数据挖掘 求职直通班

更新日期 / 2019-06-17

互联网大厂紧缺岗位，大数据领域深造历练之选

一站式高效攀上大数据领域人才顶端

【系统学核心技能】

- 大数据领域森罗万象，从数据分析、数据管道、数据仓库到数据挖掘，优达学城为每位学员提供完整系统的学习路径和行业实战项目，让有转行和晋升想法的你，结合自身学习/工作经验，寻找适合的岗位和提升方法。

【提升职场软硬实力】

- 简历和面试是求职、转行或晋升之路的第一步，优达学城将为你提供大数据行业内知名专家的简历优化分享，技术模拟面试和竞赛辅导等，全面提升你进入新领域和获得新机会的成功率。

数据挖掘，掌握大数据领域核心技能

课程 | 全球大数据领域必备定制技能

课程内容涵盖大数据分析和数据挖掘等核心理论知识，除 Python、SQL 等必备技能外，进一步学习搭建机器学习模型、数据管道和推荐引擎，以及使用 Spark 大数据处理框架。

实战 | 7 大硅谷审阅实战项目，专家逐行代码审阅

通过 Learning by doing 的学习方式，跨越理论到实践的鸿沟，为项目经历加分。

拓展 | 6 种大数据热门技能，12 个前沿案例

涵盖互联网、电子商务、金融风控、医疗、交通出行、社交网络等热门领域前沿案例在线演练，实践与应用统计学、数据分析、机器学习、深度学习、A/B 测试、推荐系统等 6 大数据科学热门技能。

求职 | 竞赛刷题，求职辅导，有效提升求职竞争力

通过热门行业实战案例精讲、 Kaggle 等数据科学挑战赛、简历优化辅导、技术模拟面试等，有效提升求职竞争力。学习者有机会获得名企实习内推机会，获取更多项目经历。

课程概览

先修要求：掌握 Python 编程，并具备 SQL 和统计学基础

课程有效期： 8 个月

学习辅导：社群答疑

第一阶段 数据分析 建议学习时长：10 周	系统学习数据分析的基础知识，共4个章节： • 应用统计学 • 数据清洗 • 探索性数据分析 • 数据可视化 **实战项目（代码审阅）：** • Google Apps 商店的数据分析 • Twitter 社交网络的数据分析 • 互联网金融等行业探索与分析数据 **简历高光案例演练：** • 【应用统计学】用户心理与行为研究 • 【数据清洗】医疗数据清洗 - 口服胰岛素临床试验 • 【数据收集】豆瓣电影数据的爬取 • 【数据探索】钻石行业暴利研究 • 【数据可视化】摩拜单车在上海使用的数据可视化
第二阶段 机器学习 建议学习时长：10 周	系统学习机器学习基础知识，共4个章节： • 机器学习基础 • 监督学习 • 非监督学习 • 深度学习 **实战项目（代码审阅）：** • 为慈善机构寻找捐赠者 • 神经网络预测共享单车使用 **简历高光案例演练：** • 【监督学习】互联网广告点击率预测 • 【非监督学习】银行与信贷公司金融风控模型搭建 • 【深度学习】社交网络情绪传播预测
第三阶段 应用数据科学 建议学习时长：6 周	深入数据科学应用领域，共4个章节： • 推荐系统与试验设计 • 大数据分析与 Spark

	• 软件工程（英/选修） • 数据工程（英/选修） **实战项目（代码审阅）：** • 为 IBM 在线社区构建推荐系统 • 大数据分析和预测用户流失 **简历高光案例演练：** • 【A/B 测试】星巴克促销活动的优化策略 • 【推荐系统】社交网络用户中的电影推荐系统 • 【业务建模】金融服务公司欧唯特的客户分层报告 • 【业务建模】星巴克用户促销广告推送优化
第四阶段 求职策略 建议学习时长：2 周	学习求职期的准备工作与应对策略，共4个章节： • 找工作的策略与心态 • 通过简历和领英档案建立你的个人品牌 • 行为和技术模拟面试 • Github 个人资料完善 **求职辅导** • 热门行业实战案例精讲 • 数据科学技术模拟面试 *【福利】名企实习内推机会
第五阶段 实践拔高	**数据科学挑战赛** 时长 2 周，帮助学生参与到正在进行中的竞赛中（如天池、Kaggle等），通过导师的指导，完成自己的第一次提交，并在不停优化中获得不错的竞赛成绩，丰富简历。
热门技能包（选修） 注：无答疑服务	该模块中，补充了与核心课程相关的部分知识内容，常常在 JD 或面试中出现，不作为学习要求，供学员根据求职目标和技术背景等自行安排，共4个章节，包含： • Leetcode 算法精讲 • GitHub使用 • Shell Workshop

第一阶段：数据分析（第1-10周）

第 0 章：欢迎学习数据挖掘求职直通班！

课程名称	学习目标
课程简介	➔ 学前准备，了解课程架构，做好准备！
在线学习第一课	➔ 在线学习简介，学习技巧，学习困境急诊室

第 1 章：使用统计学验证应用中的假设

课程标题	学习目标
概率和条件概率	➔ 了解概率基础知识 ➔ 学习正态分布和二项分布 ➔ 理解条件概率和贝叶斯规则
抽样分布和中心极限定理	➔ 了解抽样分布以及编程实现 ➔ 了解中心极限定理 ➔ 自助法（bootstrap）抽样
假设检验	➔ 了解置信区间和 p 值 ➔ 假设检验、得出结论和常见的错误类型

案例演练：用户心理与行为研究

你是一名心理系的学生，在这个项目中，你需要使用描述性统计和统计检验来分析实验心理学经典成果——斯特鲁普效应，阐述你对实验数据的理解，并根据最终结果通过统计推断得出结论。

拓展应用领域：业务决策中，判断两组数据的结果是否具有显著的差异

第 2 章：数据分析入门

课程标题	学习目标
数据分析流程	➔ 了解数据分析流程的主要步骤 ➔ 运用 Python 和 Pandas 处理多个数据集

Pandas 和 Numpy：案例分析1	➔ 对一个数据集进行完整的数据分析 ➔ 学习使用 NumPy 和 Pandas 进行数据的整理、探索、分析及可视化处理

实战项目：Google Apps 商店的数据分析

世界上最热门的 Apps 是什么？他们之间有什么共性呢？你的公司想要开发一个新的移动端的应用，作为一名数据分析师，你需要从数据中获得一些洞察，来指导公司的业务，例如：

1. 商店中哪类 Apps 最多？
2. 哪类 Apps 的平均评分最高？
3. 哪个年龄群体的用户最爱使用某类 Apps?

拓展应用领域：分析目标数据集，并从中提炼洞察

第 3 章：数据清洗

课程标题	学习目标
数据清洗入门	➔ 了解数据整理流程的各个步骤（收集、评估和清洗） ➔ 利用基本的数据收集、评估和清洗代码来整理从 Kaggle 下载的 CSV 文件
收集数据	➔ 收集不同来源的数据，包括收集文件、以编程方式下载文件、网络抽取数据和访问 API 数据等 ➔ 将不同文件格式的数据导入 Pandas, 包括平面文件（如 TSV）、HTML 文件、TXT 文件和 JSON 文件 ➔ 将收集到的数据储存在 PostgreSQL 数据库中
评估数据	➔ 使用 Pandas 以及编程的方式直观地评估数据 ➔ 区分脏数据（内容或质量问题）和乱数据（结构或整齐程度问题） ➔ 辨别数据质量问题并用矩阵进行分类：有效性、准确性、完整性、一致性和统一性
清洗数据	➔ 了解数据清洗的各个步骤（定义、编码和检验） ➔ 使用 Python 和 Pandas 清洗数据 ➔ 使用 Python 以直观的及编程的方式检验清洗代码

案例演练：豆瓣电影数据的爬取

在这个项目中，你将会从豆瓣电影的网页中获取你最爱的三个类别，各个地区的高评分电影，收集他们的名称、评分、电影页面的链接和电影海报的链接。最后对收集的数据进行简单的统计。

拓展应用领域：维基百科、拉勾网、大众点评等公开网站数据的收集

案例演练：医疗数据清洗 - 口服胰岛素临床试验

本案例中数据集并非真实，Auralin 和 Novodra 并不是真正的胰岛素产品。这些临床试验数据仅是为了本课程演示而编制的。数据质量问题由真正的医生咨询机构建立，模拟了医疗保健数据中真实常见的数据质量问题（这些问题会对护理质量、患者注册和收入产生影响）。你可以通过本案例中学习医疗行业中常见的数据质量评估和清洗思路。

拓展应用领域：医疗相关行业数据清洗的常用策略

实战项目：Twitter 社交网络的数据分析

WeRateDogs 是一个推特主，他以诙谐幽默的方式对人们的宠物狗评分。WeRateDogs 下载了他们的推特档案，并通过电子邮件发送给优达学城，专门为本项目使用。这些数据不是干净的，你需要使用 Python 以及 Python 库对其进行清洗，并创建有趣和可靠的分析与可视化。

拓展应用领域：互联网社交平台的数据清洗、分析与可视化

第 4 章：探索性数据分析

课程标题	学习目标
什么是 EDA	➔ 明确并了解探索性数据分析（EDA）的重要性
R 基本知识	➔ 安装 RStudio 软件及程序包 ➔ 编写基本 R 语言脚本以检测数据集
探索一个变量	➔ 量化并可视化数据集中的单个变量 ➔ 创建直方图和箱线图 ➔ 变换变量 ➔ 检查并识别可视化中的得失

探索两个变量	➔ 合理运用相关技巧探索数据集中任意两个变量间的关系 ➔ 创建散点图 ➔ 计算相关性 ➔ 探讨条件均值

案例演练：钻石行业暴利研究

如何挑选最优质、最保值的钻石呢？本案例中提供多枚钻石的特性及参数（如大小、颜色、净度和切工等）供你使用与探索，你会对每个特性和价格进行相关分析，也会通过回归分析预测其价格。

拓展应用领域：探索变量与变量之间的相关性，并进行合理预测

实战项目：互联网金融等行业探索与分析数据

此项目是开放式的，正确答案不止一个。正如 John Tukey 所说："某些数据和对答案的极度渴望组合起来并不能保证可以从一组给定的数据中获得合理的答案。"我们希望你提出有趣的数据问题，并且给自己一个探索的机会。你可以选择自己的数据集进行探索，我们也提供一些可选的数据集供你探索：

1. 来自 Prosper 的贷款数据
2. 拍拍贷的业务数据
3. 国家总统竞选的财产捐助
4. 葡萄酒成分与质量数据

拓展应用领域：不同领域的数据探索分析和总结，从数据中挖掘信息

第 5 章：用数据讲故事

课程标题	学习目标
数据可视化基本原理	了解数据可视化的重要性 了解不同数据类型如何进行可视化编码
设计原则	根据数据特点选择最有效的图表 有效运用色彩、形状、大小等元素
用 TABLEAU 创建可视化	熟悉使用 Tableau 基本功能，如图表、过滤器、分层结构等 创建 Tableau 计算字段

用 TABLEAU 讲故事	创建 Tableau 仪表盘和故事，展示有效的数据可视化

案例演练: 摩拜单车在上海使用的数据可视化

运用 Tableau，可视化摩拜上海用户数据中的故事和趋势。你将熟练应用可视化，还将学会如何利用视觉编码实现令人难忘的沟通。

拓展应用领域：将数据转换成图表，有效地向他人传达你的发现

第二阶段：机器学习（第 11-20 周）

第 6 章：机器学习基础

课程名称	学习目标
训练与测试模型	➔ 使用 Pandas 读取数据集，并使用 scikit-learn 训练与测试模型。
评估指标	➔ 了解用于评估模型性能的指标，如准确率、精度、召回率和 ROC 得分。
模型选择	➔ 学习如何进行交叉验证，通过学习曲线判断过/欠拟合，并学习如何使用网格搜索来训练模型。
自我评估：NumPy 与 Pandas	➔ 测试你的 NumPy 和 Pandas 技能
自我评估：模型评估与验证	➔ 测试你对模型评估与验证的理解

第 7 章：监督学习

课程名称	学习目标
线性回归	➔ 使用 Pandas 读取数据集，并使用 scikit-learn 训练与测试模型。

感知器算法	➔ 了解用于评估模型性能的指标，如准确率、精度、召回率和 ROC 得分。
决策树	➔ 学习如何进行交叉验证，通过学习曲线判断过/欠拟合，并学习如何使用网格搜索来训练模型。
朴素贝叶斯	➔ 测试你的 NumPy 和 Pandas 技能
支持向量机	➔ 测试你对模型评估与验证的理解
集成方法	➔ 通过 boosting 提升传统方法；AdaBoost
自我评估：监督学习	➔ 监督学习相关的测试题

案例演练：互联网广告点击率预测

广告点击率预估是很多广告算法工程师喜爱的战场，可以帮助广告主和广告平台更好地做决策。你将搭建机器学习模型帮助企业预测广告是否会被点击，辅助公司决策，为公司带来增量收入。

拓展应用领域：计算广告，商业建模预测

_____

实战项目：寻找慈善机构的捐助者

CharityML 需要你帮助他们建立一种模型，来识别潜在的最可能捐助的人并降低发送邮件的费用。你的目标是评估和优化几个不同的监督学习模型，以确定哪种算法能够获得最高的捐赠收益率，同时减少发送信件中的字数。

拓展应用领域：掌握不同监督学习算法的原理、特点和应用场景，保证算法的可解释性

第 8 章：非监督学习

课程名称	学习目标
聚类	➔ 学习如何聚类算法，并尝试使用 k-means 对数据进行聚类
层次聚类法与密度聚类	➔ 学习单连接聚类法和层次聚类法，DBSCAN
高斯混合模型与聚类验证	➔ 学习高斯混合模型及相关示例
特征缩放	➔ 通过案例学习特征缩放

PCA（主成分分析）	➔ 了解降维的作用，并学习 PCA 的原理和使用场景
PCA 迷你项目	➔ 使用特征脸方法和 SVM 进行脸部识别
随机投影与 ICA	➔ 学习随机投影与独立成分分析，并通过 Lab 学习如何应用这些方法
非监督学习自我评估	➔ 非监督学习相关的测试题

案例演练：银行与信贷公司金融风控模型搭建

信贷平台中，如何判断申请人的信用非常重要。你将通过机器学习算法对每一位申请人的基本信息、交易信息等数据进行甄别，减少信贷公司的坏账损失。

拓展应用领域：金融风控

第 9 章：深度学习

课程名称	学习目标
神经网络简介	➔ 学习深度学习与神经网络的基础知识。你也将在教室里亲手用 Python 实现梯度下降法与反向传播。
实现梯度下降	➔ 学习另一个误差函数，并带领你使用 NumPy 矩阵乘法实现梯度下降。
训练神经网络	➔ 你将学习如何训练神经网络，以提升其训练效果。

案例演练：社交网络情绪传播预测

你将把神经网络用于文本数据的情感研究，完全根据评论文本内容将影评归类为正面影评或负面影评。你将从本项目中学习到自然语言处理的常用策略，从数字噪音中挖掘信息。

拓展应用领域：神经网络，建模预测

_____

实战项目：神经网络预测共享单车使用

你将从零开始搭建并训练一个神经网络，并用该网络预测每日自行车租客人数，为某一共享单车预测某一天内需要的使用量，帮助他们作出管理自行车的决策。

第三阶段：应用数据科学（第 21-26 周）

第 9 章：推荐系统与试验设计

课程名称	学习目标
试验设计	➔ 统计数据在现实世界中的应用 ➔ 建立关键指标 ➔ 定义和选择试验条件和测试样本 ➔ SMART 试验模型
A/B 测试	➔ A/B 测试的原理及局限性 ➔ 新奇和新近效应 ➔ 多种分析和比较的技术 ➔ 分析 A/B 试验并写出你的发现
推荐系统中的矩阵分解技术	➔ 了解矩阵分解在机器学习中的应用 ➔ 使用 Python 实现矩阵分解技术 ➔ 解释矩阵分解结果
推荐系统引擎	➔ 使用协同过滤和矩阵分解实现推荐系统引擎 ➔ 推荐系统引擎的常见缺陷，如冷启动 ➔ 应对缺陷的审查模块
部署和应用推荐系统引擎	➔ 使用 flask 和 bootstrap 部署算法

案例演练：星巴克促销活动的优化策略

本案例来自于星巴克公司的雇员候选人的面试题目，数据来自于星巴克公司。企业通常会对广告促销进行测试，以确定某广告是否会带来更多客户购买某些特定的产品。 由于每次促销都要花公司 $0.15，因此最好仅将促销限于那些最容易接受促销的人。 本数据中包含有 7 个特征，你的任务是这七个特征中哪些特征与接收促销有帮助，并优化你的算法。

拓展应用领域：试验设计、产品优化、算法性能优化

案例演练：社交网络用户中的电影推荐系统

假设你是 Netflix 的一名数据分析师，你想要根据用户对不同电影的评分研究用户在电影品位上的相似和不同之处，了解这些评分对用户电影推荐系统是否有帮助。你将探索研究用户电影评分数据集，并学习推荐系统的经典算法——协同过滤。

_____

实战项目：为 IBM 社区构建推荐系统

IBM 有一个在线的数据科学社区，社区成员可以发布教程、文章、数据集以及分析案例。 在此项目中，您将基于用户行为和社交网络数据构建推荐引擎，以表达最可能与用户相关的内容。 您将与 IBM Watson 和 IBM Cloud 一起构建并将推荐系统部署到前端应用程序。

拓展应用领域：基于用户社交信息的推荐系统的构建和前端部署

第 10 章：大数据分析与 Spark

课程名称	学习目标
Spark 简介	➔ 什么是大数据 ➔ 认识常见的大数据生态系统 ➔ 了解何时使用与不使用 Spark
Spark 数据清洗	➔ 使用 Spark SQL 和 Spark Dataframes 操作数据 ➔ 使用 Spark 进行 ETL
调试与优化	➔ 使用 Spark WebUI 进行调试并优化代码
Spark 机器学习	➔ 使用 Spark MLlib 构建机器学习系统

实战项目：大数据分析和预测用户流失

Spotify 是一个正版流媒体音乐服务平台，该平台的用户可以选择付费订阅版或者免费版。用户可以在使用过程中自由升级或选择降级，甚至完全取消服务。你的任务是使用 Spark 框架从一个庞杂的用户行为数据中发现用户的“炸毛”信号，以及时赠送相应的优惠券，保证用户继续使用你的服务。

拓展应用领域：使用 Spark 清理和分析庞大、杂乱的数据集

案例演练：金融服务公司欧唯特的客户分层报告

在业务中，我们时常会收集到用户的人口统计（demographics）数据，这些数据本身会告诉我们大量的信息。在本项目中，您将分析由我们的合作伙伴 Bertelsmann Arvato Analytics 提供的德国某邮购公司客户的人口统计数据，并将其与一般人群的人口统计信息进行比较。您将使用无监督学习技术来执行客户细分，识别最能描述公司核心客户群的人群。这是一个真实的数据挖掘任务，本案例并没有标准答案，数据也尚未预先清理。您可以脱离项目中设定的步骤，自由选择分析数据的方法。

案例演练：星巴克用户促销广告推送优化

星巴克会对使用其移动应用的客户进行不定期促销活动的推送，比如折扣券或者买一送一，某些用户也可能未收到任何优惠。你将结合人口的统计信息、移动 App 上的交易信息，以及促销广告信息进行分析和建模，以确保星巴克在一段时间能获得最大的营业额。注：这个数据集是真正的星巴克应用程序的简化版本，因为本数据集中只有一个产品，而星巴克实际上销售了几十个产品。本案例是一个真实的数据挖掘任务。

第 11 章：软件工程（英/选修）

课程名称	学习目标
软件工程的最佳实践	➔ 编写干净、模块化、记录完备的代码 ➔ 提高效率的代码重构 ➔ 模块测试 (Unit Testing) ➔ 多脚本代码编写
面向对象的编程	➔ 什么是面向对象的编程 ➔ 学习类，并能够编写多个类的程序 ➔ 了解大型模块化 Python 包（如 Pandas 和 Scikit-learn）中的面向对象编程 ➔ 写一个 Python 的包
Web 开发	➔ 了解 Web 应用的组件 ➔ 构建使用 Flask，Plotly 和 Bootstrap 框架的Web应用程序 ➔ Web 应用程序的部署

第 12 章：数据工程（英/选修）

课程名称	学习目标
ETL 管道	➔ 了解什么是 ETL 管道 ➔ 访问和处理 CSV，JSON，日志，API 和数据库中的数据 ➔ 标准化编码和列 ➔ 标准化数据并创建虚拟变量 ➔ 处理异常值，缺失值和重复数据 ➔ 通过计算来设计和实现新功能 ➔ 构建 SQLite 数据库以存储已清理的数据
机器学习管道	➔ 了解使用机器学习管道简化数据准备和建模过程的优势 ➔ 链数据转换 ➔ 使用特征联合实现并行计算并完成更复杂的工作流 ➔ 优化参数的网格搜索管道 ➔ 构建一个完整的机器学习管道
自然语言处理管道	➔ 使用标记化，词典化和删除停用词来准备文本数据以进行分析 ➔ 转换和矢量化文本数据 ➔ 使用词袋和 tf-idf 构建功能 ➔ 使用命名实体识别和词性标注等工具提取功能 ➔ 构建 NLP 模型以执行情绪分析

立即咨询

想知道课程难度是否合适？想获得 1 对 1 学习路径规划？想了解更多课程详情？想获得不定期福利干货分享？

扫描下方二维码，立即咨询您的专属学习规划师