Python 实现大数据挖掘技术培训

【课程目标】

Python 已经成为数据分析和数据挖掘的首选语言，作为除了 Java、C/C++/C#外最受欢迎的语言。

本课程基于 Python 工具来实现大数据的数据分析和数据挖掘项目。基于业务问题，在数据挖掘标准过程指导下，采用 Python 分析工具，实现数据挖掘项目的每一步操作，从数据预处理、数据建模、数据可视化，到最终数据挖掘结束，帮助学员掌握 Python 用于数据挖掘，提升学员的数据化运营及数据挖掘的能力。

通过本课程的学习，达到如下目的：

1、全面掌握 Python 语言以及其编程思想。

2、掌握常用扩展库的使用，特别是数据挖掘相关库的使用。

3、学会使用 Python 完成数据挖掘项目整个过程。

4、掌握利用 Python 实现可视化呈现。

5、掌握数据挖掘常见算法在 Python 中的实现。

【授课时间】

2-5 天时间

（要根据学员的实际情况调整重点内容及时间）

【授课对象】

业务支持部、IT 系统部、大数据系统开发部、大数据分析中心、网络运维部等相关技术人员。

【学员要求】

课程为实战课程，要求：

1、 每个学员自备一台便携机(必须)。

2、 便携机中事先安装好 Python 3.6 版本及以上。

注：讲师现场提供开源的安装程序、扩展库，以及现场分析的数据源。

【授课方式】

语言基础 + 挖掘模型 + 案例演练 + 开发实践 + 可视化呈现

采用互动式教学，围绕业务问题，展开数据分析过程，全过程演练操作，

让学员在分析、分享、讲授、总结、自我实践过程中获得能力提升。

【课程大纲】

第一部分： Python 语言基础

目的：掌握基本的 Python 编程思想与编程语句，熟悉常用数据结构的操作

1、 Python 简介

2、 开发环境搭建

- Python 的安装
- 扩展库的安装

3、 掌握 Python 的简单数据类型

- 字符串的使用及操作
- 整数、浮点数

4、 掌握基本语句：

- if、while、for、print 等
- 基本运算：
- 函数定义、参数传递、返回值

5、掌握复杂的数据类型：列表/元组

- 列表操作：访问、添加、修改、删除、排序
- 列表切片、复制等
- 列表相关的函数、方法
- 元组的应用

6、复杂数据类型：字典

- 创建、访问、修改、删除、遍历
- 字典函数和方法

7、复杂数据类型：集合

8、掌握面向对象编程思想

- 创建类、继承类
- 模块

9、函数定义、参数传递、返回值

10、 标准库与扩展库的导入

11、 异常处理:try-except 块

演练：基本的 Python 编程语句

第二部分： Python 扩展库

目的：掌握数据集结构及基本处理方法，进一步巩固 Python 语言

1、数据挖掘常用扩展库介绍

- Numpy 数组处理支持
- Scipy 矩阵计算模块
- Matplotlib 数据可视化工具库
- Pandas 数据分析和探索工具
- StatsModels 统计建模库
- Scikit-Learn 机器学习库
- Keras 深度学习（神经网络）库
- Gensim 文本挖掘库

2、数据集读取与操作：读取、写入

- 读写文本文件
- 读写 CSV 文件
- 读写 Excel 文件
- 从数据库获取数据集

3、数据集的核心数据结构（Pandas 数据结构）

- DataFrame 对象及处理方法
- Series 对象及处理方法

演练：用 Python 实现数据的基本统计分析功能

第三部分： 数据可视化处理

目的：掌握作图扩展库，实现数据可视化

1、常用的 Python 作图库

- Matplotlib 库
- Pygal 库

2、实现分类汇总

演练：按性别统计用户人数

演练：按产品+日期统计各产品销售金额

3、各种图形的画法

- 直方图
- 饼图
- 折线图
- 散点图

4、绘图的美化技巧

演练：用 Python 库作图来实现产品销量分析，并可视化

第四部分： 数据理解和数据准备

目的：掌握数据预处理的基本环节，以及 Python 的实现

1、数据预处理

- 异常值处理：3σ 准则，IQR 准则
- 缺失值插补：均值、拉格朗日插补
- 数据筛选/抽样
- 数据的离散化处理
- 变量变换、变量派生

2、数据的基本分析

- 相关分析：原理、公式、应用
- 方差分析：原理、公式、应用
- 卡方分析：原理、公式、应用
- 主成分分析：降维

案例：用 Python 实现数据预处理及数据准备

第五部分： 分类预测模型实战

1、常见分类预测的模型与算法

2、如何评估分类预测模型的质量

- 查准率
- 查全率
- ROC 曲线

3、逻辑回归分析模型

- 逻辑回归的原理
- 逻辑回归建模的步骤
- 逻辑回归结果解读

案例：用 sklearn 库实现银行贷款违约预测

4、决策树模型

- 决策树分类的原理
- 决策树的三个关键问题
- 决策树算法与实现

案例：电力窃漏用户自动识别

5、决策树算法

- 最优属性选择算法：ID3、ID4.0、ID5.0
- 连续变量分割算法
- 树剪枝：预剪枝、后剪枝

6、人工神经网络模型（ANN）

- 神经网络概述
- 神经元工作原理
- 常见神经网络算法（BP、LM、RBF、FNN 等）

案例：神经网络预测产品销量

7、支持向量机（SVM）

- SVM 基本原理
- 维灾难与核心函数

案例：基于水质图像的水质评价

8、贝叶斯分析

- 条件概率
- 常见贝叶斯网络

第六部分： 数值预测模型实战

1、常用数值预测的模型

- 通用预测模型：回归模型
- 季节性预测模型：相加、相乘模型
- 新产品预测模型：珀尔曲线与龚铂兹曲线

2、回归分析概念

3、常见回归分析类别

4、回归分析常见算法

- 梯度上升/下降法
- 普通最小二乘法 OLS
- 局部加权线性回归 LWLR
- 岭回归（RR）
- 套索回归 Lasso
- ElasticNet 回归

第七部分： 聚类分析（客户细分）实战

1、客户细分常用方法

2、聚类分析（Clustering）

- 聚类方法原理介绍及适用场景
- 常用聚类分析算法
- 聚类算法的评价

案例：使用 SKLearn 实现 K 均值聚类

案例：使用 TSNE 实现聚类可视化

3、RFM 模型分析

- RFM 模型，更深入了解你的客户价值
- RFM 模型与市场策略

案例：航空公司客户价值分析

第八部分： 关联规则分析实战

1、关联规则概述

2、常用关联规则算法

- Apriori 算法
  - 发现频繁集
  - 生成关联规则

- FP-Growth 算法
  - 构建 FP 树
  - 提取规则

3、时间序列分析

案例：使用 apriori 库实现关联分析

案例：中医证型关联规则挖掘

第九部分： 案例实战

1、客户流失预测和客户挽留模型

2、银行欠贷风险预测模型

结束：课程总结与问题答疑。