

数萃培训课程 (中级系列)

2017 年 6 月

培训对象:

- 高校数据科学及相关专业教师
 - 企业数据分析师
 - 在校学生 (计算机, 应用数学)
-

数据科学与大数据技术专业 (代码 080910T)

2016 年 2 月, 教育部公布新增“数据科学与大数据技术”本科专业

“数据科学与大数据技术”专业 (专业代码080910T) 强调培养具有多学科交叉能力的大数据人才。该专业重点培养具有扎实的数据科学基础知识和大数据分析能力, 能够从事大数据系统开发、大数据分析、数据可视化等工作的高级专门人才。该专业包括基础课程、核心课程及选修课程三大模块。其中专业基础课程涵盖了数学、统计学、计算机科学等理论知识。培养目标:

“数据科学与大数据技术”专业, 培养德、智、体、美全面发展, 掌握数据科学的基础知识、理论、及技术, 包括面向

大数据技术与应用 (代码 610215)

2016 年 9 月, 教育部公布新增“大数据技术与应用”专科专业

“大数据技术与应用”专业 (专业代码610215) 强调培养具有大数据实践能力的大数据人才。该专业重点培养具有以下能力: 能够从事大数据系统开发、大数据分析、数据可视化等工作的高级专门人才。该专业包括基础课程、核心课程及选修课程三大模块。其中专业基础部分侧重为语言和专业基础方面的课程, 包括《大学计算机基础》、《Python 程序设计》、《数据库系统原理》等课程。培养目标:

“大数据技术与应用”专业, 培养掌握数据科学的基础知识及大数据相关技术, 掌握大数据清洗和分析常用工具的使用

思考

数据从业人员定位?

- 后端 (数据库, 数据清洗)
- 中端 (建模)
- 前端 (Web 展示, 可视化)

如何学/教?

- 初级阶段:
 - 一门主流语言 (R/Python) ⇒
 - 常用统计方法 (回归, 聚类等) ⇒
 - 数据可视化 (R/python 基本图形库与扩展库)⇒
- 中级阶段:

- 统计计算与编程 ⇒
- 高级统计分析 (时间序列, 变量选择与模型选择/评估, 贝叶斯分析) ->
- 数据挖掘/机器学习
- 高级阶段:
 - 深度学习 (文本挖掘, 自然语言处理, 社交网络及图模型, 人工智能)⇒
 - 大数据与高性能计算: 并行计算 -> 分布式计算 (Hadoop/Spark) ⇒ Scala, sparkR, Microsoft R ⇒
 - 大数据平台与开发 (javascript, node.js, gpu 编程, docker 技术)

主要教/学什么?

在校 (统计学) 学生应该主要掌握:

1. 数据库技术 (熟练 MySQL 及不同数据的转换)
2. 一门大数据分析语言 (精通能编程)
3. 大数据常用统计建模方法 (熟练并解释)
4. 大数据常用算法建模方法 (思想与使用)
5. 大数据高性能计算方法 (基本)
6. 大数据分析技术 (了解)

培养/考核技能:

1. R/python 语言基础
2. 统计基础知识
3. ETL 基础
4. 数据收集 (爬虫)
5. 数据可视化
6. 自动化报告
7. R/python 编程与开发
8. 高性能计算
9. 大数据平台使用与高性能计算

高校大数据课程设置

基础核心课程:

1. 大数据导论
2. 统计分析基础
3. 数据库与数据处理
4. 编程基础:R/python

大数据分析必修课:

1. (基于 R) 数据可视化
2. (基于 R) 统计机器学习
3. (基于 R) 编程与高性能计算
4. (基于 R) 大数据平台使用

大数据分析选修课:

1. 基于 python 的大数据统计分析
 2. 贝叶斯分析与应用
 3. 文本数据处理与爬虫技术
 4. 深度学习与应用
-

众创数萃中级大数据分析师培训大纲

基础培训课程 (常年开放)

- A-1. 大数据导论 (Schubert, 林子雨, 周宁奕, 周扬, 黄志敏, 林祯舜, 牟刚)
- A-2. 统计分析基础 (丁辉, 张日权)
- A-3. 数据库与数据处理 (郎大为, 李浩)
- A-4. R 语言基础 (汤银才, 张东, 练勇强)
- A-5. python 语言基础 (肖凯, 靳军)

A-1. 大数据导论

- 课程类型: 大数据分析普及课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 企业高层管理者
 - 数据业务主管
- 要求: 无
- 讲师: 企业数据科学家
 - Schubert, 林子雨, 周宁奕, 周扬, 黄志敏, 林祯舜, 牟刚
- 时间: 1 天/6 小时
- 价格: 5800/2800

References:

- * Data.Science.For.Dummies(2nd Ed)
 - * [dummies.com] (<http://www.dummies.com/programming/big-data/data-science/data-science-for-dummies-cheat->
-

1. 大数据概念
2. 大数据行业动态
3. 大数据案例欣赏
4. 大数据关键技术
5. 大数据分析知识库
6. 大数据人才需求分析

A-2. 统计分析基础

- 课程类型: 大数据分析师初级课程
 - 课程简介:
 - 数据是米，模型是水，想做出好吃的饭，还得用统计思想这把火。统计学是数据分析的灵魂，是互联网+信息化
 - 本课程共二大模块七个章节，分别从一维和二维数据角度讲述数据分析背后深刻的统计思想、基本原理和分析方
 - 本课程是数据分析的基础课程，参与者可以通过学习获得对数据的敏感性、统计模型的熟练性和数据分析结果解
 - 客户获益:
 - 从数据的分布与相互关系上掌握统计建模的诀窍
 - 诊断数据处理与分析存在的问题
 - 获取进一步学习统计、用好统计的
 - 适合人群:
 - 数据分析人员
 - 数据有关管理者
 - 要求: 高等数学
 - 讲师: 丁辉，张日权
 - 课时: 3 天/18 小时
 - 价格: 3000/2400
-

M1: 一维数据的统计分析

1. 数据分析与统计思想

- 概论论与统计学
- 测量/误差与随机/分布
- 概率与计算
- 数据/随机变量与分布
- 独立与相关性

2. 描述性统计分析

- 总体与样本
- 中心趋势的度量
- 离散程度的度量
- 其他特征量: 偏度、峰度、极差、异常值
- 数据的图表展示

3. 统计推断

- 常用的统计分布
- 统计推断的基本问题
- 大数定律与中心极限定理
- 数据、参数与似然函数
- 统计检验、p 值与功效
- 估计的精度与置信区间

4. 常用分布的统计推断

- 正态分布的推断: 估计与拟合
- 区间估计与样本量的确定
- t 分布与 t 检验: 单样本、两样本
- 二项分布的推断: 估计与检验
- 泊松分布的推断: 估计与检验

M2: 二维数据的统计分析

1. 二维数据的统计推断

- 二维数据的图形比较
- 相关性度量与计算
- 二维正态分布与特征量
- 二个正态总体的比较
- 二个比例比较

2. 线性回归

- 模型假设
- 简单线性回归
- 多元性回归
- 数据变换
- 多项式回归
- 回归预测与变量选择

3. 方差分析

- 单因素方差分析
- 单因素协方差分析
- 双因素方差分析
- 重复测量方差分析
- 用回归做方差分析

A-3. 数据库与数据处理

- 课程类型: 大数据分析普及课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 数据业务主管
- 要求: Excel
- 讲师: 郎大为, 李浩
- 课时: 2 天/12 小时
- 价格: 2400/2000

References:

- * 2008. Data Manipulation with R
- * Using SQL in R
- * 2016. Advanced R - Data Programming and the Cloud
- * 2016. Data Wrangling with R

1. 数据基础

- 数据的类型
- 数据的运算
- 结构化与非结构化数据
- 数据库简介
- 数据库操作语言

2. 数据库基础

- SQL 简介

- 常用 SQL 语法
 - 数据表之间的关联
 - SQL 函数与 Group
 - 常用数据库简介
 - 案例: 电商销售数据, SQL 汇总分析
3. R 语言数据获取 (一)
- 读取文件中的数据
 - 读取其他类型的数据
 - R 与数据库的连接
 - 常用数据处理 R 包介绍
 - 案例: 火车时刻表的分析
3. 数据爬虫基础
- 爬虫
 - ETL
4. 基于 R 数据整理
- dplyr 的基本函数
 - 数据汇总, 数据连接
 - dplyr 连接数据库
 - 案例: 使用 dplyr 整理汽车经销商数据

A-4. R 语言基础

- 课程类型: 大数据分析普及课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 数据业务主管
- 要求: Office/EXCEL
- 讲师: 汤银才, 张东, 练勇强
- 课时: 4 天/24 小时
- 价格: 4800/3800

M1: R 入门

1. 大数据与数据科学
- 数据科学与分析工具
 - R 及其优势
 - R 安装与配置
 - R 包安装与使用
 - R 资源与帮助
2. R 快速入门教程
- R 中的基本语法
 - R 中的数据对象及其属性
 - R 的工作空间与管理
 - R 编程基础
 - R 程序调试
3. R 编辑器与 RStudio
- R 常用编辑器

- Rstudio 功能与使用技巧
- Rstudio 进阶
- 项目管理
- Rmarkdown 与报告生成

M2: R 数据集创建与管理

1. 数据集的创建

- 常用数据结构与创建
- 数据的存储
- 数据的读取
- apply 系列函数
- 用于数据的常用函数

2. 数据管理

- 缺失值的处理
- 数据的排序 (sort, order, rank)
- 数据集的合并与子集提取
- 数据表数据的切片、切块与组合

3. R 与其他系统交互

- 与 Excel 的连接 (xlsx 等)
- 数据库操作: sqldf 包简介
- 从关系型数据库 MySQL 读取数据
- 从非关系型数据库 NoSQL 读取数据
- R 与 Java

M3: R 绘图初步

1. 基本的绘图命令

- 大趋势: 信息可视化
- R 绘图基础: 低级与高级绘图命令
- 基本绘图函数: plot, points, lines, curve
- 绘图三要素设置详解 (颜色, 点型, 线型)
- 绘图信息补充 (title, text, legend, axis)

2. 一维数据的可视化

- 常用统计分布与 4 类函数
- 一维离散变量的分布图示
- 一维连续变量的分布图示
- 一维连续分布诊断图
- 非参数密度估计与展示

3. 二维数据的可视化

- 二个离散变量的分布图示
- 二个混合变量的分布图示
- 二个连续变量的分布图示
- 多变量的可视化

M4: R 数据探索与比较分析

1. 数据的描述性统计分析

- 常用描述性统计量及其计算
- 单个连续型变量描述性统计量的获取
- 分组计算描述性统计量

2. 相关性度量

- 变量的类型与转换
- 两个定性变量之间的关联性
- 两个有序变量之间的关联性
- 两个定量变量之间的关联性
- 定性变量与定量变量之间的关联性

3. 相关性检验

- 组间差异比较
 - 独立样本的 t 检验
 - 非独立样本的 t 检验
 - 组间差异的非参数检验
- 分类变量比较
 - 列联表的生成
 - 联合分布、边际分布与条件分布
 - 独立性检验 (卡方检验, Fisher 精确检验, McNemar 检验, Cochran-Mantel-Haenszel 检验)

M5: R 统计建模

1. 回归模型

- lm() 函数中的公式表示
- 一元线性回归
- 多元线性回归
- 回归预测

- 分位数回归

2. 广义线性模型

- 广义线性模型概述
- glm() 函数介绍
- logistic 回归
- Poisson 回归

3. 模型的检验与比较

- 回归模型诊断
- 变量选择
- 模型比较
- 异常值判断
- 预测与交叉验证

A-5. Python 基础

- 课程类型: 大数据分析普及课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 企业高层管理者
 - 数据业务主管
- 要求: Office/EXCEL
- 讲师: 肖凯, 靳军
- 课时: 4 天/24 小时
- 价格: 4800/3000

- Reference

- * 2014. Python Data Analysis
 - * 2014. matplotlib Plotting Cookbook
 - * 2015. Python Data Science Essentials
 - * Regression Analysis with Python
-

- 第1讲：数据分析方法概述及相关工具
 - 认识数据
 - 数据分析的步骤和原则（确定/分解/评估/决策）
 - 相关工具概述及对比（excel/spss/R/matlab/python/Java）
- 第2讲：python 环境和基础语法
 - 安装 anaconda 套件
 - 基础环境 jupyter/ipython
 - 基本数据结构
 - 基本语法
 - 迭代器
 - 函数
 - python 在业界的应用案例分享
- 第3讲：数据操作与计算
 - numpy 数组与操作
 - numpy 统计函数
 - numpy 线性代数
 - 科学计算和最优化 (scipy)
 - 强大灵活的数据结构 pandas DataFrame
- 第4讲：绘图与可视化
 - matplotlib 中绘图
 - pandas 中绘图
 - Charts 中绘图
 - ggplot, seaborn
 - 交互式可视化 bokeh
 - 在线工具 plot.ly
 - 高级应用案例分享
 - 绘制地图
 - 绘制 3D 图形
- 第5讲：统计分析库
 - 概率和统计分析
 - 时间序列分析
 - 简单回归分析 (statsmodels,scikit-learn)
 - 多元回归分析 (statsmodels,scikit-learn)
 - 多项式回归
 - logistic 回归 (statsmodels)

中级培训课程 (定期开放)

- B-1. 数据可视化（谢佳标，王旭，魏鹏）
- B-2. 高级统计分析（丁辉，徐安察，李洪成）
- B-3. 贝叶斯分析与应用（徐安察，汤银才，张东）
- B-4. 统计机器学习（谢佳标，尹志，王旭）
- B-5. 网络爬虫与文本挖掘（尹志，靳军）

B-1. 数据可视化

- 课程类型: 大数据分析师中级课程
 - 课程简介:
 - 客户获益:
 - 适合人群:
 - 数据从业人员
 - 数据业务主管
 - 基于 R 大数据分析与开发的用户
 - 要求: R 基础
 - 讲师: 谢佳标, 王旭, 魏鹏
 - 课时: 6 天/36 小时
 - 价格: 8000/6400
-

M1: 静态绘图的提升

1. R 绘图系统

- R graphics 基本绘图命令 (复习)
- 区域分割与绘图
- 图形输出常见问题处理
- 常用统计在不同系统中的比较

2. 图形的渲染

- rainbow 函数
- 高质量图形渲染库 Cairo
- RColorBrewer 扩展包
- scales: brewer.pal

M2: 常用绘图系统

1. lattice 绘图系统

- 一个简单的 lattice 例子
- 图形参数设置
- 面板函数
- 图形的叠加: 条件变量与条件变量设置
- 其他常用统计图

2. ggplot2 绘图系统

- 最简单的绘图函数: qplot
- ggplot2 的语法: 以散点图为例
- 图层与统计图展示
- ggthemes 主题包介绍

M3: 交互式绘图

1. 交互式绘图包

- rCharts (nPlot, hPlot, mPlot)
- recharts
- plotly
- bokeh/rbokeh

2. 基于 GIS 的动态可视化

M4: 开发与应用

1. shiny 基础

- shiny 简介
- 构建 APP
- 部件设计
- 部署分享 shinyapp

2. shiny 高级开发

- 高级 shiny 技巧
- shinydashboard
- 开发案例

B-2. 高级统计分析

- 课程类型: 大数据分析师中级课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 数据业务主管
 - 基于 R 大数据分析与开发的用户
- 要求: 统计分析基础
- 讲师: 丁辉, 徐安察, 李洪成
- 课时: 3 天/18 小时
- 价格: 4600/3000

M1. 时间序列分析

1. 随机过程与时间序列
 - 时间序列的特征
 - 时间序列的描述性分析
 - 白噪声与平稳性
2. 平稳时间序列模型
 - AR 模型
 - MA 模型
 - ARMA 模型
3. 非平稳序列模型
 - ARIMA 模型
 - 季节效应
4. 异方差模型
 - ARCH 模型
 - GARCH 模型
 - EGARCH 模型
 - SV 模型

M2. 多元统计分析

1. 多变量回归
 - 多变量回归分析
 - 协方差分析
2. 多变量降维
 - 岭回归与 Lasso
 - 主成分分析
3. 多变量分类
 - 分类
 - 判别分析

- 聚类分析

M3. 重抽样方法

1. 随机数据的产生
 - 常用随机数据的产生
 - 重要性抽样
2. 蒙特卡罗方法
 - 蒙特卡罗积分与方差减少技术
 - 自助法
 - jackknife

M4. 最优化方法

1. 极值问题
 - 极大似然估计
 - 一维最优化问题及其求解
 - 多维最优化问题及其求解
 - Laplace 近似
 - EM 算法
2. 其他优化算法
 - 线性规划
 - 遗传算法
 - 图优化

B-3. 贝叶斯分析与应用

- 课程类型: 大数据分析师中级课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 数据业务主管
 - 基于 R 大数据分析与开发的用户
- 要求: R 基础, 统计分析基础
- 讲师: 徐安察, 汤银才, 张东
- 课时: 6 天/36 小时
- 价格: 6800/5500

M1. 贝叶斯分析入门

1. 贝叶斯分析概述
 - 先验信息与提取
 - 贝叶斯公式
 - 后验推断
 - 估计
 - 检验
 - 预测
2. 单参数贝叶斯模型
 - 二项分布: 成功率
 - 正态分布: 均值/方差
 - 指数分布
 - 泊松分布

M2. 贝叶斯分析提高

1. 多参数贝叶斯模型

- 正态分布
 - 多项分布
2. MCMC 方法
 - 贝叶斯计算
 - 抽样方法 (复习)
 - M-H 算法
 - Gibbs 抽样
 - 常用的 M-H 算法
 3. MCMC 软件:
 - BUGS, WinBUGS, OpenBUGS 及在 R 中的实现
 - JAGS 与 rjags, runjags
 - Stan 与 rstan

M3. 实用贝叶斯模型

1. 回归模型
 - 线性模型
 - 最小二乘估计
 - 基于 Jeffreys 先验分析
 - 基于 G-先验分析
2. 广义线性模型
 - logit 模型
 - Probit 模型
 - 对数线性模型
3. 分层贝叶斯模型
 - 分层贝叶斯模型的构建
 - 正态分布
 - logistic 回归模型
4. 混合模型
 - 有限混合
 - EM 解决方法
 - MCMC 解决方法
 - 未知混合个体

M4. 贝叶斯方法的应用

1. 时间序列分析
 - AR 模型
 - MA 模型
 - ARMA 模型
 - 陷马尔可夫模型
2. Capture-Recapture 模型
3. 贝叶斯网络
 - 有向图
 - 贝叶斯网络

B-4. 统计机器学习

- 课程类型: 大数据分析师中级课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 数据业务主管

- 基于 R 大数据分析 & 开发的用户
 - 要求: R 基础, 高维统计分析
 - 讲师: 谢佳标, 尹志, 王旭
 - 课时: 6 天/36 小时
 - 价格: 5800/4500
-

M1: 认识数据挖掘

1. 数据挖掘概述

- 数据挖掘过程
- 数据挖掘对象
- 数据挖掘方法
- 数据挖掘应用
- 无监督与有监督学习

2. 模型评估与选择

- 分类的性能评价
- 混淆矩阵
- 风险图
- ROC 曲线及相关图表 (plotROC, pROC, ROCR)
- 利用 caret 包比较 ROC 曲线
- 交叉验证
- K 折交叉验证基本原理
- 利用 e1071 包完成交叉验证
- 利用 caret 包完成交叉验证

M2: 聚类分析

1. 几类常用的聚类方法

- K-means
- K-medoids
- 系谱聚类
- 密度聚类
- 期望最大化聚类
- 相关 R 包: cluster
- 实例:

2. 隐变量模型

- 概述
- 混合模型
- 隐马尔柯夫模型
- 聚类分析
- 实例:

M3: 分类

1. 判别分析

- 线性判别分析
- 朴素贝叶斯分析
- kNN
- 相关的 R 包: caret
- 实例:

2. 决策树

- 树的构建
- CART
- C4.5
- 相关的 R 包: rpart

- 实例:
- 3. 随机森林
 - 基本原理
 - 相关的 R 包:
 - randomForest, randomForestSRC, ggRandomForests,
 - gbm, glmnet, ranger
 - 可视化分析
- 4. 支持向量机
 - 基本原理
 - 相关的 R 包
 - 可视化分析
- 5. 集成学习
 - Bagging
 - AdaBoost
 - xgboost

B-5. 网络爬虫与文本挖掘

- 课程类型: 大数据分析师中级课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 数据业务主管
 - 大数据分析与应用的用户
- 要求: R/python 基础
- 讲师: 尹志, 靳军
- 课时: 4 天/24 小时
- 价格: 4800/3000

- Reference

- * 2014. XML and Web Technologies for Data Sciences with R(BookZZ.org)
- * 2015. Automated Data Collection with R

M1. 网络爬虫基础

1. 技术准备
 - 网络通信基础
 - HTTP 协议简介
 - Web 开发知识介绍
 - 网站分析知识介绍
2. 开发环境与语言
 - 开发环境安装与使用 (Anaconda 套件与 PyCharm)
 - Python 基础数据结构 (元组/列表/字符串/字典)
 - Python 基础语法 (条件/循环/函数/类/模块)
 - 常用 Python 库使用案例分享
 - Python 技巧与实践分享

M2. 数据爬取与存储

1. 网络爬虫工具库 - 基础 Python 爬虫库 (urllib/Requests) - 认识正则表达式 - “漂亮”的爬虫库-Beautiful Soup - 静态网页爬取案例分享 - Selenium 与 “幻影”浏览器- PhantomJS - Ajax 和 DHTML 网站爬取 - 动态网页爬取案例分享 - 利用 API 进行数据采集
2. 网络爬虫存储

- 文件读取与保存
- 关系数据库存储-MySQL
- 爬虫配合 MySQL 存储案例分享
- 分布式存储-NoSQL 数据库
- 爬虫配合 MongoDB 存储案例分享
- HDFS 简介

M3. 网络爬虫提升

1. 分布式爬虫
 - 多线程爬虫
 - 多进程爬虫
 - 爬虫队列设计
 - 集群化爬取
2. 网络爬虫框架
 - Python 网络爬虫框架介绍
 - Scrapy 基本使用
 - Scrapy 进阶使用
 - 爬虫框架使用案例分享
3. 网络爬虫突破
 - 模拟登录
 - 常见验证码突破
 - 爬虫代理池
 - 各类网页内容处理
 - 爬取移动端 APP 技巧
 - 设计健壮的网络爬虫

M4. 文本挖掘

1. 文本挖掘技术基础
 - 文本挖掘全流程概述
 - 自然语言处理库 (NLTK)
 - TextBlob 文本处理库介绍
 - 中文分词介绍 (jieba)
 - 词云介绍
2. 文本挖掘技术进阶
 - 文本挖掘预处理技术
 - 文本特征处理
 - 文本聚类
 - 主题模型
 - 基于深度学习的文本挖掘
 - 文本挖掘案例分享

拓展培训课程 (不定期开放)

- C-1. 数据治理 (郎大为, 唐力, 李浩)
- C-2. 深度学习 (尹志, 魏鹏, 王旭)
- C-3. R 语言编程与开发 (谢佳标, 王旭)
- C-4. 大数据平台技术与应用 (尹志, 谢佳标, 刘逸铭)
- C-5. Python 大数据分析 (肖凯, 李浩)

C-1. 数据治理

- 课程类型: 大数据分析师高级课程
- 课程简介:

- 客户获益:
 - 适合人群:
 - 数据从业人员
 - 数据业务主管
 - 大数据分析与开发的用户
 - 要求: R/python 基础
 - 讲师: 郎大为, 唐力, 李浩
 - 课时: 3 天/18 小时
 - 价格: 3800/3000
-

C-2. 深度学习

- 课程类型: 大数据分析师高级课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 数据业务主管
 - 大数据分析与开发的用户
- 要求: R/python 基础, 统计机器学习
- 讲师: 尹志, 魏鹏
- 课时: 4 天/24 小时
- 价格: 4800/4000

——~

M1: 基础知识与准备

1. 深度学习背景
 - 什么是深度学习
 - 传统机器学习局限性
 - 深度学习反思
2. 深度学习框架
 - 八大深度学习框架概述
 - tensorflow,caffe 环境的准备
3. 人工神经网络
 - 感知器
 - 激活函数 (sigmoid, ReLu 等对比)
 - 梯度下降算法
 - 反向传播算法介绍与详细推导

M2: 深度神经网络与应用

1. 计算机视觉与卷积神经网络
 - 计算机视觉背景
 - 卷积神经网络训练细节 (卷积操作和池化操作)
 - 图像检测与分割
 - caffe, MXNet 使用
 - 案例 (卷积神经网络): 手写数字识别
 - AlexNet, GoogLeNet 解读
2. 自然语言处理与循环神经网络
 - 自然语言处理背景
 - 循环神经网络训练细节
 - 不同语言翻译 (Translation)
 - caffe, TensorFlow 使用

- 案例 (循环神经网络): 生成手写字符
- 长短时记忆 (LSTM) 训练原理

M3. 无监督学习网络结构

1. 限制性波尔茨曼机
 - 波尔茨曼机网络结构
 - 限制性波尔茨曼机网络结构
 - CD 算法
 - 限制性波尔茨曼机协同过滤上的应用
2. 深度信念网络
 - 贝叶斯网络
 - 表示
 - 推理 (精确推理, 近似推理)
 - 学习 (参数估计与结构学习)
 - 深度信念网络
 - 基本网络结构
 - 参数学习
 - 应用
3. 自动编码器
 - 自动编码器结构与原理
 - 堆叠自动编码器结构与原理
 - 系数自动编码器结构与原理
4. 案例 (无监督学习): 图片数据

C-3. R 语言编程与开发

- 课程类型: 大数据分析师高级课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 数据业务主管
 - 大数据分析 with 开发的用户
- 要求: R 基础, 数据可视化
- 讲师: 谢佳标, 王旭
- 课时: 6 天/36 小时
- 价格: 6800/5200

- References:

- * 2016. Faster! _HigheR! _Stronger!
- * 2016. Efficient R programming
- * 2015. R High Performance Programming
- * 2013. Seamless R and C++ Integration with Rcpp

M1: 编程篇

1. R 编程基础
 - R 中的基本数据结构
 - R 中的控制语句: for, while, repeat
 - 函数的构建与调用
 - 基本的 debugging 方法/函数
 - Scoping rules
2. 提升 R 的性能

- R 的缺陷与克服
 - 向量化编程/函数
 - apply 系列函数:
 apply, lapply, sapply, tapply, mapply, split
 - 其他 Apply 方式
3. R 性能监控
- 监控时间: system.time
 - 本地缓存工具 memoise
 - 性能监控工具 Rprof
 - summaryRprof
 - 性能可视化工具 lineprof

M2: 开发篇

1. R 包开发
- R 包的构成
 - R 包重新编译与安装
 - Rstudio 中开发 R 包
 - 标准化 R 包开发流程
 - 案例: R 语言天气可视化
 - 案例: 每日中国天气
2. 基于 HTMLWidgets 应用与开发
3. 基于 shiny 应用与开发
4. 跨平台通信
- Rserver 与 Java
 - Rsession 与 Java
 - rJava
 - Node.js 与 R

C-4. 大数据平台技术与应用

- 课程类型: 大数据分析师中级课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 数据业务主管
 - 大数据分析与开发的用户
- 要求: R 基础, 数据库基础
- 讲师: 尹志, 谢佳标, 刘逸铭
- 课时: 6 天/36 小时
- 价格: 6800/5200

1. Linux 操作基础

2. Hadoop 基础

- Hadoop 介绍
- hadoop 集群搭建 (centos/ubuntu/linux/mac)
- HDFS 原理
- HDFS Shell 操作实战
- YARN 介绍

3. Hadoop 下的常用工具

- Sqoop
 - Sqoop 介绍与安装
 - Sqoop 基础
 - Sqoop 导入实战
 - Sqoop 增量导入
 - Sqoop 导出实战
 - Sqoop job
- Hive
 - Hive 架构
 - Hive 环境搭建
 - Hive 实战
 - Hive 工作原理
 - 基于 Hive 的日志分析
 - 日志分析的 ETL 自动调度
- HBase
 - HBase 架构
 - HBase 安装
 - HBase 实战
 - Hive 与 HBase 集成实战
- kylin
 - kylin 架构
 - kylin 搭建
 - kylin 实战
 - kylin 性能调优
- spark
 - Spark 架构
 - Spark 工作原理
 - Spark 集群搭建
 - Spark 开发环境

4. Rhadoop/scalaR

5. SparkR

- Spark SQL
- Spark Mlib
- SparkR
- 实战

C-5. Python 大数据分析

- 课程类型: 大数据分析师高级课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 数据业务主管
 - 大数据分析与开发的用户
- 要求: R/python 基础, 统计机器学习
- 讲师: 肖凯, 李浩
- 课时: 4 天/24 小时

- 价格: 5800/4800
 - References
 - 2015. Python Data Science Essentials
 - 2014. Python Data Analysis
 - 2014. Mastering Machine Learning with scikit — learn
-

- 第 1 讲: python 编程与提高 (复习)
 1. 基本语法 (条件/循环/函数/类/模块)
 2. python 语言编程的最佳实践经验
 3. Python 爬虫
 4. 最简单的制作 python 包的方法
- 第 2 讲: 数据挖掘概览
 1. 数据挖掘与模型
 2. 建模流程和步骤
 3. 常见算法
 4. 高级应用案例解析
- 第 3 讲: 典型数据挖掘项目示范
 1. 基于 titanic 数据集预测生存概率
 2. 演示如何进行绘图探索
 3. 特征处理和建模
- 第 4 讲: 降维与异常值处理
 - PCA
 - LFA, LDA, ICA
 - RBM
 - 异常值判别
 - 正则化方法
 - 特征工程
- 第 5 讲: 常用机器学习算法
 - scikit-learn
 - 聚类: k-Means
 - k-近邻
 - 二分类: 基于 logistic 回归
 - 朴素贝叶斯分类器
 - 感知机到 SVM 分类器
 - 感知机到 ANN
 - 决策树
- 第 6 讲: 机器学习进阶
 - 参数调优
 - 集成学习
 - 神经网络
 - 深度学习
- 第 7 讲: 数据挖掘案例讲解
 - 基于 MNIST 数据集识别数字
 - 展示如何进行特征构造
 - 常规机器学习算法和深度学习算法的效果差异
 - 机器学习的结果评估
 - 深度学习的最新进展分享
- 第 8 讲: 现代分析技术的应用
 - 文本挖掘 (NLTK)
 - 社交网络分析 (NLTK)
 - 空间数据的展示与地理信息分析

众创数萃大数据实战训练营

- D-1. 大数据统计分析实战训练营 (R/python)(谢佳标, 李洪成, 张东)
- D-2. 大数据分析师实战训练营 (R)(李舰, 郎大为, 练勇强)
- D-3. 大数据分析师实战训练营 (python)(肖凯, 尹志, 李浩)
- D-4. web 可视化工程师实战训练营 (周宁奕, 刘逸铭)
- D-5. 量化金融分析师实战训练营 (张家齐, 李孟育, 靳军)

D-1. 大数据统计分析实战训练营 (R/python)

参考:

- 纽约数据科学研究院 Data Science Bootcamps
- 约翰霍普金斯大学 Data Science 专项课程
- datasociety

-
- 课程类型: 大数据分析师高级课程
 - 课程简介:
 - 客户获益:
 - 适合人群:
 - 数据从业人员
 - 大数据分析与开发的用户
 - 要求: R/python 基础, 统计基础
 - 讲师: 谢佳标, 李洪成, 张东
 - 课时: 10 天/60 小时
 - 价格: 10800/8800

– Reference:

* 2011. Using R for data management, statistical analysis, and graphics

M1: 数据分析师工具

- 数据处理的常规工具 (数据库, 软件, 展示)
- 训练营主要工具
 - R, RStudio
 - markdown, Rmarkdown
 - python, jupyternotebook
 - git, GitHub

M2: 收集与清洗数据

- 数据存贮系统
- 寻找与获取来源的数据
 - 网络 (web)
 - 社交网络 (微信, 博客)
 - API
 - 数据库 (MySQL)
- 数据清洗与处理

M3: 数据的探索性分析

- 特征量的提取
- 探索数据的分布
- 探索数据之间的关系
- 探索高维数据的信息
- 数据的动态可视化
- 探索数据的空间分布

M4: 数据的建模、预测与验证

- 模型的类型
- 预测流程
- 交叉验证
- 模型的评估
 - 回归模型评估
 - 分类模型评估

M5: 数据的相关性探索

- 相关分析
 - 简单相关关系
 - 自相关分析
 - 偏相关分析
 - 互相关分析
 - 典型相关分析
- 聚类分析
- 关联分析
 - 关联规则挖掘
 - 序列模式挖掘
- 主成分分析

M6: 数据的回归分析与优化

- 简单线性回归
- 多元线性回归
- 回归模型的诊断
 - 残差分析
 - 变量选择
 - 模型比较
 - 多重共线性
 - 离群值检测
- 回归的改进: 正则化
- Logistic 回归

M7: 复杂回归分析

- 梯度提升回归树
- 神经网络
- 支持向量基
- 决策树
 - ID3 算法
 - C4.5/5.0 算法
 - CART 算法
- 集成与随机森林

M8. 时间序列分析

1. 时间序列概述
 - 时间序列的特征

- 时间序列的描述性分析
 - 白噪声与平稳性
2. 平稳时间序列模型: ARMA 模型
 3. 非平稳序列模型: ARIMA 模型
 4. 异方差模型: GARCH 类模型

D-2. 大数据分析师实战训练营 (R)

- 课程类型: 大数据分析师高级课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 大数据分析与开发的用户
- 要求: R 基础, 统计基础
- 讲师: 李舰, 郎大为, 练勇强
- 课时: 10 天/60 小时
- 价格: 10800/8800

- References:

* 2012.Customer and Business Analytics---Applied Data Mining for Business Decision Making Using R

* 2014. R FOR DATA SCIENCE (Dan Toomey)

- 课程目标:
1. 大数据分析
 - 大数据的基本特点
 - 大数据的分析技术
 - 大数据的存储与管理
 - Hadoop、Spark 等工具的应用
 2. R 语言应用
 - R 语言统计分析
 - R 语言可视化应用
 - R 语言数据挖掘与机器学习
 - R 语言的高性能运算

M1. R 基础

1. 数据基础与 R 入门
 - 认识数据
 - 数据分析方法概述
 - 大数据与数据科学概述
 - 常见分析工具概述
 - R 漫谈
 - R 工作环境介绍
 - Rstudio 简介
 - R 常用操作
 - R 在业界应用案例分享
2. R 语言数据操作与编程基础
 - R 基础数据结构
 - 函数操作与函数式编程
 - 数据的读入与写出

- 文件的操作
 - R 与数据库操作
 - 常用函数介绍 (数据操作、字符处理、日期处理)
 - 编程与控制语句 (条件、循环、**apply** 操作)
 - **dplyr** 与数据处理
 - 案例 1: 编写一个模拟排队的函数
 - 案例 2: 制作一个 R 包
 - 案例 3: 电商数据的清洗
 - 案例 4: 网站文本数据的清洗
3. 统计模型
- 线性回归与预测
 - 模型诊断
 - 回归扩展 (非线性模型、**logistic** 模型、**lasso**)
 - 主成分分析和因子分析
 - 聚类分析和判别分析
 - 多维变量的探索
 - 时间序列分析简介
 - 蒙特卡罗方法简介
 - 案例 5: 足球比赛数据分析
 - 案例 6: 销量数据的预测
4. 数据挖掘与机器学习
- 无监督学习介绍
 - 关联规则
 - 案例 7: 零售数据的关联规则挖掘
 - 分类算法: 从 **logistic** 回归说起
 - 机器学习结果评估
 - 多重交叉验证
 - 常用分类方法 (决策树、随机森林、支持向量机)
 - 神经网络与深度学习
 - 案例 8: 足球比赛数据分析的机器学习
5. 数据可视化
- 描述性统计与统计图形介绍
 - R 中图形设备与作图方式
 - 常用图形参数介绍
 - **ggplot2** 介绍
 - 动态可视化示例与业界进展
 - **shiny** 介绍
 - 数据分布的研究 (直方图、QQ 图、热图)
 - 数据关系的探索 (散点图与相关分析, 箱线图与因子分析, 马赛克图与残差表分析)
 - 统计图应用案例及常见误区
 - 案例 9: 地理数据的可视化
 - 案例 10: 使用 **shiny** 开发一个小型动态分析系统
6. 现代分析技术的应用
- 自然语言处理与文本挖掘
 - 案例 11: 网络舆情的文本挖掘
 - 社交网络分析
 - 案例 12: 诗人的社会关系
 - 空间数据的展示与地理信息分析
 - 图像分析技术简介
 - 最优化方法与运筹学简介

D-3. 大数据分析师实战训练营 (python)

- 课程类型: 大数据分析师高级课程
- 课程简介:
- 客户获益:
- 适合人群:
 - 数据从业人员
 - 大数据分析与开发的用户
- 要求: R 基础, 统计基础
- 讲师: 肖凯, 尹志, 李浩
- 课时: 10 天/60 小时
- 价格: 10800/8800
- 课程目标: 掌握 python 语言
 - 统计分析
 - 可视化应用
 - 数据挖掘与机器学习
 - 文本挖掘

M1. python 基础

1. 数据分析方法概述及相关工具
 - 认识数据
 - 数据分析的步骤和原则 (确定/分解/评估/决策)
 - 相关工具概述及对比 (excel/spss/R/matlab/python/Java)
2. python 环境和基础语法
 - 安装 anaconda 套件
 - 基本数据结构 (列表/字符串/字典)
 - 基本语法 (条件/循环/函数/类/模块)
 - 基础环境 linux-shell/IDLE/notebook
 - python 与 ipython
 - python 语言编程的最佳实践经验
 - python 在业界的典型应用

M2. 常用工具库

1. 基础工具库
 - 数值计算 numpy
 - 绘图与可视化 matplotlib 与 Chart
 - 数据操作 pandas
 - 高级应用案例分享
2. 统计分析库
 - 概率和统计分析 (statsmodels)
 - 科学计算和最优化 (scipy)
 - 线性回归和 logistic 回归 (statsmodels)

M3. python 数据挖掘

1. 数据挖掘初步
 - 数据挖掘与模型
 - 建模流程和步骤
 - 常见数据挖掘算法
 - 特征工程

- 正则化方法
 - 主成分分析 (PCA)
 - 高级应用案例解析
2. 数据挖掘算法详解
- 决策树
 - 集成学习
 - 参数调优
 - 感知机
 - 神经网络
 - 深度学习
- M4. 数据挖掘案例讲解
1. titanic 数据集
- 绘图探索: 可视化
 - 特征处理和建模
 - 生存概率预测
2. MNIST 数据集: 数字识别
- 特征构造
 - 常规机器学习算法和深度学习算法的效果差异
 - 机器学习的结果评估
 - 深度学习的最新进展分享

M5. python 爬虫与文本挖掘

1. Python 爬虫
- 数据采集 (HTML 解析, API 使用)
 - 存储数据 (MySQL)
 - 基础爬虫库 (urllib, request)
 - 认识正则表达式 (re 模块)
 - 漂亮的爬虫库 (BeautifulSoup)
 - 方便的现成框架 (Scrapy)
2. 文本挖掘技术基础
- 文本挖掘全流程概述
 - 自然语言处理库 (NLTK)
 - TextBlob 文本处理库介绍
 - 中文分词介绍 (jieba)
 - 词云介绍 (wordcloud)
3. 文本挖掘技术进阶
- 文本挖掘预处理技术
 - 文本特征处理
 - 文本聚类
 - 主题模型
 - 基于深度学习的文本挖掘
 - 文本挖掘案例分享

M6. 现代分析技术 - 社交网络分析 - 空间数据的展示与地理信息分析 - 图像分析技术简介

D-4. web 可视化工程师实战训练营

参考:

- 可视化工程师修炼手册: 跟着我爬虫+数据库+可视化八个快动作, 大数据文摘, 2017-06-04

-
- 课程类型: 大数据分析师高级课程
 - 课程简介:

在大数据时代，最为火爆的技术型岗位当属数据分析师(科学家)、算法工程师和全栈(前端)工程师。全栈(前端)工程师

本课程由五大模板组成，逐步展示前后端web技术栈，

通过对爬虫、数据库、数据清洗、可视化、部署上线等技术链路的技能学习和对地理可视化、**graph**可视化等专类可视

本课程是大数据分析可视化实战训练营，学员将在课程案例中学习与体验，在随堂作业与结业项目中进行网站数据爬取

- 客户获益:
- 适合人群:
 - 数据从业人员
 - 大数据分析与开发的用户
- 要求: 数据库基础, javascript 基础
- 讲师: 周宁奕, 刘铭逸
- 课时: 10 天/60 小时
- 价格: 10800/6800
- 课程目标:
 1. 全面理解前后端 web 技术栈
 2. 全程学习爬虫、数据库、数据清洗、可视化、部署上线等技术链路
 3. 实操体验地理可视化、**graph** 可视化等专门工具

M1: Web 数据可视化概览

1. 可视化的纵与横
 - 数据可视化基础
 - 数据产业链路
 - 技术预备知识
 - Nodejs
 - HTML5 基础
 - SQL
 - 案例:
 - 基于地图的房价分析系统, mapbox
 - 基于图与文本搜索的邮件门分析工具, palantir 与本拉登
2. 服务器与 web 基础
 - web 生态简介
 - 服务器的原理
 - 服务器简介与操作
 - HTML5 基础

M2: 爬虫与数据处理

1. Node.js 爬虫技术
 - 如何获取数据, 几种途径与特点
 - Node.js 爬虫的基本原理、实现方法、应用方式
 - 实战(队列、请求池、模拟登录、ip 代理)
 - 扩展与延伸
 - 案例: 房价数据爬取
2. 关系型数据库 Postgres SQL
 - 数据库选型
 - 导入数据、查询数据、更新数据、表结构与索引
 - SQL 实战
 - 延伸: ORM、客户端与工具

- 案例：
 - 上海房价分析
 - 上海轨道交通分析

M3: 图表可视化

1. D3.js 图表基础

- 使用 D3.js 开发 3 种基础图表
- 开发实时更新的动态基础图表
- 作业：自身图表库的构建、
- 基于 echart 的可视化
- 基于 d3 的通用图表可视化

2. D3.js 图表进阶

- 延伸色彩的算法实现

M4: webGIS 数据可视化

- 地理数据可视化分类
- 基于 D3.js 和 leaflet 的地理可视化
- 用 dat.gui 实现参数控制
- 更多延伸地理投影、色彩搭配
- 案例：
 - 莆田黑医院可视化
 - 房价项目可视化

M5: 图可视化 (graph layout)

- 最简单的乒乓球游戏 (JS/CSS 动画)
- 物理系统的横向扩展：
 - canvas
 - konva.js
 - 盗梦空间
 - 爱舍尔的画与脑洞打开的游戏扩展 (双曲面空间与投影)
- 物理引擎与图布局计算：
 - 泡泡图 circle packing 问题
 - 用 Gephi 优化图可视化算法
- 基于 D3.js 实现力引导图可视化
- 案例：
 - 希拉里邮件门分析
 - 社交网络分析插件

D-5. 量化金融分析师实战训练营

-
- 课程类型: 大数据分析师高级课程
 - 课程简介:
 - 客户获益:
 - 适合人群：
 - 数据从业人员
 - 大数据分析与开发的用户
 - 要求: R/python 基础, 统计基础
 - 讲师: 张家齐, 李孟育, 靳军
 - 课时: 10 天/60 小时
 - 价格: 10800/8800