件；
  BETWEEN/IN/NOT操作符限定了查询值的范围；
  AND/OR操作符表达搜索条件间的逻辑关系；
  数据汇总和分组可以使用GROUP BY语句。

2.3 SQL 语言高级技巧

SQL进阶操作

- 子查询和组合查询

子查询：可以嵌套在主查询的SELECT,WHERE,FROM,GROUP BY等位置，使用子查询可以写出具有更复杂功能的SQL语句，使得查询更加灵活。

组合查询：使用UNION操作符，连接多个SELECT语句，把多条查询结果当做一条组合查询返回，大大简化了查询的复杂程度。

- 表联结（JOIN）

关系型数据库的设计方便了处理和提高了存储的效率，然而却带来了一个问题，就是跨表的查询。联结是一种机制，用于关联不同的表。

基本概念：主键、外键。

种类：内部联结；自然联结；外部联结；带聚集函数的联结。

- LIKE操作符和正则表达式

LIKE操作符支持的通配符：可以用于匹配搜索值的某一部分，来实现数据的过滤;

正则表达式：使用REGEXP进行匹配，正则表达式的函数主要分为三大类，对应三类不同的功能：模式匹配、替换、拆分。它们之间是相辅相成的。

第三部分
概率统计知识

数据分析是基于适当的统计分析方法对海量数据进行整理、探索数据内部结构、提取信息、形成结论的一门学问。理解概率论、统计学的理论是理解数据分析模型的重要基础。

3.1 概率论基础

概率学是用于研究随机现象的学问。掌握好概率论的知识，可以帮助你更好地理解接下来数据分析方法背后的数学模型。

概率学基础

入门概率论，你需要理解三大概率（联合概率、条件概率、边际概率）、事件独立性的含义、Bayes公式的应用。

概率分布

在这个部分，你需要了解概率分布函数、累计概率分布函数、正态分布、二项式分布、泊松分布、超几何分布等的相关概念与应用环境。

采样及中心极限定理

熟悉样本整体、采样、采样分布和中心极限定理的概念，可以更好地理解概率和频率是怎么联系在一起，能更深入地掌握统计抽样的思想。

3.2 描述统计学

统计学是一门古老的学问，主要分为描述统计和推断统计。统计学是处理数据的学科，也是数据分析的基础。你需要掌握统计学的基本概念、统计图、统计量、数据描述方法、置信区间、假设检验。

统计学基本概念

变量和样本是统计学建立的基础，变量用于形容事件的某种特征，样本是总体一部分元素的集合。现实中无法穷尽所有的元素，只能对选取的样本进行分析，所以样本的选择尤为重要。

常用统计图

统计图是用于描述数据的图形，可以直观地展示数据的特征，一目了然。常用统计图包括有：条形图、直方图、散点图、箱线图、统计地图等。

基本统计量

统计量是根据样本数据计算出来的，是样本的函数，用于对数据进行分析和检验。常见的统计量有：平均数、中位数、方差、标准差、Z-score等。

数据的描述方法

描述统计是描述、总结变量的基本情况的统计，研究反映客观现象的数据，并通过图表形式对数据进行可视化，进而综合概括与分析得出反映客观现象的规律性数量特征。

- 将数据资料转化为图表 ,直观展示数据的分布情况。通常使用频数分布表、直方图、折线图、条形图、频数分布图等图表；
- 分析数据，了解各变量内观察值的集中和分散情况。描述集中趋势的有：平均数、中位数、众数、几何平均数和调和平均数等，描述分散趋势的有：标准差、方差、最大值、最小值、全距、平均差和四分差等；
- 表示数据与常态分配偏离情况，使用偏态与峰度。

3.3 推断统计学

置信区间

置信区间是统计学的概念，是由样本统计量构造的总体参数估计区间，是假设检验的基础，了解过置信区间、整体平均值的置信区间、整体比例的置信区间这些概念与构造方式后，你就可以进入到假设检验的部分。

假设检验

假设检验：在一定的假设条件下，由样本推断总体的方法，常用的假设检验方法有u检验、t检验、卡方检验、F检验，秩和检验等。假设的构造、检验方法的运用、1型和2型错误可以帮助你理解如何检验参数的可信程度。

相关性与回归分析

相关性描述和回归分析常常用于衡量变量之间的关系，这也是数据分析的重要任务。在掌握一些相关分析的定义、相关系数（皮尔逊相关系数）、相关性非因果以及回归分析中的参数解读、评价指标、检验（hold-out 检验、交叉检验）等内容之后，你会对数据预测以及数据建模有一个初步的认识，并可以开始把你所学到的知识运用到实际的预测建模当中。

第四部分
利用Python进行数据分析和可视化

熟悉掌握相关计算机技能、理论知识之后，你可以开始深入学习Python是如何利用丰富的数据分析、可视化包来完成相关的分析任务。

4.1 数据分析（NumPy / Pandas）

NumPy可以实现对于数组和矢量的操作，Pandas专注于结构化（表格化）的数据操作、处理和运算，这两个库为使用Python进行数据分析提供了简洁、丰富的指令，使Python成为数据分析的利器。

4.2 Python数据可视化

数据可视化是一个非常重要的部分，选择合适的方式可以让你的分析结果一目了然。常用的Python可视化工具包有：matplotlib、seaborn、plotly等。

第五部分
机器学习基础

在学习完数据分析的理论基础和计算机技能后，可以开始对机器学习的探索。机器学习旨在发明计算机算法，使得我们的工具——计算机可以从数据中自动分析获得规律，并利用规律对未知数据进行预测，利用数据来解决问题。学习本部分的知识可以使你获得对数据分析核心部分——算法有更进一步的理解。

5.1 监督学习

用已知分类的一组数据来调整分类器的参数，从而提升分类器的性能。

回归：线性回归和逻辑回归

- 线性回归 利用最小二乘函数对两个或以上变量之间相互依赖的线性关系进行探究；
- 逻辑回归：在线性回归的基础上，加入了函数映射，将函数值由原本的连续值映射到0,1区间，再进行求解。

分类：一般的分类方法

常见的分类方法有决策树、朴素贝叶斯、随机森林、支持向量机、人工神经网络、KNN、集成学习等。

5.2 无监督学习

相对于监督学习，无监督学习使用的训练数据是没有分类结果的，且是基于代价评判基础的。常见的无监督学习方法有：聚类分析（K-means聚类）、非负矩阵因式分解、自组织映射等。

5.3 机器学习在Python上的实现

scikit-learn是一个开源学习模块，为用户提供了许多机器学习的算法接口，可以使用Python直接调用。

如何开始学习

阅读完这份技能清单之后，相信你对数据分析师所需的基本技能有了更加深入的了解。如果你更加坚定了学习这门技术的想法，那么你需要针对上述的技能，做一些初步的学习规划并严格执行。

当然我们也非常推荐你加入DC学院推出的《数据分析师（入门）》课程，这门课将完全以时下流行的数据分析案例为导向，在解决具体的问题中学习理论，加上导师step by step的操作，相信你可以很快上手去独立完成一些项目。

比如学习完python你可以编写自己的爬虫去获取你想要的数据，学完数据库的知识，你可以完成一些数据的基本操作和提取，最终你可以基于统计和机器学习方法完成大数据的分析，并形成专业的分析报告。

我们针对性地准备了课后资料和练习题目，你可以随时检测自己的学习效果，还有与知识点完美匹配训练赛，提交答案即可获得评分，同时可以查看自己的排名情况。

点此进入课程《数据分析师（入门）》

当然，你也可以按照自己的计划去学习这些技能，但是要记住，一定要在两三个月内坚持学习，直到你可以独立去做一些实际的事情。

毕竟，坚持总是需要很多理由，而放弃，一个就够了。

最后，希望你在数据分析的路上披荆斩棘！

附：《数据分析师（入门）》课程大纲

实际课程会根据课程安排作细节调整

第一章：开启数据分析之旅

1) 数据分析的一般流程及应用场景
2) Python 编程环境的搭建及数据分析包的安装

第二章　获取你想要的数据

1) 获取互联网上的公开数据集
2) 用网站 API 爬取网页数据
3) 爬虫所需的 HTML 基础
4) 基于 HTML 的爬虫，Python（Beautifulsoup）实现
5) 网络爬虫高级技巧：使用代理和反爬虫机制
6) 应用案例：爬取豆瓣 TOP250 电影信息并存储

第三章　数据存储与预处理

1) 数据库及 SQL 语言概述
2) 基于 HeidiSQL 的数据库操作
3) 数据库进阶操作：数据过滤与分组聚合
4) 用 Python 进行数据库连接与数据查询
5) 其他类型数据库：SQLite&MongoDB
6) 用 Pandas 进行数据预处理：数据清洗与可视化

第四章　统计学基础与 Python 数据分析

1）探索型数据分析：绘制统计图形展示数据分布
2）探索型数据分析实践：通过统计图形探究数据分布的潜在规律（Seaborn 实现）
3）描述统计学：总体、样本和误差，基本统计量
4）推断统计学：概率分布和假设检验
5）验证型数据分析实践：在实际分析中应用不同的假设检验（scipy 实现）
6）预测型数据分析：回归、分类、聚类
7）预测型数据分析：用特征选择方法优化模型
8）预测型数据分析实践：用 scikit-learn 实现数据挖掘建模全过程
9）预测型数据分析实践：用 rapidminer 解决商业分析关键问题
10）高级数据分析工具：机器学习、深度学习初探

第五章　报告撰写及课程总结

1) 养成数据分析的思维
   问题的重要性（为什么要分析这个问题？）
   问题的准确定义（可以以假设检验的方式写出）
   如何选择分析所使用的数据集（数据来源是否可靠，内容是否充分？）
   问题分析所采用的方法（方法是否适用？）
   数据分析预处理（如何生成训练集、测试集）
   分析结果所采用的评价指标
   实验结果细粒度分析（如找到关键的自变量因素）
   要清楚分析所使用的方法以及数据集的局限在哪里

2) 应用案例及报告撰写
   实践展示数据分析的全流程
   数据分析报告撰写的技巧

3) 课程回顾以及一些拓展

关于DC学院

class.pkbigdata.com

DC学院为DataCastle（专业的数据科学学习社区）旗下的在线学习平台。在DC学院，你可以在高效的课程框架下学习，挑战企业的真实技术问题，和小伙伴在线竞技，提升自己的江湖地位。我们相信学习没有捷径，但是DC可以帮你少走弯路。

体系课

为用户提供实际岗位的系统知识学习服务，让用户快速掌握细分领域的核心技术，能够独立完成相应的数据科学任务，并胜任对应岗位的工作。

专题课

针对数据科学领域最常遇到或者广泛应用的问题，通过以实践为导向的形式，帮助用户快速获得解决具体问题的方法和能力。

DC学院师资

核心授课团队由电子科技大学教授周涛领衔，长江学者张小松和国家千人计划入选者徐增林担任长期课程指导老师。核心师资主要为领域顶尖数据科学家、名校老师以及企业技术合伙人，并不断吸引行业技术领袖参与数据科学知识的传播。

[点此进入 DC学院]

DC学院
class.pkbigdata.com

数据分析师成长手册

更多数据科学课程，上 DC 学院
class.pkbigdata.com