Python 大数据题库说明书 V1.0

一、题库说明

1. 题库简介

本题库基于人民邮电出版社出版的教材《Python 金融数据分析与挖掘实战》、《Python 大数据分析与挖掘实战(微课版)》和配套资源，以及学生课程设计和竞赛作品加工而成。题库覆盖 Python 基础、Numpy、Pandas、Matplotlib、Scikit-learn、关联规则、金融数据挖掘、地理信息数据挖掘、地铁交通数据挖掘、图像识别、文本挖掘、GUI 可视化开发、深度学习、网络爬虫众多领域主题，是 Python 大数据学习的极佳选择，全库 155 道编程题。题库内容包括：**题目、数据和程序、讲解视频三部分，视频陆续在网易云课堂上更新，网校连接如下：**
https://study.163.com/provider/480000002230206/index.htm?share=2&shareId=480000002230206

2. 团队介绍

本题库由“Python 大数据学习吧”项目团队开发，团队由 2 位指导教师和若干位学生组成，旨在建立一个新型的创新学习联盟，实现资源共享、互帮互助、共同进步。为教师提供 Python 大数据分析与挖掘相关教材、同步导学视频课程、教学 PPT、教案、数据和程序、教学大纲、题库等丰富的教学资源；为学生提供丰富的题库视频资源，联合参加学科竞赛、 创新创业、课程设计与毕业设计等互助学习及研讨活动。

诚邀各位教师加盟，加盟教师可利用题库资源联合出版教材、开发课程、项目申报和各类创新教学活动，鼓励加盟教师与学生们在本校成立创新学习团队，加盟到“Python 大数据学习吧”这个大家庭中，一起丰富和完善 Python 大数据题库资源，最终实现产学研用和获得收益。

本团队指导教师 1，黄恒秋，CPDA 项目数据分析师，2011.7-2014.6 就职于深圳市国泰安信息技术有限公司，从事 CSMAR 数据库分析师、软件策划及设计相关工作。2014.9-今，广西民族师范学院数理与电子信息工程学院专任教师，从事数据分析与挖掘、数学建模、Python 语言、MATLAB 语言、高等数学相关课程教学工作。第一作者发表中文核心期刊论文 3 篇，其中 EI 源刊 1 篇。作为第一主编出版教材《Python 金融数据分析与挖掘实战》和《Python 大数据分析与挖掘实战（微课版）》2 部。2019 年组织参加第七届“泰迪杯”数据挖掘挑战赛，获全国一等奖 1 项，二等奖 2 项，三等奖 4 项。2019 年组织参加第一届广西大学生人工智能设计大赛（大数据建模赛道）获一等奖 1 项，二等奖 2 项，三等奖 6 项。2020 年组织参加第二届广西大学生人工智能设计大赛（AI 建模创新赛道）获二等奖 8 项，三等奖 8 项。

本团队指导老师 2，莫洁安，广西民族师范学院数理与电子信息工程学院专

任教师，CPDA 项目数据分析师、大数据分析师（高级），主要研究方向是大数据算法与深度学习、人工智能优化等，涉足领域有金融大数据、图像和文本处理等，发表大数据相关论文 2 篇，主要授课课程是数学建模、数据挖掘与分析、python 爬虫等。分别作为第四主编和第二主编出版教材《Python 金融数据分析与挖掘实战》和《Python 大数据分析与挖掘实战（微课版）》。曾指导全国大学生数学建模竞赛、广西人工智能竞赛等等各类竞赛获得一等奖一项、二等奖 4 项等等。

联系人：陈东民、龙华南

邮箱：1764659690@qq.com、973544910@qq.com

3. 说明书

库名	主题	题号
语法基础	Python 基本数据类型、数据结构、基本数据操作、条件语句、循环语句、函数	1.1~1.10
科学计算	数组定义、数据访问及操作、数组切片、数据存储	2.1~2.6
数据处理	序列及数据框定义、数据访问及操作、切片、分组计算、外部文件读取、大数据文件分块读取、数据关联和合并、数据计算、随机抽样、排序等	3.1~3.11
图像绘制	散点图、线性图、柱状图、直方图、箱线图、饼图、子图的绘制	4.1~4.7
机器学习	缺失值处理、数据规范化、线性回归、逻辑回归、神经网络分类及回归、支持向量机分类及回归、K 均值聚类、主成分分析及应用	5.1~5.11
关联规则	布尔数据集构建、一对一和多对一关联规则挖掘	6.1~6.2
金融挖掘	财务报表及财务指标、股票交易、行业及行业指数相关数据的集成、指数构造、上市公司净利润增长率和股票收益率等基本指标计算	7.1~7.10
金融挖掘	价、量指标的可视化图形绘制	7.11~7.12
金融挖掘	基于总体规模与投资效率指标、基于成长与价值指标的上市公司综合评价	7.13~7.14
金融挖掘	投资组合收益率、波动率的计算	7.15~7.16
金融挖掘	沪深 300 指数预测	7.17
金融挖掘	国际指数关联规则挖掘分析	7.18~7.20
金融挖掘	上市公司盈利能力聚类分析及实证检验	7.21~7.23
金融挖掘	MA、MACD、KDJ、RSI、BIAS、OBV 相关指标的计算，股票涨跌趋势预测模型构建及量化投资实证检验	7.24~7.32
金融挖掘	股票价格关键点提取、形态特征计算、形态聚类、收益率计算，高收益率形态预测及量化投资实证检验	7.33~7.39
金融挖掘	行业联动与轮动关联分析、轮动关联规则挖掘与量化投资模型构建及实证检验	7.40~7.44

金融挖掘	上市公司财务风险预警模型构建	7.45~7.47
金融挖掘	量化择时模型构建	7.48~7.49
金融挖掘	股票联动分析	7.50~7.51
金融挖掘	行业盈利状况可视化分析	7.52~7.53
金融挖掘	上市公司综合能力聚类分析及实证检验	7.54~7.56
金融挖掘	上市公司透明度综合评价	7.57
金融挖掘	上市公司业绩连续高增长预测模型构建	7.58~7.60
金融挖掘	上市公司高送转预测模型构建	7.61~7.64
地理信息	众包任务定价优化方案	8.1~8.7
地理信息	基于 GPS 行车轨迹数据的常规运输线路识别	8.8~8.9
地铁交通	地铁日客流量预测	9.1~9.4
地铁交通	地铁小时客流量预测	9.5~9.7
地铁交通	地铁站点功能性分类研究	9.8~9.11
图像识别	手写体图像识别	10.1~10.2
图像识别	人脸图像识别	10.3~10.4
图像识别	纸币图像识别	10.5~10.6
文本挖掘	上市公司新闻标题情感分类模型构建及应用	11.1~11.6
系统开发	PyQt5 界面应用开发环境配置	12.1
系统开发	纸币面额识别系统	12.2
系统开发	上市公司综合评价系统	12.3
深度学习	深度学习环境搭建	13.1
深度学习	多层神经网络、卷积神经网络和循环神经网络基本用法	13.2~13.3
深度学习	基于卷积神经网络的纸币面额识别模型	13.4
深度学习	基于循环神经网络的上市公司新闻情感分类模型	13.5
网络爬虫	上市公司百度新闻标题数据爬取	14.1
网络爬虫	腾讯视频电影评论数据爬取	14.2
网络爬虫	前程无忧招聘信息数据爬取	14.3
网络爬虫	苏宁易购商品信息爬取	14.4

二、题库内容

1. 语法基础

1.1 现实中的数据主要有数值和文本两种形式，在 Python 中可用数值和字符串两种基本数据类型来定义，请给出例子。
[知识点及要求]基本数据类型

1.2 现实应用中的数据是由多个数值和文本组成，利用 Python 中的简单数据结

构可以实现连续存储，比如列表和元组，为了访问及检索的方便可对存储的值设置唯一的标识键，可通过字典来实现，请给出例子。
[知识点及要求]基本数据结构

1.3 用简单的数据类型（字符串）和基本数据结构（列表、元组、字典）可实现数据的存储，请给出单值、连续多值和不连续多值的访问例子。
[知识点及要求]基本数据结构访问

1.4 建一个空列表 L1，并向列表中添加数值“3”和文本“Python”两个元素，同时定义另外一个列表 L2=[1,2,3,4]，并将 L2 添加到 L1 的后面。
[知识点及要求]列表 append 和 extend 方法

1.5 定义一个嵌套列表 L=[5,[4,'myself'],(1,2,4),'learn']和一个空字典 D，用 for 循环方式将列表 L 中的元素作为值依次填充至字典 D 中，其中标识键用 a,b,c,d 来表示。
[知识点及要求]循环语句 for 和字典 setdefault 方法

1.6 某银行一年定期存款利率为 3%，期末本金和利息一起存入下一个年度，如果现存入 1 万元，需要经过多少年才使得本金和利息达到 1.8 万元。
[知识点及要求]循环语句 while

1.7 编写一段程序代码，利用条件语句实现成绩的分级，其中 90~100 为优秀，80~89 为良好，70~79 为中等，60~69 为及格，0~59 为不及格。
[知识点及要求]条件语句 if

1.8 定义一个函数用来计算长方体的表面积和体积，其中输入参数为长 L，宽 K 和高 H，返回结果为表面积 S 和体积 V。
[知识点及要求]函数定义及应用

1.9 用 for 循环依次打印 2017 年 11 月和 12 月的自然日期，并在控制台下显示出来。
[知识点及要求]字符串连接

1.10 将列表 L=['113.980 22.566', '113.940 22.686', '113.957 22.576', '114.244 22.564 ']中的经纬度字符型数据，按经度和纬度拆分出来并转换为数值类型，分别存储为两个不同的列表 L1 和 L2。
[知识点及要求]字符串拆分和子串查找

2. 科学计算

2.1 对比 Python 的基本数据类型（列表、元组、字典等），数组具有更灵活的数据存储方式，比如一维数组和二维数组或者矩阵，特别是对于数值型数据来说更有优势，根据给出的列表 L1=[1,2,3,4,0.1,7]和嵌套列表 L2= [[1,2,3,4],(5,6,7,8)]，请利用 numpy 包中的 array()函数将其定义为一维数组和二维数据。

[知识点及要求]赋值定义较复杂数据结构：数组

2.2 在编程过程中，预先定义一些数组变量用来保存程序产生的结果是非常必要的，请给出利用 numpy 包中 ones ()、zeros()、arange()函数定义的例子。
[知识点及要求]内嵌函数定义较复杂数据结构：数组

2.3 现有数组 A=np.array([1,3,3.1,4.5])和 B= np.array([[1,2,3,4],(5,6,7,8),[9,10,11,12]])，请给出求解 A 的最大值、最小值、正弦值、余弦值、长度、A 乘 B 和 B 长度的程序
[知识点及要求]数组运算

2.4 现有数组 A=np.array([[1,2,3,4],[5,6,7,8],[9,10,11,12],[13,14,15,16]])，编程实现如下功能：1）将 6、7、14、16 这四个元素顺序切片出来构成一个 2*2 数组；2）以第 0 列元素大于 5 构造逻辑索引记为 I，以 I 为索引切片数组 A 的第 1、3 列元素获得新的数组，记为 B。
[知识点及要求]数组切片

2.5 现有数组 A=np.array([[1,2,3,4],[5,6,7,8],[9,10,11,12],[13,14,15,16]]) 和 B= np.array([1,1,1,1])，请将数组 A 和数组 B 进行水平连接获得新数组 C，即 C 的前 4 列来源于 A，最后一列来源于 B。
[知识点及要求]数组连接

2.6 用 for 循环将以下数据文件读取到 Python 中，并依次进行垂直连接，最终获得完整的数据集，并记为 X。（数据文件见文件夹”2.6”，截图如下）

名称	修改日期	类型	大小
X2010_0331.npy	2020/10/30 10:07	NPY 文件	29 KB
X2010_0630.npy	2020/10/30 8:55	NPY 文件	30 KB
X2010_0930.npy	2020/10/30 8:56	NPY 文件	29 KB
X2010_1231.npy	2020/10/30 8:58	NPY 文件	28 KB
X2011_0331.npy	2020/10/30 9:00	NPY 文件	29 KB
X2014_0331.npy	2020/10/30 10:52	NPY 文件	29 KB
X2014_0630.npy	2020/10/30 10:53	NPY 文件	31 KB

[知识点及要求]数组二进制文件存取

3. 数据处理

3.1 与 Python 基本数据类型（列表、元组、字典等）和主要针对数值数据储存的 numpy 数组相比，pandas 包中提供了支持数值和文本混合数据类型更加有效的存储方式，比如序列和数据框。现有列表 L1=[1,-2,2.3,'hq']、L2=[‘kl’,’ht’,’as’,’km’]和元组 T1=(1,8,8,9)和 T2=(2,4,7,’hp’)，请给出值为 L1，采用默认索引和指定索引（a,b,c,d）两种方式的序列定义方法，以及索引为 a,b,c,d，列名和值分别为 L1、L2、T1、T2 及其值的数据框构造方法。
[知识点及要求]赋值定义较复杂数据结构：序列和数据框

3.2 在实际数据建模应用中，常常需要读取外部数据文件，比如 Excel 文件、TXT 文件和 CSV 文件，请编程实现以下任务：1）请读取“一、车次上车人数统计表.xlsx”中的 sheet2 数据，用一个数据框 df1 来表示；2）请读取文本文件 txt1 中的数据，用一个数据框 df2 来表示；3）大容量文件的读取需要采用分块读取的方式来处理数据，比如 csv 文件常用来存放大容量文件。请采用分块读取的方式读取“data.csv”文件，每次读取 20000 行，读取出来的数据分别用数据框 A1，A2，A3，A4……等来表示。
[知识点及要求]外部数据文件读取：Excel、TXT、Csv。

3.3 序列和数据框作为 pandas 包中两种非常重要的数据结构，同时他们之间也有紧密的联系，数据框可以视为由多个序列组成，它们具有相同的索引，取出数据框中的一列则为序列。在数据处理中，往往是采用不同的数据结构进行相互转化，并利用特定数据结构中的方法计算和处理数据。请读取地铁站点进出站客流数据表（Data.xlsx），完成以下任务：1）取出第 0 列，通过去重的方式获得地铁站点编号的个数；2）采用数据框中的 groupby 分组计算函数，统计出每个站点每天的进站人数和出站人数，计算结果采用一个数据框 df 来表示，其中列标签依次为站点编号、日期、进站人数和出站人数；3）计算出每个站点国庆节期间（10.1~10.7）的进站人数和出站人数。
[知识点及要求]数据框逻辑索引切片和基本切片方法，groupby 分组计算函数应用。

3.4 数据处理过程中经常需要对多个数据集按键进行关联，pandas 包中提供了 merge()函数实现两个数据框之间的关联，包括内连接、左连接和右连接，请根据以下定义的两个字典 dict1 和 dict2，完成如下任务：1）将两个字典转化为数据框；2）对两个数据框给出左连接、右连接和内连接的实现代码，同时简要说明其基本思想。
dict1={'code':['A01','A01','A01','A02','A02','A02','A03','A03'],'month':['01','02','03','01','02','03','01','02'],'price':[10,12,13,15,17,20,10,9]}
dict2={'code':['A01','A01','A01','A02','A02','A02'],'month':['01','02','03','01','02','03'],'vol':[10000,10110,20000,10002,12000,21000]}
[知识点及要求]数据框内连接、左连接和右连接关联操作

3.5 数据处理过程中对数据集进行合并也是经常发生的，pandas 包中提供了 concat()函数实现两个数据框的水平合并和垂直合并，请根据以下定义的三个字典 dict1、dict2 和 dict3，完成如下任务：1）将三个字典转化为数据框 df1、df2、df3；2）df1 和 df2 进行水平合并，合并后的数据框记为 df4；3）df3 和 df4 垂直合并，并修改合并后的 index 为按默认顺序排列。
dict1={'a':[2,2,'kt',6],'b':[4,6,7,8],'c':[6,5,np.nan,6]}
dict2={'d':[8,9,10,11],'e':['p',16,10,8]}
dict3={'a':[1,2],'b':[2,3],'c':[3,4],'d':[4,5],'e':[5,6]}
[知识点及要求]数据框垂直和水平合并操作

3.6 定义列表 L=[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15]，并转化为序列 S，采用序列

中的方法，实现周期为 10 的移动求和、求平均值、求最大值、求最小值的计算。
[知识点及要求]序列移动计算方法应用

3.7 数据处理过程中对数据集的拆分、部分字段或部分数据的提取是一种常见的操作，这种操作有时我们也叫切片，数据框中有两种方式可以实现切片，通过索引实现（iloc）和列标签实现（loc）。请读取地铁站点进出站客流数据表（Data.xlsx），并完成如下任务：1）采用索引实现的方式，获取 135 站点 10 月 1 日~10 月 7 日早上 9~11 点 3 个时刻的进站客流量数据；2）采用列标签实现方式，获取 135 站点 10 月 1 日~10 月 7 日早上 9~11 点 3 个时刻的进站客流量数据。
[知识点及要求]数据框两种常见切片方法

3.8 读取股票交易数据表（data.xlsx），完成如下任务：1）提取 600000.SH 代码交易数据，并按交易日期从小到大进行排序；2）对整个数据表按代码、交易日期从小到大进行排序。
[知识点及要求]数据框排序

3.9 读取地铁站点进出站客流数据表（Data.xlsx），统计计算获得每个站点每个时刻（除去周末和节假日）的总进站客流量和总出站客流量，用一个数据框来表示，列名依次为：站点编号、时刻、总进站客流、总出站客流，并将结果导出到 Excel 表格中，命名为“各站点各时刻进出站客流数据.xlsx”。
[知识点及要求]数据框综合应用案例

3.10 现实生活中抽签是一种比较公平有效的选择或者分配方式，现有 30 个课程设计选题需要分配给 30 个同学，请你写一个程序，实现不重复的随机抽签功能，从而帮助同学们进行抽签。
[知识点及要求]序列及简单随机抽样

3.11 某题库有选择、填空、判断、计算和应用 5 种题型，每种题型题号从 1 开始依次按顺序编号，其中选择题 70 道，填空题 80 道，判断题 50 道，计算题 30 道，应用题 20 道。现有 40 个同学参加考试，要求每个同学从 5 种题型中随机抽取 1 道题目组成试卷，请编程实现给出每个同学试卷的具体题目编号。
[知识点及要求]序列及较复杂抽样

4. 图像绘制

4.1 读取 3.9 获得的“各站点各时刻进出站客流数据.xlsx”，绘制站点 155 各时刻进站客流散点图。
[知识点及要求]散点图绘制

4.2 读取 3.9 获得的“各站点各时刻进出站客流数据.xlsx”，绘制站点 157 各时刻进站客流线形图。
[知识点及要求]线性图及绘制

4.3 读取 3.9 获得的“各站点各时刻进出站客流数据.xlsx”，绘制站点 157 各时刻

进站客流柱状图。
[知识点及要求]柱状图及绘制

4.4 读取 3.9 获得的“各站点各时刻进出站客流数据.xlsx”，绘制站点 157 各时刻进站客流直方图。
[知识点及要求]直方图及绘制

4.5 读取 3.9 获得的“各站点各时刻进出站客流数据.xlsx”，绘制站点 157 各时刻进站客流饼图。
[知识点及要求]饼图及绘制

4.6 读取 3.9 获得的“各站点各时刻进出站客流数据.xlsx”，绘制各站点在 9 时刻进站客流的箱线图。
[知识点及要求]箱线图及绘制

4.7 读取 3.9 获得的“各站点各时刻进出站客流数据.xlsx”，将 155、157、151、123 四个站点在各时刻的进站客流，用一个 2*2 的子图，绘制其线性图。
[知识点及要求]子图及绘制

5. 机器学习

5.1 读取“银行贷款审批数据.xlsx”表，自变量为 x1~x15，决策变量为 y（1-同意贷款，0-不同意贷款），其中 x1~x6 为数值变量，x7~x15 为名义变量，请对 x1~x6 中存在的缺失值用均值策略填充，x7~x15 用最频繁值策略填充。
[知识点及要求]缺失值填充

5.2 在 5.1 基础上，对 x1~x6 数值型变量作均值-方差标准化处理，需要注意的是 x7~x15 名义变量不需要作标准化处理。
[知识点及要求]数据标准化

5.3 在 5.2 基础上，取数据集前 600 条记录作为训练数据，后 90 条记录作为测试数据，构建支持向量机模型，输出其模型准确率和预测准确率。
[知识点及要求]支持向量机分类模型及其应用

5.4 在 5.2 基础上，取数据集前 600 条记录作为训练数据，后 90 条记录作为测试数据，构建逻辑回归模型，输出其模型准确率和预测准确率。
[知识点及要求]逻辑回归模型及其应用

5.5 在 5.2 基础上，取数据集前 600 条记录作为训练数据，后 90 条记录作为测试数据，构建神经网络模型，输出其模型准确率和预测准确率。
[知识点及要求]神经网络分类模型及其应用

5.6 在发电场中电力输出（PE）与 AT（温度）、V（压力）、AP（湿度）、RH（压强）有关，相关测试数据见“发电场数据.xlsx”文件，请完成以下任务：1）利用线

性回归分析命令，求出 PE 与 AT、V、AP、RH 之间的线性回归关系式系数向量（包括常数项）和拟合优度（判定系数），并在命令窗口输出；2）今有某次测试数据 AT=28.4、V=50.6、AP=1011.9、RH=80.54，试利用构建的线性回归模型预测其 PE 值。
[知识点及要求]线性回归模型及其应用

5.7 基于 5.6 的数据集，构建神经网络回归模型，输出其模型准确率，并针对测试数据 AT=28.4、V=50.6、AP=1011.9、RH=80.54，预测其 PE 值。
[知识点及要求]神经网络回归模型及其应用

5.8 基于 5.6 的数据集，构建支持向量机回归模型，输出其拟合优度，并针对测试数据 AT=28.4、V=50.6、AP=1011.9、RH=80.54，预测其 PE 值。（备注：需对采用的核函数进行说明，比如高斯核、线性核、多项式核或者 sigmoid 核等）
[知识点及要求]支持向量机回归模型及其应用

5.9 读取 3.9 获得的“各站点各时刻进出站客流数据.xlsx”，将数据按 id 为站点编号、字段为时刻、值为进站客流构造新的数据变量，如下：

站点编号	7	8	9	10	11	……	24
155 157 ……	500 600 ……						

对以上数据做主成分分析，并写出主成分的表达式及说明其意义
[知识点及要求]主成分提取及分析

5.10 读取“农村居民人均可支配收入来源 2016.xlsx”数据表，其中数据来源于 2016 年《中国统计年鉴》，对表中给出的我国内陆 31 个地区做主成分分析，并基于主成分进行综合排名。
[知识点及要求]基于主成分分析的综合评价

5.11 读取“农村居民人均可支配收入来源 2016.xlsx”数据表，其中数据来源于 2016 年《中国统计年鉴》，对表中给出的我国内陆 31 个地区做 K-均值聚类分析（K=4），并在控制台中输出聚类结果和各个类的聚类中心。
[知识点及要求]K 均值聚类算法及其应用

6. 关联规则

6.1 将以下超市的购买记录：

I1：西红柿、排骨、鸡蛋、毛巾、水果刀、苹果
I2：西红柿、茄子、水果刀、香蕉
I3：鸡蛋、袜子、毛巾、肥皂、苹果、水果刀
I4：西红柿、排骨、茄子、毛巾、水果刀
I5：西红柿、排骨、酸奶、苹果

I6：鸡蛋、茄子、酸奶、肥皂、苹果、香蕉
I7：排骨、鸡蛋、茄子、水果刀、苹果
I8：土豆、鸡蛋、袜子、香蕉、苹果、水果刀
I9：西红柿、排骨、鞋子、土豆、香蕉、苹果

转换为布尔数据集，其中数据集用数据框来表示，数据框中的字段名称即为商品名称，如果商品在某个购买记录中出现用 1 来表示，否则为 0
[知识点及要求]布尔数据集构建

6.2 针对以下布尔数据集，请编程计算规则“A->B”和“A,B->C”的支持度和置信度。

A	B	C
1	1	0
0	1	1
1	0	0
1	1	1
1	1	1
1	0	0
1	1	1
0	1	1
1	0	0
1	1	1
1	1	0
1	1	1
1	1	0

[知识点及要求]基于布尔数据集的一对一和多对一关联规则挖掘

7. 金融挖掘

7.1 现有反映上市公司发展能力、经营能力、现金流量、盈利能力四个方面的财务指标数据表，请以股票代码、会计年度为键，对以上 4 个数据表集成为一个数据表。
[知识点及要求]上市公司财务指标数据关联集成

7.2 从股票日交易数据表中计算其一、二、三、四个季末的收盘价，并添加到 7.1 的集成数据表中。
[知识点及要求]由日频率数据计算季频率数据

7.3 将 7.1 集成数据表中的财务指标，添加到股票日交易数据表中，注意财务指标数据用最近公布的进行填充。
[知识点及要求]交易和财务日、季不同频率周期的数据集成

7.4 计算所有上市公司股票 2015~2018 年的净利润增长率，输出净利润增长率连续 4 年大于 30%的上市公司股票简称。

[知识点及要求]上市公司财务数据处理、简单指标计算及结果整理

7.5 基于 7.4 的数据，按行业分类表中的三级行业名称作为划分标准，计算出每个行业 2016~2018 年的净利润增长率，并输出净利润增长率连续 3 年大于 20%的行业。
[知识点及要求]行业和上市公司数据处理、指标计算及结果整理

7.6 读取互联网和医药板块的个股交易数据，并对停牌的股票用最近交易日填充，构建这两个板块的全样本流通市值加权指数，其中日指数值=日流通市值总和/基准日流通市值总和。
[知识点及要求]股票交易数据处理、基准日期获取、数据填充、指标构造及计算

7.7 对股票交易数据表中停牌数据利用最近交易日数据进行填充。关于股票某个交易日是否停牌，可与交易日历表进行关联查询获得。
[知识点及要求]基于基准日期的股票交易数据填充

7.8 根据 7.7 中填充完整的股票交易数据表，计算各股票的周收益率。
[知识点及要求]股票周最大最小交易日获取及收益率计算

7.9 根据 7.7 中填充完整的股票交易数据表，计算各股票的月收益率。
[知识点及要求]股票月最大最小交易日获取及收益率计算

7.10 根据 7.7 中填充完整的股票交易数据表，计算各股票的年收益率。
[知识点及要求]股票年最大最小交易日期获取及收益率计算

7.11 设计 2*1 子图，绘制股票代码 600000,2017 年 1 月~10 月的每日收盘价、交易量走势图，其中上方子图模块绘制收盘价走势图，下方模块绘制交易量走势图。
[知识点及要求]价、量融合及子图绘制

7.12 根据 7.7 中填充完整的股票交易数据表，在一个图像上绘制股票代码 000001 的每日收盘价和 5、10、20、60、120 收盘价移动平均值走势图
[知识点及要求]移动平均线绘制及指标计算

7.13 请根据以下提供的反映上市公司总体规模与投资效率方面指标数据，按年度对上市公司进行综合评价，输出排名前 20 的上市公司股票简称。

字段名称	字段中文名称	字段说明
B001101000	营业收入	企业经营过程中确认的营业收入。
B001300000	营业利润	与经营业务有关的利润
B001000000	利润总额	公司实现的利润总额
B002000000	净利润	公司实现的净利润
A001000000	资产总计	资产各项目之总计
A001212000	固定资产净额	固定资产原价除去累计折旧和固定资产减值准备之后的净额
F050501B	净资产收益率	净利润／股东权益余额

F091001A	每股净资产	所有者权益合计期末值 / 实收资本期末值
F091301A	每股资本公积	资本公积期末值 / 实收资本期末值；
F090101B	每股收益	净利润本期值 / 实收资本期末值

[知识点及要求]基于总体规模与投资效率指标的上市公司综合评价

7.14 请根据以下提供的反映上市公司成长与价值方面指标数据，按年度对上市公司进行综合评价，输出排名前 20 的上市公司股票简称。

字段名称	字段中文名称	字段说明
F100101B	市盈率	今收盘价当期值 /（净利润上年报值 / 实收资本期末值）
F100301B	市现率	今收盘价当期值 /（经营活动产生的现金流量净额上年报值 / 实收资本期末值）
F100401A	市净率	今收盘价当期值 /（所有者权益合计期末值 /实收资本期末值）
F081001B	净利润增长率	（净利润本年期单季度金额-净利润上一个单季度金额）/净利润上一个单季度金额
F081601B	营业收入增长率	（营业收入本年期单季度金额-营业收入上一个单季度金额）/营业收入上一个单季度金额
F051501B	营业净利率	净利润/营业收入
F050501B	净资产收益率	净利润/股东权益余额
PEG	市盈率增长率	（当前市盈率-上期市盈率）/上期市盈率

[知识点及要求]基于成长与价值指标的上市公司综合评价

7.15 采用 7.13 的综合评价方法，基于 2016 年的财务指标数据进行综合评价，获得排名前 20 的股票代码构建投资组合，以 2017 年 5 月 1 日至 12 月 31 日为持有期，计算投资组合的收益率，并与同期的沪深 300 指数收益率比较。其中收盘价采用考虑现金红利再投资的收盘价可比价（Adjprcwd）进行计算。
[知识点及要求]投资组合构建及收益率计算

7.16 根据 7.8~7.10 的股票周、月、年收益率，计算其波动率，即收益率的标准差。
[知识点及要求]波动率定义及指标计算

7.17 今有沪深 300 指数 2014 年的交易数据，其数据结构如下所示。

Indexcd	Idxtrd01	Idxtrd02	Idxtrd03	Idxtrd04	Idxtrd05	Idxtrd06
000300	2014-01-02	2323.43	2325.99	2310.65	2321.98	451942.9
000300	2014-01-03	2311.97	2314.84	2280.89	2290.78	597826.5
000300	2014-01-06	2286.37	2286.37	2229.33	2238.64	663004
000300	2014-01-07	2222.31	2246.79	2218.65	2238	437531
……	……	……	……	……	……	……

字段依次表示指数代码、交易日期、开盘价、最高价、最低价、收盘价、成交量。
请计算如下指标：
A1（收盘价 / 均价）：即收盘价 / 过去 10 个交易日的移动平均收盘价
A2（现量 / 均量）：即成交量 / 过去 10 个交易日的移动平均成交量

A3（收益率）：(当日收盘价 - 前日收盘价)/ 前日收盘价
A4（最高价 / 均价）：最高价 / 过去 10 个交易日的移动均平均收盘价
A5（最低价 / 均价）：最低价 / 过去 10 个交易日的移动平均收盘价
A6（极差）：最高价 - 最低价（衡量波动性）
A7（瞬时收益）：收盘价 - 开盘价
Y（决策变量）：后交易日收盘价 - 当前交易日收盘价，如果大于 0，记为 1；如果小于等于 0，记为 -1。
同时对指标 A1~A7 作标准化处理：(当前值 - 均值)/ 标准差，最终得到以下标准的数据结构形式：

ID	A1	A2	A3	A4	A5	A6	A7	Y
1								
2								
3								
4								
5								
6								
……								

取后 30 行数据作为测试样本，剩下数据的作为训练样本，利用支持向量机进行训练及测试，输出模型准确率和预测准确率，注意需说明采用的核函数。
[知识点及要求]沪深 300 指数预测影响因素指标定义、计算及支持向量机回归模型应用。

7.18 读取国际指数交易数据表，构建一个字典，键为国际指数代码，值为国际指数涨跌情况序列，其中序列的 index 为交易日期，值为涨跌标识数组，1 表示较前交易日上涨，否则用 0 表示。
[知识点及要求]国际指数涨跌情况字典数据变量构建

7.19 基于 7.18 构建的字典，以沪深 300 指数交易日期为基准，关联其他国际指数，并获得国际指数涨跌情况的布尔数据集。
[知识点及要求]基于基准日期的国际指数涨跌情况布尔数据集构建

7.20 基于 7.19 的布尔数据集，挖掘两两之间的关联规则，其中支持度和置信度要求大于 0.3 和 0.95，并输出其关联规则。
[知识点及要求]基于布尔数据集的一对一关联规则挖掘算法设计及程序实现

7.21 基于 2015 年计算机行业（申银万国标准）上市公司盈利能力数据进行主成分分析（累计贡献率 95％以上），提取其主成分，并写出主成分的表达式和说明主成分的意义。
[知识点及要求]上市公司盈利能力主成分提取

7.22 基于 7.21 提取的主成分进行 K 均值聚类分析，其中 K=4，并输出聚类结果和聚类中心。
[知识点及要求]基于盈利能力主成分的 k 均值聚类

7.23 基于7.22的聚类结果，计算2015年每个类别的上市公司平均净利润增长率，其中每个类别的上市公司平均净利润=该类别的上市公司净利润总和/该类别上市公司的数量。
[知识点及要求]基于盈利能力主成分聚类结果的实证检验

7.24 以中国建筑（股票代码：601668）2017 年的交易数据为例，编写一个程序脚本（非函数形式），计算 5、10、20 日移动平均线 MA 指标值，并输出到 Excel 表格中。其计算公式参考如下：

$$MA_t(n)=\frac{1}{n}C_t+\frac{n-1}{n}MA_{t-1}(n)$$

C_t为第t日股票价格；n为天数，一般取5, 10, 20；t为时间
[知识点及要求]移动平均线指标的计算

7.25 以中国建筑（股票代码：601668）2017 年的交易数据为例，编写一个程序脚本（非函数形式），指数平滑异同平均线 MACD 指标值，并输出到 Excel 表格中。其计算公式参考如下：

$$MACD_t=2\times(DIF_t-DEA_t)$$

$$DIF_t=EMA_t(12)-EMA_t(26)$$

$$DEA_t=\frac{2}{10}DIF_t+\frac{8}{10}DEA_{t-1}$$

$$EMA_t(n)=\frac{2}{n+1}C_t+\frac{n-1}{n+1}EMA_{t-1}(n)$$

[知识点及要求]指数平滑异同平均线指标的计算

7.26 以中国建筑（股票代码：601668）2017 年的交易数据为例，编写一个程序脚本（非函数形式），随机指标 K、D、J 值，并输出到 Excel 表格中。其计算公式参考如下：

$$K_t=\frac{2}{3}K_{t-1}+\frac{1}{3}RSV_t$$

$$D_t=\frac{2}{3}D_{t-1}+\frac{1}{3}K_t$$

$$J_t=3D_t-2K_t$$

$$RSV_t(n)=\frac{C_t-L_n}{H_n-L_n}\times100\%$$

H_n、L_n分别表示n日内最高收盘价和最低收盘价, n=9

[知识点及要求]随机指标 K、D、J 的计算

7.27 以中国建筑（股票代码：601668）2017 年的交易数据为例，编写一个程序脚本（非函数形式），计算 6、12、24 天相对强弱指标 RSI 值，并输出到 Excel 表

格中。其计算公式参考如下：

$$RSI_t(n)=\frac{A}{A-B}\times 100\%$$

公式中，$A=n$ 日内收盘涨数；$B=n$ 日内收盘跌数；$n=6,12,24$

[知识点及要求]相对强弱 RSI 指标的计算。

7.28 以中国建筑（股票代码：601668）2017 年的交易数据为例，编写一个程序脚本（非函数形式），计算 5、10、20 天乖离率指标 BIAS 值，并输出到 Excel 表格中。其计算公式参考如下：

$$\text{乖离率}=\frac{\text{当日收盘价}-n\text{日平均价}}{\text{n日平均价}}\times 100\%, n=5,10,20$$

[知识点及要求]乖离率指标 BIAS 的计算。

7.29 以中国建筑（股票代码：601668）2017 年的交易数据为例，编写一个程序脚本（非函数形式），计算能量潮 OBV 指标值，并输出到 Excel 表格中。其计算公式参考如下：

$$\text{今日}OBV=\text{前一日}OBV+\text{sgn}\times\text{今日的成交量}$$

其中，sgn 是符号函数，其数值由下面的式子决定：

若今日收盘价≥昨日收盘价，$\text{sgn}=+1$

若今日收盘价<昨日收盘价，$\text{sgn}=-1$

[知识点及要求]能量潮指标 OBV 的计算

7.30 以中国建筑（股票代码：601668）2017 年的交易数据为例，编写一个程序脚本（非函数形式），计算其涨跌趋势指标值。其方法如下：下一日收盘价减去当日收盘价，若大于 0，则下日股价呈现上涨趋势，用 1 表示，反之则股价呈现下跌趋势，用-1 表示。

[知识点及要求]涨跌趋势指标的定义及计算

7.31 以 7.24~7.29 计算的指标作为自变量（X），7.30 计算的指标作为因变量（Y），训练数据时间段为 2017 年 1 月至 11 月，测试数据时间段为 2017 年 12 月，构建支持向量机分类模型，输出其模型准确率和预测准确率。

[知识点及要求]基于技术指标的股票价格涨跌趋势预测模型构建（支持向量机模型）

7.32 以 7.24 数据表中所有的股票代码构建投资组合，对每只股票代码每年的交易数据计算 7.24~7.30 中的指标值，当年 1 月~11 月份为训练数据，并对 12 月份的数据做出预测，如果预测结果为 1，则设计一个量化投资策略：以当天收盘价

买入，下一个交易日卖出，计算投资收益率，最终获得投资组合的收益率（投资组合中所有股票投资收益率之和）。
[知识点及要求]基于股票价格涨跌趋势预测模型的量化投资策略设计与实现。

7.33 股票价格走势主要由一些关键性价格点构成，提取其关键价格点，可以有效的分析其形态特征。一般地，连续三个价格序列 P(i-1)、P(i)、P(i+1)，|P(i)-(P(i-1)+P(i+1))/2|值越大，P(i)成为关键点的可能性就大。请获取上汽集团（股票代码：600104）2017 年 4 月 1 日~7 月 31 日的交易数据，提取其关键价格点 12 个（包括原始价格序列的第一个和最后一个价格点），并通过可视化方式将原始价格和关键点价格走势图绘制出来。
[知识点及要求]股票价格关键点的定义、提取、计算及可视化。

7.34 根据 7.33 提取的关键价格点，构造股票价格走势形态特征，构造方法如下：
1）首先计算两个关键价格点之间的正切值（斜率），公式如下：

$$\tan\theta = \frac{p_2 - p_1}{x_2 - x_1}$$

其中 p_1 和 p_2 分别表示前后两个关键点，x_1 和 x_2 为关键点对应的下标。

2）涨跌幅划分：

上涨幅度大：	tan 值>0.5
上涨幅度较大：	tan 值介于 0.2~0.5 之间
上涨：	tan 值介于 0.1~0.2 之间
平缓：	tan 值介于-0.1~0.1 之间
下跌：	tan 值介于-0.2~-0.1 之间
下跌幅度较大：	tan 值介于-0.5~-0.2 之间
下跌幅度大：	tan 值<-0.5

3）股票价格走势形态特征表示，即将以上涨跌幅标签化为 7、6、5、4、3、2、1。
[知识点及要求]基于股票价格关键点的形态特征定义及计算

7.35 利用 7.13 基于总体规模与投资效率指标的综合评价方法，取 2016 年的财务指标数据进行综合评价，获得排名前 400 的股票作为研究样本，然后根据 7.34 的方法，提取这 400 只股票 2017 年 5 月 1 日~8 月 31 日和 2017 年 6 月 1 日~9 月 30 日两个计算周期的价格形态特征数据，同时作垂直合并形成完整的价格形态特征数据集，在其计算过程中也把关键价格点数据归一化并保存下来，以便后面使用。
[知识点及要求]投资组合的构建及形态特征数据的计算

7.36 对 7.35 提取的完整价格形态特征数据作 K 最频繁值聚类分析，K=20，即将 400 只股票样本聚为 20 类，并对每类股票以特征计算周期之后一个月为持有期，计算每类股票平均收益率。每类股票平均收益率=该类所有股票收益率之和/该类股票总数。
[知识点及要求]投资组合股票形态聚类及类平均收益计算。

7.37 对 7.36 的聚类结果转化为两类，其中平均收益率排名前 5 的类其股票记为 1 类，其余的为-1 类。以股票价格形态特征指标数据为自变量 X，转化后的两类为因变量 Y，并作为训练数据集构建支持向量机模型，输出其模型准确率。
[知识点及要求]训练数据集的构建

7.38 基于 7.35 的 400 只股票样本，提取 2017 年 6 月 1 日~9 月 30 日的价格形态特征指标数据，作为测试数据的自变量 X1，利用 7.37 构建的支持向量机模型对其进行预测，如果预测结果为 1，则表示该只股票出现的价格形态属于高收益类别，即在未来一个月内投资可能获得较高的收益。故设计一个量化投资策略：如果预测结果为 1，则以下一个月期初收盘价买入，期末收盘价卖出，计算该股票的收益率。所有预测结果为 1 的股票投资收益率之和即为该投资组合的收益率。为了评价该投资组合收益率的优劣，计算同期沪深 300 指数收益率作为评价基准。
[知识点及要求]测试数据集的构建及量化投资策略设计实现。

7.39 基于 7.35 归一化后的关键点价格数据和 7.36 计算的股票类平均收益率（未来一个月作为持有期的收益率），绘制每类股票价格形态走势和未来一个月作为持有期的收益率可视化图形，以便直观地观察哪些是对投资有价值的形态。为了提高可读性，每类股票价格形态绘制其前面 5 只股票即可。
[知识点及要求]股票价格形态与收益率可视化分析

7.40 现有 2010 年 1 月 4 日至 2017 年 3 月 7 日申银万国 34 个行业指数交易数据，共 1741 个交易日，其中部分指数不满足 1741 个交易日，请将其剔除。在此基础上，首先定义一个数据框 D，键为指数代码，值为指数的日涨跌标识数组，其中当日指数较前一个交易日上涨记为 1，否则为 0；其次，将字典 D 转换为数据框，index 为对应的交易日期。最终获得了指数日涨跌情况的布尔型数据集。
[知识点及要求]指数日涨跌情况布尔数据集构建

7.41 现有 2010 年 1 月 4 日至 2017 年 3 月 7 日申银万国 34 个行业指数交易数据，共 1741 个交易日，其中部分指数不满足 1741 个交易日，请将其剔除。在此基础上，首先利用交易日历数据表找出每周的最小交易日和最大交易日，其次定义一个数据框 D，键为指数代码，值为指数的周涨跌标识数组，其中当周最大交易日指数较最小交易日指数上涨记为 1，否则为 0；再次，将字典 D 转换为数据框，index 采用默认序号即可，表示周次。最终获得了指数周涨跌情况的布尔型数据集。
[知识点及要求]指数周涨跌情况布尔数据集构建。

7.42 现有 2010 年 1 月 4 日至 2017 年 3 月 7 日申银万国 34 个行业指数交易数据，共 1741 个交易日，其中部分指数不满足 1741 个交易日，请将其剔除。在此基础上，首先利用交易日历数据表找出每月的最小交易日和最大交易日，其次定义一个数据框 D，键为指数代码，值为指数的月涨跌标识数组，其中当月最大交易日指数较最小交易日指数上涨记为 1，否则为 0；再次，将字典 D 转换为数据框，index 采用默认序号即可，表示月次。最终获得了指数月涨跌情况的布尔型数据集。

[知识点及要求]指数月涨跌情况布尔数据集构建

7.43 对 7.40~7.42 的日、周、月指数涨跌情况布尔型数据集进行行业轮动关联规则挖掘。所谓轮动，是指当期（日、周、月）指数上涨，会引起下期（日、周、月）也上涨的现象。请输出支持度大于 0.3、置信度大于 0.7 的关联规则。
[知识点及要求]行业轮动关联规则挖掘

7.44 根据输出的关联规则，设计量化投资策略。比如有满足条件的轮动关联规则 **A 行业->B 行业**，则对该规则设计以下投资策略：如果当期 A 行业指数上涨，则取 B 行业排名前 20 的股票构建投资组合，对投资组合中的股票以下期期初收盘价买入，下期期末收盘价卖出，计算股票投资收益率，投资组合中所有股票投资收益率即为该投资组合的收益率。（注：由于行业指数不能交易，所以取行业排名前 20 的股票构建投资组合近似代替指数，排名方法为 7.13 基于总体规模与投资效率指标的综合评价方法，指标数据取最近年度数据）
[知识点及要求]基于行业轮动关联规则挖掘的量化投资策略设计与实现。

7.45 企业财务风险预警是企业风险预警系统的一个重要组成部分。企业财务风险判断标准如下：1）连续两年年报显示净利润为负值；2）当年净资产收益率或者总资产净利润率为负值，满足条件之一即为风险企业，记为 1，否则为非风险企业（两个条件同时不满足），记为 0。其财务特征变量如下：流动比率、速动比率、现金比率、产权比率、利息保障倍数、盈利现金比率、总资产报酬率、净资产收益率、存货周转率、应收账款周转率、总资产周转率、主营业务鲜明率、资本保值增值率、净资产增长率，依次表示为 x1~x14。其中：
现金比率= 货币资金÷ 流动负债
盈利现金比率= 经营活动的现金净流量÷ 净利润
主营业务鲜明程度= 主营业务利润÷ |净利润|
请根据提供的数据，筛选出风险企业及财务特征变量（自变量 X），同时构造风险标识变量（因变量 Y=1），用一个数据框表示，记为 A1.
[知识点及要求]财务风险特征指标和标识指标数据的构造

7.46 基于 7.45 的数据，随机筛选出与风险企业数量相同的非风险企业，包括财务特征变量（自变量 X），同时构造非风险标识变量（因变量 Y=0），用一个数据框表示，记为 A0.
[知识点及要求]随机抽样、非财务风险特征指标和标识指标数据的构造

7.47 对 7.45 和 7.46 的 A1 和 A0，进行垂直合并，获得完整的数据集 A，用数据框来表示，同时修改 index 为默认序号。对数据集 A 按 80%训练，20%测试，随机划分训练集和测试集，构建支持向量机模型，输出模型的准确率和预测准确率。
[知识点及要求]风险和非风险企业数据合并、划分训练和测试集、模型构建

7.48 读取沪深 300 指数 2016 年~2017 年的交易数据表和交易日历数据表，并计算以下指标：
A1 周最高价：周内指数成交的最高价。
A2 周最低价：周内指数成交的最低价。

A3 成交额：一周内指数成交额。
A4 周收益率：（本周收盘价-上周收盘价）/上周收盘价。
A5 上周收益率：上一周的周收益率。
A6 前周收益率：上上周的周收益率。
A7 上周成交额：上一周的成交额。
A8 近四周平均成交额：在最近的四周内的平均成交额。
Y（因变量）：下周收盘价 – 本周收盘价，如果大于 0，记为 1；如果小于等于 0，记为 -1。
（提示：本题需先根据交易日历表，找出每周的最小交易日和最大交易日，进而计算以上周频率指标）
[知识点及要求]周最大最小交易日提取、周频率指标的计算

7.49 基于 7.48 计算的 A1~A8（自变量 X）和 Y（因变量）指标数据，按相同的周次作对齐处理，同时按 80%训练，20%测试，随机划分训练集和测试集，构建支持向量机模型，输出模型的准确率和预测准确率。
[知识点及要求]基于周频率数据的量化择时模型

7.50 基于 7.15 的股票交易数据（trd_2017.xlsx）和财务指标数据（data.xlsx）。首先利用 7.13 中总体规模与投资效率指标综合评价方法，获得排名前 30 的股票作为研究样本。其次以其中一个股票代码的交易日期为基准，关联其他股票交易数据，使得所有股票交易数据具有相同的交易日期，其数据集记为 Data。
[知识点及要求]股票投资组合构建、基准日期提取及数据处理

7.51 基于 7.50 的 Data，定义一个字典 D，键为股票代码，值为每只股票的涨跌指标数组，并将字典 D 转化为数据框（即布尔型数据集），数据框的 index 为交易日期。其中涨跌指标=当日收盘价-前日收盘价，如果大于 0，表示上涨，记为 1，否则记为 0。基于获得的布尔型数据集，挖掘两两股票之间的关联规则，并将置信度最大的 20 条关联规则输出到 Excel 表格中。
[知识点及要求]股票投资组合中股票涨跌情况布尔数据集构建及联动关联规则挖掘。

7.52 读取申银万国行业分类表和公司净利润数据表，计算获得每个行业 2012~2017 年（共 6 个年度）的净利润增长率，并输出每年净利润增长率最大的 8 个行业及增长率数据到 Excel 表格中。
[知识点及要求]行业净利润增长率的计算

7.53 基于 7.52 的数据，用 3*2 子图，采用直方图绘制出每年净利润增长率最大的 8 个行业，子图的横轴为行业名称，纵轴为净利润增长率，标题为对应年份。
[知识点及要求]行业净利润增长率可视化分析

7.54 提取上市公司 2015 年的总体规模与投资效率指标数据，指标含义如下：

字段名称	字段中文名称	字段说明
B001101000	营业收入	企业经营过程中确认的营业收入。
B001300000	营业利润	与经营业务有关的利润

B001000000	利润总额	公司实现的利润总额
B002000000	净利润	公司实现的净利润
A001000000	资产总计	资产各项目之总计
A001212000	固定资产净额	固定资产原价除去累计折旧和固定资产减值准备之后的净额
F050501B	净资产收益率	净利润 / 股东权益余额
F091001A	每股净资产	所有者权益合计期末值 / 实收资本期末值
F091301A	每股资本公积	资本公积期末值 / 实收资本期末值；
F090101B	每股收益	净利润本期值 / 实收资本期末值

对指标做主成分分析（要求累计贡献率在 95%以上），并写出主成分的表达式和说明其意义。
[知识点及要求]上市公司综合竞争力主成分分析

7.55 基于 7.54 提取的主成分，采用极差法进行标准化处理，并对标准化后的数据进行 K-均值聚类分析，并输出其聚类中心。
[知识点及要求]基于主成分的上市公司综合竞争力聚类分析

7.56 基于 7.55 的聚类结果，每一类股票作为投资组合，计算持有期为 2016-05-01~2016-12-31 的总收益率。其中股票收益率=（持有期最大交易日的收盘价-持有期最小交易日的收盘价）/持有期最小交易日的收盘价。投资组合收益率=投资组合中各股票收益率之和。收盘价采用考虑现金红利再投资的收盘价可比价进行计算。
[知识点及要求]上市公司综合竞争力聚类分析结果实证检验

7.57 上市公司的透明度可以从经营状况，股权结构和治理结构方面进行综合评价，选取的指标如下：
总资产净利润率（X1）： 净利润 / 总资产余额
速动比率（X2）：（流动资产－存货） / 流动负债。
总资产增长率 （X3）：（资产总计本期期末值—资产总计本期期初值） /（资产总计本期期初值）。
两权分离度（X4）：控制权与所有权之间的差值。
实际控制人性质（X5）：实际控制人性质分为国有企业、民营企业、自然人、非自然人、组织或个人等。
高管持股数量（X6）：高级管理人员持有股数量，包括兼任情况。
董事长与总经理兼任情况（X8）：1=同一人； 2=不同一人
董事会人数（规模）（X9）：董事会中董事总人数，包括独立董事。
现有工业企业上市公司 34 家，对指标数据作主成分分析，并基于提取的主成分进行综合排名（要求提取的主成分累计贡献率在 95%以上），并将排名结果输出到 Excel 表格中。（注：排名结果包括股票简称和得分）。
[知识点及要求]基于财务与治理指标的上市公司透明度综合评价

7.58 现有归属于母公司所有者的净利润，每股收益、每股经营活动净现金流，营业毛利率，应收账款周转率，4 个数据表 2008 年~2018 年的年度报告数据。请计算上市公司 2011 年~2018 年共 8 年的归属母公司所有者净利润增长率数据，

同时在计算过程中把2011~2014年连续4年归属于母公司所有者净利润增长率大于30%且每年基本每股收益大于0的股票记为+1类，否则记为0类，作为训练数据集的因变量（Y）供后面使用。同理计算2015~2018年，作为测试数据集的因变量供后面使用。
[知识点及要求]上市公司业绩连续高增长因变量的构造

7.59 在7.58基础上，计算获得因变量（Y）对应的自变量（X）值。对于训练数据集，计算对应股票2011~2014年归属母公司所有者净利润增长率、每股经营活动净现金流、营业毛利率、应收账款周转率4个指标数据的平均值作为X。同理针对测试数据集，计算对应股票2015~2018年的自变量（X）值。
[知识点及要求]上市公司业绩连续高增长自变量的构造

7.60 基于7.58~7.59获得的训练数据集和测试数据集，包括因变量和自变量，构建支持向量机分类模型（注意类平衡策略），并对测试数据进行分类预测，最终输出模型的准确率和预测准确率。
[知识点及要求]上市公司业绩连续高增长预测模型的构建。

7.61 基于2018年日交易数据表，设计一个简单算法，用于检测上市公司在2017年度报告中是否有“高送转”行为，同时获得上市公司高送转情况数据，其中有高送转行为的记为1，否则为0，作为因变量Y。
[知识点及要求]基于交易数据检测上市公司高送转行为。

7.62 基于2017年三季度的数据，探索影响年度高送转行为因素指标，包括每股收益、每股公积金、每股净资产、每股未分配利润、净利润增长率、股价、流通股本、总股本，请根据提供的每股指标数据表、净利润增长率数据表、月交易数据表，计算获得以上8个指标数据，作为自变量X。
[知识点及要求]利用三季度数据，构建影响年度高送转行为因素指标。
7.63 根据7.61~7.62，将自变量数据X和因变量数据Y，关联合并集成至一个数据集中。同时，由于高送转公司和非高送转公司这两类数量差异较大，请对非高送转公司进行随机抽样，使得其数量与高送转公司数量一致。
[知识点及要求]数据关联集成、不平衡类数据的随机抽样

7.64 基于7.63的数据，按80%训练，20%测试，划分训练集和测试集，构建支持向量机模型、逻辑回归模型，输出模型的准确率和测试样本的预测准确率。
[知识点及要求]利用当年三季度数据作为模型输入，当年年报是否高送转作为输出，构建高送转预测模型。

8. 地理信息

8.1 安装地理信息可视化包folium，并读取“附件1：已完成任务数据.xlsx”和“附件2：会员信息数据.xlsx”数据表，将任务和会员的位置在地图上标注出来，以观察任务和会员是否在同一个区域上。
[知识点及要求] 地理信息可视化包folium及其应用

8.2 针对每个任务，计算其在 5 公里区域范围内的任务数量、任务平均定价、会员数量、会员平均信誉值、会员可预订任务总量，共 5 个指标数据。
[知识点及要求] 基于经纬度地理坐标数据的指标计算

8.3 针对每个任务，计算其在 5 公里区域范围内在 6:30、6:33~6:45、6:48~7:03、7:06~7:21、7:24~7:39、7:42~7:57、8:00，每个时段的会员可预订任务量，共 7 个指标数据。
[知识点及要求] 基于经纬度地理坐标数据的指标计算。

8.4 将 8.2 和 8.3 计算的指标数据水平集成，获得各任务 12 个指标完整数据集，同时将该数据集分成两部分：一部分是已完成的任务数据，另一部分为未完成的任务数据。已完成的任务假设定价是合理的，能被广大会员接受。未完成的任务假设定价不合理，需要重新进行定价。请以完成的任务 12 个指标数据作为自变量 X，其定价作为因变量 Y，构建定价模型，并以未完成的任务 12 个指标数据作为定价模型输入变量，输出即为重新定价结果。
[知识点及要求] 任务定价模型的构建

8.5 以 8.4 各任务 12 个指标完整数据集和附件一定价数据作为自变量 X，任务完成情况作为因变量 Y，构建分类模型。以未完成的任务 12 个指标数据和重新定价结果作为模型输入，输出即为重新定价后任务的完成情况。
[知识点及要求] 任务定价模型的应用

8.6 对未完成的任务重新定价效果从两个方面评估：成本增加额和新增任务完成量。请给出具体程序计算实现，其中成本增加额=新定价总和-原定价总和，新增任务完成量=8.5 模型输出为 1 的数量*模型准确率。
[知识点及要求] 任务定价调整方案评价

8.7 请对附件 3 的新任务给出定价，并对其是否能完成给出评估。
[知识点及要求] 任务定价调整方案应用

8.8 请对给出的 10 辆车 GPS 行车轨迹数据，找出其常规经停地点。
[知识点及要求] 基于 GPS 行车轨迹数据的常规经停地点识别

8.9 基于 8.8 的常规经停地点，划分行车线路，并绘制出其常规行车线路图。
[知识点及要求] 基于 GPS 行车轨迹数据的常规运输线路图绘制

9. 地铁交通

9.1 现有某城市 2015 年 8 月至 11 月的地铁刷卡数据，每个月的数据已按日期时间从小到大排序好，CSV 格式，文件较大。请通过分块读取的方式，基于二分法查找思想，设计一个高效算法，快速找出每天最后一条刷卡记录序号。
[知识点及要求]大数据文件分块读取技术、二分法查找

9.2 读取数据前 1 万条记录，通过去重方式，获得所有地铁站点编号，基于 9.1

结果，统计获得每个地铁站点每天的客流量数据。注意客流量=进站客流量+出站客流量。
[知识点及要求]序列去重方法、数据框切片及计算

9.3 探索影响地铁客流量的因素，包括天气（温度、雨雪、风速等）、是否周末、是否节假日等。请根据提供的数据，设计并计算相关指标。
[知识点及要求]天气、周末、节假日数据指标设计及计算

9.4 以 9.3 计算的指标作为自变量 X，9.2 计算的客流量数据作为因变量 Y，构建地铁日客流量预测模型。其中以 8、9、10 月及 11 月 1 日~23 日数据作为训练样本，11 月 24 日~30 日作为测试样本。
[知识点及要求]地铁日客流量预测模型构建

9.5 利用 111 月份的地铁刷卡数据，基于分块读取技术和二分法查找思想，设计一个算法，找出每个地铁站点每天 6:00~23:00，间隔 1 小时为一个统计时间段的最后一条刷卡记录序号。
[知识点及要求]分块读取技术和二分法查找思想进一步拓展到小时频度数据处理

9.6 基于 9.5 的结果，统计获得每个地铁站点每天各个统计时间段的客流量。
[知识点及要求]数据框切片及计算

9.7 基于 9.3 计算的指标和各时段序号作为自变量 X，9.6 的结果为因变量 Y，构建每天各统计时间段的客流量预测模型。其中 11 月前 23 天的数据为训练数据，后 7 天的为测试数据。
[知识点及要求]移动平均计算、小时为频率的客流量预测模型构建

9.8 基于 9.5 的结果，统计获得每个地铁站点每天各个统计时间段的进站客流量。
[知识点及要求]数据框切片及计算

9.9 基于 9.8 的结果，统计出每个站点各统计时间段的平均进站客流量，其中周末和节假日不统计。
[知识点及要求]数据框切片、分组统计

9.10 基于 9.9 的结果进行主成分分析，并对提取的主成分说明其意义和写出表达式，其中累计贡献率要求在 95%以上。
[知识点及要求]各站点各时段客流特征主成分提取

9.11 基于提取的主成分进行聚类分析，输出其聚类中心，并对地铁站点进行功能性分类，比如居住型、就业型...等。
[知识点及要求]基于客流特征对站点进行功能性划分

10. 图像识别

10.1 读取机器学习包中的手写体数字图像数据集，并对数据集进行初步探索分析，并在控制台中输出探索结果，包括图像像素数据集大小、图像标签取值，同时将第 1 个图像绘制出来。数据集读取代码参考如下图所示：

```
2 import sklearn.datasets
3 a=sklearn.datasets.load_digits()
```

a - Dictionary (5 elements)

Key	Type	Size	Value
DESCR	str	1	Optical Recognition of Handwr… ========== ..
data	float64	(1797, 64)	array([[0., 0., 5., ...… [0., 0., ...
images	float64	(1797, 8, 8)	array([[[0., 0., 5., ...… [0., 0. ...
target	int32	(1797,)	array([0, 1, 2, ..., 8, 9, 8])
target_names	int32	(10,)	array([0, 1, 2, 3, 4, 5, 6, 7…

[知识点及要求]图像数据认识、可视化

10.2 对图像像素数据集和图像标签数据集，按 80%训练和 20%测试进行随机划分，构建支持向量机分类模型，输出模型的准确率和测试集的预测准确率。
[知识点及要求]基于全像素特征的简单图像识别模型构建

10.3 读取机器学习包中的人脸识别图像数据集，对数据集进行探索分析，同时以 80％训练和 20%测试，构建支持向量机分类模型，输出模型的准确率和测试数据集的预测准确率。
数据集读取参考代码如下：

```
1 # -*- coding: utf-8 -*-
2 import sklearn.datasets
3 a=sklearn.datasets.fetch_olivetti_faces()
```

a - Dictionary (4 elements)

Key	Type	Size	Value
DESCR	str	1	Modified Olivetti faces dataset.
data	float32	(400, 4096)	array([[0.30991736, 0.36776859, …
images	float32	(400, 64, 64)	array([[[0.30991736, 0.36776859,…
target	int32	(400,)	array([0, 0, 0, ..., 39, 39, 39])

[知识点及要求]基于全像素特征的人脸识别模型构建

10.4 基于 10.3 的人脸识别数据集，对像素特征数据做主成分分析，并提取主成分，要求累计贡献率在 95％以上。基于提取的主成分数据，按 80%训练和 20%

测试，构建支持向量机分类模型，输出模型准确率和测试数据集的预测准确率。
[知识点及要求]基于像素主成分的人脸识别模型

10.5 现有 1 元、5 元、10 元、20 元、50 元、100 元共 6 种面额的纸币彩色图像，请计算 R、G、B 三个颜色通道的一阶、二阶、三阶颜色矩，共 9 个特征指标数据，记为自变量 X，同时构造纸币面额标签数据集，记为 Y。
[知识点及要求]图像颜色特征提取

10.6 现有 3 张 5 元、10 元、50 元的纸币彩色图像。请利用 10.5 的数据构建支持向量机分类模型，输出模型准确率，同时对 3 张纸币进行识别。
[知识点及要求]基于颜色特征的图像识别模型

11. 文本挖掘

11.1 安装结巴分词包 jieba 和文本处理包 gensim，读取上市公司新闻标题训练集和测试集文本，并进行分词处理，同时构造情感分类标签作为因变量 Y。
[知识点及要求]第三方包 jieba 和 gensim 的安装，分词应用

11.2 读取停用词文件，对分词后的训练集和测试集文本去掉停用词。
[知识点及要求]文本数据预处理：去停用词

11.3 在 11.2 基础上，继续去掉训练集和测试集文本中的数值。
[知识点及要求]文本数据预处理：去数值

11.4 对训练集和测试集文本构造词向量，并计算其逆向词频矩阵，作为训练集和测试集的自变量，记为 X 和 X1。
[知识点及要求] 文本处理包 gensim 和机器学习包 sklearn 中的文本特征提取模块应用

11.5 基于 11.1 的情感分类标签 Y 和 11.4 的训练数据 X，构建支持向量机模型，输出模型准确率，并对测试数据 X1 进行情感分类预测。
[知识点及要求]文本分类模型构建

11.6 对测试集进行人工打情感分类标签，并评估模型的预测效果。
[知识点及要求]模型评估

12. 系统开发

12.1 以 Pycharm 和 Anaconda 为开发环境，配置 Qtdesigner 界面设计师模块和 Python 程序代码生成模块。
[知识点及要求]PyQt 开发环境的配置

12.2 以 10.5 和 10.6 为基础，设计一个纸币面额识别系统，要求实现功能为：1)

导入和展示模块。点击导入按钮，弹出文件选择框，选中待识别的纸币图像后能在系统界面上显示出来；2)识别模块。点击识别按钮，能对导入的纸币进行识别并把面额值在系统界面上显示出来；3)识别系统编译成可执行文件 EXE。
[知识点及要求]界面设计、功能实现及编译成 EXE

12.3 按申银万国行业分类标准，利用 7.13 基于总体规模与投资效率指标的上市公司综合评价方法，获得 2015~2017 年每年每个行业排名前 20 的上市公司股票简称和综合得分。基于以上需求，设计一个上市公司综合评价系统，实现功能如下：1）以树结构展示所有待选择的行业；2）从树中选择指定行业之后，通过下拉列表框选择年份，则在表格中显示选中行业指定年份的排名前 20 上市公司股票简称和综合得分；3）系统编译成可执行文件 EXE。
[知识点及要求]界面设计、功能实现及编译成 EXE

13 深度学习

13.1 通过 Anaconda 的 Navigation 导航创建 Tensorflow2.x 环境，并安装 Tensorflow2.x 深度学习包。在此基础上，通过 Navigation 导航或激活其环境后用 pip 命令，向 Tensorflow2.x 环境安装绘图包 Matplotlib、机器学习包 Scikit-learn、数据处理包 Pandas 和图像处理包 pillow。
[知识点及要求] Tensorflow 深度学习环境搭建、相关包的安装。

13.2 读取 Tensorflow2.X 内置的 mnist 服饰数据集（注：如网速太慢，可用以下载好的数据集），构建多层神经网络分类模型和卷积神经网络分类模型。
[知识点及要求] 多层神经网络模型和卷积神经网络模型的基本使用方法

13.3 读取 Tensorflow2.X 内置的 imdb 电影评论数据集（注：如网速太慢，可用以下载好的数据集），构建多层神经网络分类模型和循环神经网络分类模型。
[知识点及要求] 多层神经网络模型和循环神经网络模型的基本使用方法

13.4 基于 11.1 的上市公司新闻标题训练数据集，构建循环神经网络分类模型，并对测试数据集的情感倾向进行分类。
[知识点及要求] 实际应用场景的文本情感分类模型构建及应用

13.5 基于 10.5 的纸币图像数据集，按 80%训练和 20%测试，构建卷积神经网络分类模型，并对测试数据集进行分类。
[知识点及要求] 实际应用场景的图像分类模型构建及应用。

14 网络爬虫

14.1 爬取我国所有上市公司百度新闻标题最近一个月的数据，内容包括新闻标题、发布日期、新闻来源、网址，并将结果整理输出到 Excel 表格中。
[知识点及要求]上市公司新闻标题原始数据收集、整理。

14.2 爬取腾讯视频官网电影《中国机长》最近 10 页的评论文本数据，并将结果整理输出到 Excel 表格中。
[知识点及要求]电影评论原始数据收集、整理。

14.3 爬取前程无忧招聘网站中职位关键词为：Python、大数据、数据挖掘、数据分析、算法工程师共 5 个，每个关键词爬取 50 页，内容包括职位、薪资、地区、公司、公司性质、行业类型，最终把结果整理输出到 Excel 表格中。
[知识点及要求]招聘网站关键职位信息原始数据收集、整理。

14.4 爬取苏宁易购网站中某商品（比如：华为手机）信息，内容包括商品图片、商品价格、营业店铺、评论数、好评率，并将结果整理输出到 Excel 表格中。
[知识点及要求]购物网站商品信息原始数据收集、整理。