CPDA 数据分析师专业技术考试大纲

（2019 年修订版）

考试介绍

一、考试目标

CPDA 数据分析师专业技术考试主要测试考生是否具备数据分析基础知识，是否了解数据分析工作流程及数据分析技术，是否具备利用数据分析知识及思维解决实际业务问题的能力。

重点考查内容包括数据分析理论知识、数据分析算法与模型、数据分析应用。在数据分析理论知识中，考查考生对数据分析基础知识的掌握；数据分析算法与模型，考查考生对数据分析常用算法与模型相关参数的掌握，对各种数据分析算法与模型的应用与评估能力；数据分析应用，考查考生是否能够熟练应用数据分析技术，根据企业决策需求与业务场景，借助数据分析工具与算法，以数据驱动的思维模式解决实际问题的能力。

考试范围涉及数据分析统计基础、企业战略管理、数据获取、数据预处理、数据可视化、数据分析思维、客户数据分析、产品数据分析、供应链概述、采购数据分析、生产制造数据分析、物流数据分析、营销数据分析、实业投资分析、金融投资分析等。

二、考试科目及考试形式

考试科目包括数据分析理论知识、数据分析算法与模型、数据分析应用，考试方式分为理论机考和实操机考，满分都为 100 分。

科目	考试方式	题型	总分	及格分	考试时长
数据分析理论知识	机考	单选题、多选题 、判断题	100 分	60 分	90 分钟
数据分析算法与模型	机考	操作计算题	100 分	60 分	120 分钟
数据分析应用	机考	应用分析题（操作+分析）	100 分	60 分	120 分钟
CPDA 数据分析师的认证考核采取全国统一时间，每年四次； 其他特殊情况以考前通知为准； 具体信息，请查询官方网站 www.chinacpda.com。					

表 1　考试科目及考试形式

三、教材及相关课程

《数据分析基础》（含《战略管理基础》）、《客户与产品数据分析》（原《营销数据分析》）、《供应链优化与投资分析》、《CPDA 数据分析师课程讲义》以及必修远程课程是数据分析师考生必修必考教材与资料。

课程分面授与远程教学。面授时间为 6 天，每天上课时间为 6 小时，一般为每周的周六和周日授课；远程学习有必修与选修两部分，为学员自主选课，必修课程在考试范围内，选修课程不做考试要求，课程有效期均为自选择开通日起一个自然年。教学平台登录 Ai.datahoop.cn，进入数据分析试算平台和网课中心。

科目一　数据分析理论知识

数据分析理论知识是对考生数据分析基础知识的掌握程度的测试，考试范围涉及数据分析统计基础、企业战略管理、数据获取、数据预处理、数据可视化、数据分析思维、客户数据分析、产品数据分析、供应链概述、采购数据分析、生产制造数据分析、物流数据分析、营销数据分析、实业投资分析、金融投资分析等内容基础理论知识的考查。

试卷题型结构：

判断题	共 15 题，每题 1 分	共 15 分
单选题	共 30 题 每题 1.5 分	共 45 分
多选题	共 20 题 每题 2 分	共 40 分

表 2　《数据分析理论知识》试卷题型结构

数据分析理论知识考试内容：

一、数据分析统计基础

1、考试内容：概率统计基本概念

考试样题：概率大的事件一定发生，概率小的事件一定不发生。（ ）

正确答案： 错误

2、考试内容：概率统计的参数估计、假设检验

考试样题：对于来自正态总体的 n 个简单随机样本 X，S^2 是 n 个样本的样本方差，σ^2 是总体方差，那么比值 $(n-1)S^2/\sigma^2$ 可近似服从（ ）

A、 自由度为 n-1 的 t 分布

B、 自由度为 n 的 χ^2 分布

C、 自由度为 n-1 的 χ^2 分布

D、 自由度为 n-1 的 F 分布

正确答案： C

3、考试内容：数据及其分类

考试样题：定量属性可以是整数值或者是连续值。（ ）

正确答案： 正确

4、考试内容：数据分析基本方法

考试样题：为调查我国城市女婴出生体重：北方 n1=5385，均数为 3.08kg，标准差为 0.53kg；南方 n2=4896，均数为 3.10kg，标准差为 0.34kg，经统计学检验，p=0.0034<0.01，这意味着（ ）

A、 南方和北方女婴出生体重的差别无统计学意义

B、 南方和北方女婴出生体重差别很大

C、 由于 P 值太小，南方和北方女婴出生体重差别无意义

D、 南方和北方女婴出生体重差别有统计学意义但无实际意义

正确答案： D

二、数据分析工具的使用

1、考试内容：SQL 工具的常见命令

（1）考试样题：SQL 语句中删除表的命令是（ ）

A、 DROP TABLE

B、 DELETE TABLE

C、 ERASE TABLE

D、 DELETE DBF

正确答案： A

（2）考试样题：如下表 student 中,如何筛选 type 为包含数学或语文的记录?（ ）

ID	type	score
A01	数学	78
A02	语文	76
A03	英语	90
A04	数学	68
A05	英语	84

A、 select * from student where type=“数学” and type=“语文”

B、 select * from student where type=“数学” or type=“语文”

C、 select * from student where type in （“数学” , “语文”）

D、 select * from student where type in （“数学”、“语文”）

正确答案： BC

2、考试内容：Python 工具的常见命令

考试样题：以下哪个表达式在 Python 中是非法的？（ ）

A、x=y=z=1

B、x=(y=z+1)

C、x,y=y,x

D、x+=y

正确答案： B

三、企业战略与数据分析

1、考试内容：企业战略常见模型与方法

（1）考试样题：通用（GE）矩阵对业务模块在市场中的竞争情况和发展成长情况进行分析，指定业务模块的发展战略，包括对某些业务的发展方向做出调整。（ ）

正确答案： 正确

（2）考试样题：波特五力模型中五个压力来源是供应商议价能力、购买者的议价能力、行业新进入者的威胁、替代产品的威胁及企业内部的管理压力。（ ）

正确答案： 错误

2、考试内容：企业战略应用

考试样题：大数据背景下，数据支撑业务的目的是（ ）

A、建立数据科学 B、完成数据应用 C、配备数据硬件 D、吸纳数据人才

正确答案： B

四、数据获取

1、考试内容：数据库与数据结构

（1）考试样题：下列选项中哪个属于 NoSQL 数据库？（ ）

A、BigTable B、HBase C、Oracle D、MongoDB

正确答案： ABD

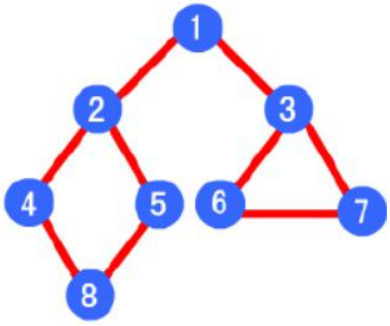
（2）考试样题：数据库其实就是一个应用软件。（ ）

正确答案：错误

2、考试内容：数据获取原理与策略，包括外部数据获取、内部数据获取等

考试样题：广度优先搜索的遍历顺序为：1->2->3->4->5->6->7->8（　）

正确答案：正确



3、考试内容：数据获取技术的应用

考试样题：智能健康手环的应用开发，体现了（　）的数据采集技术的应用。

A、 统计报表　　B、 网络爬虫　　C、 API 接口　　D、 传感器

正确答案： D

4、考试内容：抽样调查技术，包括抽样调查基本概念、抽样方法、抽样误差和精度描述、抽样实施步骤等

（1） 考试样题：抽样误差是指（　）

A、抽样调查中所存在的误差

B、 由于抽样的不同方法而产生的误差

C、 抽样调查中的工作误差

D、样本统计值与总体参数值之间存在的误差

正确答案： D

（2）考试样题：对于概率抽样，下列说法不正确的是（　）

A、也叫随机抽样

B、每个单位都有一定的机会被抽中

C、每个单位被抽中的概率是已知的，或者可以被计算出来

D、每个单位被抽中的概率都是相同的

正确答案： D

五、数据预处理

1、考试内容：数据预处理方法

考试样题：将原始数据进行集成、变换、维度规约、数值规约是在以下哪个步骤的任务？（ ）

A、数据获取　B、分类和预测　C、数据预处理　D、数据可视化

正确答案： C

2、考试内容：数据清洗与异常值处理方法

（1）考试样题：缺失值的处理方法有哪些？（ ）

A、用平均值填充

B、忽略缺失记录

C、以任意数据填充

D、用默认值填充

正确答案： ABD

（2）考试样题：假设 12 个销售价格记录组已经排序如下：5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215 ,使用等宽划分（宽度为 50）方法将它们划分成四个箱，求 15 在哪个箱子里？（ ）

A、第一个　B、第二个　C、第三个　D、第四个

正确答案： A

3、考试内容：数据集成方法

考试样题：下列关于数据重组的说法中，错误的是（ ）

A、数据重组是数据的重新生产和重新采集

B、数据重组能够使数据焕发新的光芒

C、数据重组实现的关键在于多源数据融合和数据集成

D、数据重组有利于实现新颖的数据模式创新

正确答案： A

4、考试内容：数据规约方法

考试样题：以下属于数据规约方法的是（ ）

A、数据离散化　B、数据标准化　C、噪声数据识别　D、数据压缩

正确答案： AD

5、考试内容：数据变换方法

（1） 考试样题：以下属于数据变换方法的是（ ）

A、最小最大标准化 B、 零均值标准化 C、 小数定标标准化 D、中位数标准化

正确答案： ABC

（2）考试样题：最大最小值标准化法也叫极值法，该方法适用于已知数据集的最小值或最大值情况。（ ）

正确答案： 正确

六、数据可视化

考试内容：基本图表及其使用技巧

（1）考试样题：柱状图通常用来比较变量，而条形图经常用来呈现变量的分布。（ ）

正确答案：错误

（2）考试样题：10 家公司月销售额数据（万元）分别为：72，63，54，54，29，26，25，23，23，20。下列哪个图形适合描述这些数据？（ ）

A、茎叶图 B、直方图 C、饼图 D、折线图

正确答案：A

（3）考试样题：描述急性心肌梗死患者的凝血酶浓度 X 与凝血时间 Y 的关系，宜绘制（ ）

A、条形图 B、直方图 C、散点图 D、饼图

正确答案：C

七、描述性数据分析思维场景问题

1、考试内容：场景应用类题型--数据降维问题

考试样题：如果要对问卷中的原始变量进行分解，从中归纳出潜在的“类别”，可以采用（ ）

A、回归分析 B、因子分析 C、聚类分析 D、联合分析

正确答案：B

2、考试内容：场景应用类题型--用户分类型问题

考试样题：对客户的生命周期进行分类主要使用以下哪个方法?（ ）

A、聚类分析 B、判别分析 C、逻辑回归 D、线性回归

正确答案：A

3、考试内容：场景应用类题型--关联性问题

（1）考试样题：某超市研究销售记录发现，购买牛奶的人很大概率会购买面包，这种属于数据挖掘的哪类问题？（ ）

A、聚类分析 B、关联规则 C、分类分析 D、自然语言处理

正确答案： B

（2）考试样题：利用 Apriori 算法计算频繁项集可以有效降低计算频繁集的时间复杂度。在以下的购物篮中产生支持度不小于 3 的候选 3-项集，在候选 2-项集中需要剪枝的是（ ）ID 项集。

ID 项集

1 面包、牛奶

2 面包、尿布、啤酒、鸡蛋

3 牛奶、尿布、啤酒、可乐

4 面包、牛奶、尿布、啤酒

5 面包、牛奶、尿布、可乐

A、啤酒、尿布 B、啤酒、面包 C、面包、尿布 D、啤酒、牛奶

正确答案： BD

4、考试内容：算法模型类题型—主成分分析/因子分析

考试样题：因子分析的主要作用有（ ）

A、对变量进行降维

B、对变量进行判别

C、对变量进行聚类

D、以上都不对

正确答案： A

5、考试内容：算法模型类题型—聚类算法

（1）考试样题：在聚类分析当中，簇内的相似性越大，簇间的差别越大，聚类的效果就越差。（ ）

正确答案： 错误

（2）考试样题：下列关于聚类挖掘技术的说法中，错误的是（ ）

A、不预先设定数据归类类目，完全根据数据本身性质将数据聚合成不同类别

B、要求同类数据的内容相似度尽可能小

C、要求不同类数据的内容相似度尽可能小

D、与分类挖掘技术相似的是，都是要对数据进行分类处理

正确答案： B

6、考试内容：算法模型类题型—关联规则算法

（1）考试样题：支持度表示前项与后项在一个数据集中同时出现的频率。（ ）

正确答案： 正确

（2）考试样题：Apriori 算法用下列哪个做项目集(Itemset)的筛选？（ ）

A、最小信赖度(Minimum Confidence)

B、最小支持度(Minimum Support)

C、交易编号(TransactionID)

D、购买数量

正确答案： B

八、预测性数据分析思维场景问题

1、考试内容：场景应用类题型--分类型/预测型问题

考试样题：一位母亲记录了儿子 3～9 岁的身高，由此建立的身高与年龄的回归直线方程为 y=7.19x+73.93，，据此可以预测这个孩子 10 岁时的身高，则正确的叙述是（ ）

A、身高一定是 14583mm B、身高超过 14600mm

C、身高低于 14500mm D、身高在 14583mm 左右

正确答案： D

2、考试内容：算法模型类题型--准确率指标

（1）考试样题：分类模型的误差大致分为两种：训练误差（training error）和泛化误差（generalization error）。（ ）

正确答案： 正确

（2）考试样题：召回率反映的是预测为正中的样本中正例的概率。（ ）

正确答案： 错误

（3）考试样题：以下两种描述分别对应哪两种对分类算法的评价标准？（ ）

(a)警察抓杀人犯，描述警察抓的人中有多少个是杀人犯的标准。

(b)描述有多少比例的杀人犯给警察抓了的标准。

A、Precision, Recall B、Recall, Precision

C、 Precision, ROC D、Recall, ROC

正确答案： A

3、考试内容：算法模型类题型—数据预测

考试样题：（ ）是根据市场现象的历史资料，运用科学的数学方法建立预测模型，使市场现象的数量向未来延伸，预测市场现象未来的发展变化趋势，预计或估计市场现象未来表现的数量。

A、因果预测法 B、趋势预测法（时间序列分析法）

C、定性预测法 D、定量预测法

正确答案： B

4、考试内容：算法模型类题型—决策树算法

（1）考试样题：在决策树中，随着树中结点数变得太大，即使模型的训练误差还在继续减低，但是检验误差开始增大，这是出现了模型拟合不足的问题。（ ）

正确答案： 错误

（2）考试样题：我们可以用哪种方式来避免决策树过度拟合的问题？（ ）

A、利用修剪法来限制树的深度

B、利用盆栽法规定每个节点下的最小的记录数目

C、利用逐步回归法来删除部分数据

D、目前并无适合的方法来处理这问题

正确答案： AB

5、考试内容：算法模型类题型—KNN 算法

考试样题：KNN 算法更适合于（ ）的分类问题。

A、重复时间 B、稀有事件 C、规则事件 D、相近事件

正确答案：B

6、考试内容：算法模型类题型—支持向量机算法

（1）考试样题：SVM 算法更适用于稀有事件的分类问题，如客户流失、欺诈侦测等。

正确答案：错误

（2）考试样题：SVM 算法中，对于线性不可分的情况，通过使用非线性的映射函数可以将低维不可分的样本转化到高维空间使其线性可分，这样的非线性映射函数称为（ ）

A、激活函数　B、核函数　C、超函数　D、转换函数

正确答案：B

7、考试内容：算法模型类题型—神经网络算法

考试样题：（　）是模拟生物神经网络进行信息处理的一种数学模型。

A、集成学习　　B、人工神经网络　　C、支持向量机　　D、以上三者都是

正确答案：B

8、考试内容：算法模型类题型—逻辑回归算法

考试样题：用逻辑回归方法得到的分析结果中，其中预测为正类的有 102 个，其中 78 个预测正确。预测为负类的有 115 个，其中 83 个预测正确。那么正类的 precision 和 recall 各是（　）

A、0.76　0.71　　B、0.72　0.76　C、0.77　0.72　　D、0.71　0.76

正确答案：A

9、考试内容：算法模型类题型—朴素贝叶斯算法

考试样题：Naive Bayes 是属于数据挖掘中的什么方法？（　）

A、聚类　　B、分类　　C、时间序列　　D、关联规则

正确答案：B

10、考试内容：算法模型类题型—集成学习算法

考试样题：关于集成学习以下说法正确的是？（　）

A、Adaboost 相对于单个弱分类器而言通过 Boosting 增大了模型的 Bias

B、随机森林相对于单个决策树而言通过 Bagging 增大了模型的 Variance

C、我们可以借鉴类似 Bagging 的思想对 GBDT 模型进行一定的改进，例如每个分裂节点只考虑某个随机的特征子集或者每棵树只考虑某个随机的样本子集这两个方案都是可行的

D、GBDT 模型无法在树维度通过并行提速，因为基于残差的训练方式导致第 i 棵树的训练依赖于前 i-1 棵树的结果，故树与树之间只能串行

正确答案：CD

九、客户数据分析

1、考试内容：客户画像分析与 RFM 模型

考试样题：以下哪些变量是使用 RFM 方法构造出来的？（　）

A、最近 3 期境外消费金额 B、最近 6 期网银消费交易笔数

C、信用额度 D、距最近一次逾信用额度期的月数

正确答案：ABD

2、考试内容：客户运营分析与 AARRR 模型

考试样题：产品运营中的 AARRR 模型指什么？（ ）

A、AARRR 指：产品访问量、降低跳失率、提高使用时长、付费转化率、回访率

B、AARRR 指：获取用户、提高活跃度、提高留存率、获取收入、自传播

C、AARRR 指：活跃用户、购买用户、付费转化、使用时长、跳失率

D、AARRR 指：新用户获取、老用户留存、付费转化、单用户贡献度、付费频率

正确答案：B

3、考试内容：客户关系管理模型

（1）考试样题：CRM 对企业的作用，包括（ ）

A、营销智能 B、销售自动化 C、提高效率 D、建立学习型组织

正确答案：ABC

（2）考试样题：根据客户生命周期，在潜在期阶段，关于客户服务下列说法正确的是（ ）

A、客户与企业关系获得快速发展，开始准备建立长期稳定的客户关系

B、双方互惠互利关系真正建立

C、该阶段服务的好坏直接关系到能否获得消费者的青睐，意味着能否有效地获取市场份额

D、可以采用多媒体广告策略、赠送、现场问答等多种手段

正确答案：D

十、产品数据分析

1、考试内容：产品功能分析 KANO 模型

考试样题：KANO 模型定义的顾客需求层次有（ ）

A、兴奋型需求 B、喜爱型需求 C、基本型需求 D、期望型需求

正确答案：ACD

2、考试内容：产品价格分析 PSM 模型

考试样题：在采用 PSM 模型进行价格定位时，要求受访者回答下列哪些问题？（ ）

A、开始觉得便宜的价格

B、开始觉得贵的价格

C、觉得价格太便宜以至于消费者会怀疑其质量而不接受这个产品或服务

D、觉得价格太贵以至于消费者不会考虑购买此产品或服务

正确答案：ABCD

3、考试内容：产品销量分析巴斯模型

考试样题：巴斯模型计算过程中涉及下列哪些参数？（ ）

A、公司在第 t 年之前的销售量的总和

B、公司产品在某一地区的未来市场总量

C、创新系数

D、模仿系数

正确答案：ABCD

4、考试内容：产品数据化运营分析

考试样题：下列不属于广告计费方式的专业术语的是（ ）

A、CPD B、CTR C、CPA D、CPB

正确答案：D

十一、供应链优化

1、考试内容：供应链管理

考试样题：（ ）主要体现供应链的物理功能，即以最低的成本将原材料转化成零部件、半成品、产品，以及在供应链中的运输等。

A、有效性供应链 B、反应性供应链 C、稳定供应链 D、动态供应链

正确答案：A

2、考试内容：市场营销与预测

考试样题：产品改良、市场改良和营销组合改良等决策适用于产品生命周期的（ ）

A、介绍期 B、成长期 C、成熟期 D、衰退期

正确答案：C

3、考试内容：时间序列算法

考试样题：乘法模型是分析时间序列最常用的理论模型。这种模型将时间序列按构成分解为（ ）等四种成分，各种成分之间（ ），要测定某种成分的变动，只须从原时间序列中（ ）

A、长期趋势、季节变动、循环波动和不规则波动；保持着相互依存的关系；减去其他影响成分的变动

B、长期趋势、季节变动、循环波动和不规则波动；缺少相互作用的影响力量；减去其他影响成分的变动

C、长期趋势、季节变动、循环波动和不规则波动；保持着相互依存的关系；除去其他影响成分的变动

D、长期趋势、季节变动、循环波动和不规则波动；缺少相互作用的影响力量；除去其他影响成分的变动

正确答案：C

4、考试内容：回归分析算法

（1）考试样题：在多元线性回归模型中，解释变量可以有相关性。（ ）

正确答案：错误

（2）考试样题：在一元线性回归中，判定系数 R^2 的平方根等于自变量和因变量的相关系数。（ ）

正确答案：正确

十二、投资数据分析

1、考试内容：现金流量表分析

（1）考试样题：能使经营现金流量减少的项目是（ ）

A、无形资产摊销　　B、出售长期资产利得

C、存货增加　　D、应收账款减少

正确答案：C

（2）考试样题：在企业处于高速成长阶段，投资活动现金流量往往是（ ）

A、流入量大于流出量　　B、流出量大于流入量

C、流入量等于流出量　　D、不一定

正确答案：B

（3）考试样题：现金流入量是指项目引起的企业现金收入得的增加额。（ ）

正确答案：正确

2、考试内容：企业盈利能力分析

（1）考试样题：总资产报酬率是指（ ）与平均总资产之间的比率。

A、利润总额 B、息税前利润 C、净利润 D、息前利润

正确答案：B

（2）考试样题：（ ）是反映盈利能力的核心指标。

A、总资产报酬率 B、股利发放率

C、总资产周转率 D、净资产收益率

正确答案：D

3、考试内容：企业运营能力分析

（1）考试样题：从资产流动性方面反映总资产效率的指标是（ ）

A、总资产产值率 B、总资产收入率 C、总资产周转率 D、产品销售率

正确答案：C

（2）考试样题：流动资产占总资产的比重是影响（ ）指标变动的重要因素。

A、总资产周转率 B、总资产产值率 C、总资产收入率 D、总资产报酬率

正确答案：A

4、考试内容：企业偿债能力分析

考试样题：如果流动比率大于 1，则下列结论成立的是（ ）

A、速动比率大于 1 B、现金比率大于 1

C、营运资金大于 0 D、短期偿债能力绝对有保障

正确答案：C

5、考试内容：企业发展能力分析

考试样题：下列项目中，不属于企业资产规模增加的原因的是（ ）

A、企业对外举债 B、企业实现盈利 C、企业发放股利 D、企业发行股票

正确答案：C

科目二　数据分析算法与模型

数据分析算法与模型，考查考生对数据分析常用算法与模型相关参数的掌握，对各种数据分析算法与模型的应用与评估能力。

考试范围涉及以下内容：

（1）机器学习算法与模型，包括监督学习算法中逻辑回归算法、决策树、朴素贝叶斯算法、KNN 算法、支持向量机算法、神经网络算法等；非监督算法中各种聚类算法、降维算法以及关联规则算法等。

（2）客户&产品数据分析算法与模型，包括客户数据分析的 RFM 模型、AARRR 模型；产品数据分析的 KANO 模型、PSM 模型、巴斯模型、漏斗模型、对应分析方法模型、协同过滤算法等。

（3）供应链分析算法与模型，包括供应商选择--层次分析法，生产优化--规划求解法，物流优化—节约里程法、市场预测分析—回归算法与时间序列算法等。

（4）投资数据分析算法与模型，包括实业投资中现金流量表、投资收益分析与投资风险分析；金融投资中量化选股模型、量化择时模型等。

试卷结构：4 道数据分析算法与模型操作计算题，每题 25 分，共 100 分。

数据分析算法与模型考试内容：

一、机器学习算法与模型

1、考试样题：关联分析算法与模型试题

以下购物篮数据展示的是 10 个订单中购买 a、b、c、d、e 五种产品的情况（1 代表购买，0 代表未购买），据此进行分析。数据预览如下：

购物篮 ID	a	b	c	d	e
1	1	0	0	1	1
2	1	1	1	0	1
3	1	1	0	1	1
4	1	0	1	1	1
5	1	0	1	0	1
6	0	1	0	1	1
7	0	0	1	0	0

8	1	1	1	0	0
9	1	0	0	1	1
10	1	1	0	0	1

数据文件（考试系统提供）：购物篮数据.csv

要求（25 分）：

（1）设 minsupport=40%，利用 Apriori 算法写出所有的频繁项目集，并指出其中支持度最大的二项频繁项目集。（10 分）

（2）在第一问基础上设 minconfidence=60%，找出所有的有效强关联规则。（15 分）

参考答案：

（1）将正确数据格式的数据导入 DATEHOOP 进行关联分析，根据题目要求，分析参数设置如下：最小支持度=0.4。得到所有频繁项目集：

支持度排序	项目一		项目二	支持度	置信度	提升度
1	{}	-->	a	0.8	0.8	1
2	{}	-->	b	0.5	0.5	1
3	{}	-->	c	0.5	0.5	1
4	{}	-->	d	0.5	0.5	1
5	{}	-->	e	0.8	0.8	1
6	a	-->	e	0.7	0.875	1.09375
7	e	-->	a	0.7	0.875	1.09375
8	d	-->	e	0.5	1	1.25
9	e	-->	d	0.5	0.625	1.25
10	a	-->	d	0.4	0.5	1
11	e	-->	ad	0.4	0.5	1.25
12	a	-->	de	0.4	0.5	1
13	de	-->	a	0.4	0.8	1
14	ad	-->	e	0.4	1	1.25
15	ae	-->	d	0.4	0.571428571	1.142857143
16	b	-->	e	0.4	0.8	1
17	d	-->	a	0.4	0.8	1
18	e	-->	b	0.4	0.5	1
19	b	-->	a	0.4	0.8	1
20	a	-->	b	0.4	0.5	1
21	c	-->	a	0.4	0.8	1
22	a	-->	c	0.4	0.5	1

其中支持度最大的二项频繁项目集为{a,e}/{e,a}。

（2）设置最小支持度为 0.4，最小置信度为 0.6，满足最小支持度和置信度条件下提升度>1 的为有效强关联规则，本题目要求的有效强关联规则如下：

6	d	-->	e	0.5	1	1.25
7	ad	-->	e	0.4	1	1.25

8	a	-->	e	0.7	0.875	1.09375
9	e	-->	a	0.7	0.875	1.09375
15	d	-->	ae	0.4	0.8	1.14285714
16	e	-->	d	0.5	0.625	1.25

2、考试样题：逻辑回归算法与模型试题

根据鸢尾花数据，变量包括萼片长、萼片宽、花瓣长、花瓣宽以及花的类型，分析问题。

X1 萼片长（cm）	X2 萼片宽（cm）	X3 花瓣长（cm）	X4 花瓣宽（cm）	y 花的类型
数值型	数值型	数值型	数值型	0 和 1 分别代表两种不同的类型

数据文件（考试系统提供）：鸢尾花训练集数据.csv
鸢尾花预测数据.csv

要求（25 分）：

（1）根据训练数据，用类型_num 作为因变量 Y，其他变量作为自变量 X，做逻辑回归，写出逻辑回归的方程（结果保留两位小数，逻辑回归参数不要调整）。（5 分）

（2）根据模型结果，写出伪随机数发生器种子为 2，测试集占比 20%情况下逻辑回归的训练和测试混淆矩阵，以及准确率和召回率以及 Accuracy 的值。（12 分）

（3）根据预测数据，依据训练模型结果预测，写出预测结果。（8 分）

参考答案：

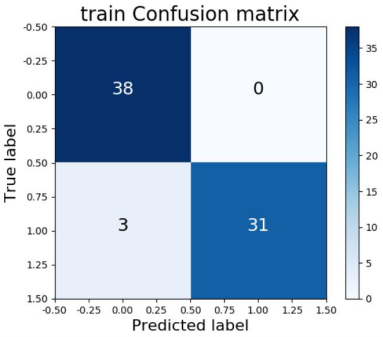
将正确数据格式的数据导入 DATEHOOP 进行逻辑回归分析：

（1）设萼片长为 x1，萼片宽为 x2，花瓣长为 x3，花瓣长为 x4，根据训练数据得到回归方程为：

Ln{p(y=1)/p(y=0)}=0.99+0.26X1+0.72X2-0.77X3-20.04X4

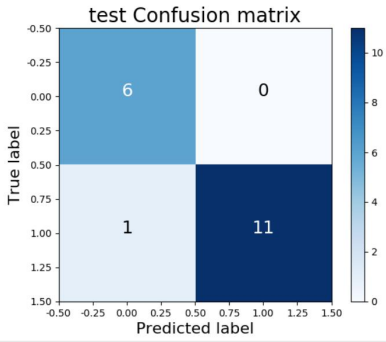
（2）参数设置伪随机数发生器种子为 2，测试集占比为 20%进行逻辑回归分析

训练集--混淆矩阵



训练集混淆矩阵：

测试集--混淆矩阵



测试集混淆矩阵：

训练集准确率和召回率以及 Accuracy 的值：

	Precision	Recall
0	0.9268	1
1	1	0.9118

Accuracy：0.9583

测试集准确率和召回率以及 Accuracy 的值：

	Precision	Recall
0	0.8571	1
1	1	0.9167

Accuracy：0.9444

（3）根据上述模型进行数据预测，得到的预测结果如下表所示：

萼片长（cm）	萼片宽（cm）	花瓣长（cm）	花瓣宽（cm）	y-predict
7.7	3	6.1	2.3	0
7.1	3	5.9	2.1	0
6.7	3.1	4.4	1.4	1
6.2	2.9	4.3	1.3	1
6	2.2	5	1.5	0
6.9	3.2	5.7	2.3	0
6.3	3.3	6	2.5	0
5.5	2.4	3.7	1	1
7.2	3.6	6.1	2.5	0
6.9	3.1	4.9	1.5	1

二、客户&产品数据分析算法与模型

1、 考试样题： PSM 算法与模型试题

某中高档户外运动品牌为寻找合理的促销折扣力度，对 300 名消费者进行了调查，在回收的 264 份有效问卷中，对各档折扣分别持“比较便宜”、“太便宜了”、“还是有点贵”、“还是太贵”四种态度的人数做了统计，数据预览如下：

折扣	人数（人）			
	Q1、比较便宜	Q2、太便宜了	Q3、还是有点贵	Q4、还是太贵
1 折或以下	0	102	0	0
2 折	21	74	0	0
3 折	45	48	21	0
4 折	69	11	18	0
5 折	45	13	45	18
6 折	26	13	26	8
7 折	50	3	66	26
8 折	8	0	48	82
9 折	0	0	30	77
9 折以上	0	0	10	53
人数总计	264	264	264	264

数据文件（考试系统提供）：中高档户外运动品牌消费调查数据.xlsx

要求（25 分）：

（1）受访者对各档折扣分别持四种态度的人数累计百分比和曲线草图。(10 分)

（2）受访者对各档折扣分别持接受、保留接受、不接受态度的人数百分比。（5 分）

（3）请为该品牌选取合理的折扣区间，及最优折扣。（10 分）

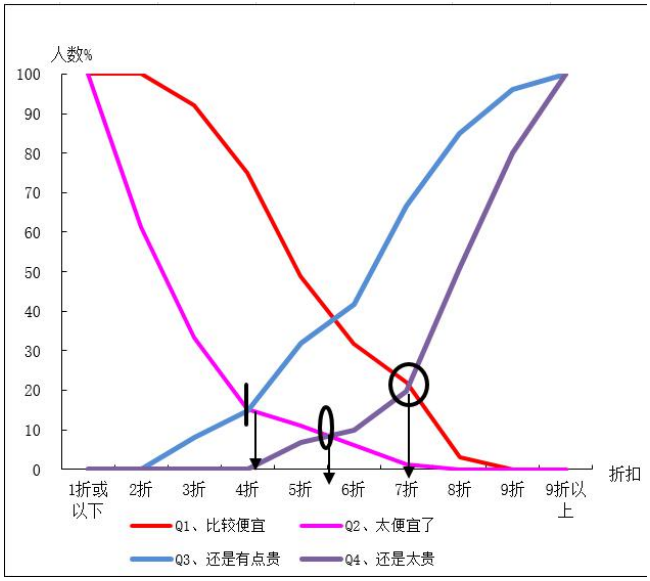
参考答案：

（1）受访者对各档折扣分别持四种态度的人数累计百分比:

折扣	人数比例（%）				累计人数比例(%)			
	Q1、比较便宜	Q2、太便宜了	Q3、还是有点贵	Q4、还是太贵	Q1、比较便宜	Q2、太便宜了	Q3、还是有点贵	Q4、还是太贵
1 折或以下	0	39	0	0	100	100	0	0
2 折	8	28	0	0	100	61	0	0
3 折	17	18	8	0	92	33	8	0
4 折	26	4	7	0	75	15	15	0
5 折	17	5	17	7	49	11	32	7
6 折	10	5	10	3	32	6	42	10

7 折	19	1	25	10	22	1	67	20
8 折	3	0	18	31	3	0	85	51
9 折	0	0	11	29	0	0	96	80
9 折以上	0	0	4	20	0	0	100	100
人数比例总计	100	100	100	100				

受访者对各档折扣分别持四种态度的人数累计曲线草图



（2）受访者对各档折扣分别持接受、保留接受、不接受态度的人数百分比:

折扣	人数比例%		
	可接受者	有保留接受者	不接受者
1 折或以下	0	0	100
2 折	0	39	61
3 折	0	67	33
4 折	10	75	15
5 折	19	63	18
6 折	27	58	16
7 折	11	68	21
8 折	12	37	51
9 折	4	16	80
9 折以上	0	0	100

（3）根据第（1）（2）步分析，该品牌合理的折扣区间为 4 折到 7 折，其中 5.5 折为最优折扣。

2、考试样题： KANO 算法与模型试题

某厨卫公司要开发一款燃气灶产品，列举出 5 个可作为卖点的功能属性：防风、定时、防干烧、不沾油、快速而准确地打火。该公司的产品设计人员不知道该主要开发哪项功能，分析师小李向公司提出了使用 KANO 模型对上述五个功能进行调研分类的想法，并得到了公司的支持。假设你是小李：

（1）请你绘制 KANO 模型图来介绍对功能属性分类的思路（5 分）

（2）请你对燃气灶的防干烧功能属性设计调查问题（5 分）

（3）针对燃气灶的防干烧功能，受访者有多少种可能的回答组合，请写出每一种回答组合所对应的属性类别符号（符号见题注）（5 分）

（4）假设基于对 240 名受访者的调研，得到下表，请算出这 5 种功能各自的 worse 系数和 better 系数，并基于这两个系数判断这 5 中功能的属性类别（5 分）

人数 \ 功能	O	I	M	A
防风	118	30	29	63
防干烧	47	78	30	85
定时	22	140	8	70
快速而准确地打火	45	20	107	68
不沾油	69	51	29	89

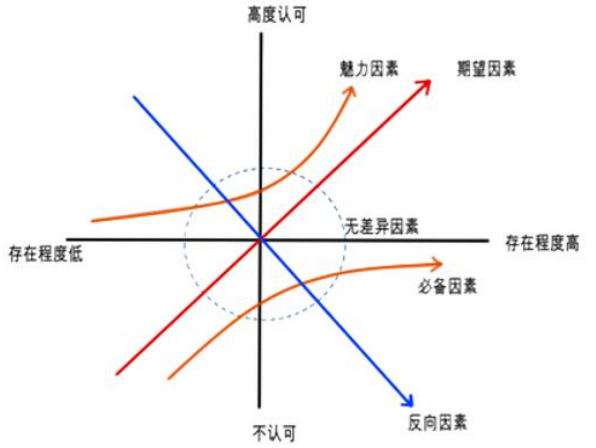
（5）请对该燃气灶的这 5 项功能开发提出建议（5 分）

数据文件（考试系统提供）：燃气灶功能开发数据.xlsx

注：魅力因素属性用符号 A 表示；必备因素属性用符号 M 表示；期望因素属性用符号 O 表示；可有可无无差异因素属性用符号 I 表示；用户讨厌的反向因素属性用 R 表示；有问题的回答用 Q 表示

参考答案：

（1）



分为期望因素（O），表示具备某功能满意度会提升，反之则满意度下降；必备因素（M），表示不具备某功能满意度会明显下降，而具备某功能满意度不会大幅提升；无差异因素（I），表示具备或不具备某功能对满意度没影响；魅力因素（A）表示不具备某功能满意度不会明显下降，而具备某功能满意度会大幅提升；反向因素（R）表示不具备某功能满意度会提升，反之则满意度下降。

（2）设置调查问卷如下：

如果燃气灶有此功能，你感觉如何？	如果燃气灶没有该功能，你感觉如何？
1.我喜欢	1.我喜欢
2.理应如此	2.理应如此
3.无所谓	3.无所谓
4.我能忍受	4.我能忍受
5.我不喜欢	5.我不喜欢

（3）针对燃气灶的防干烧功能，受访者可能产生 25 种组合的答案，具体见下表：

KANO 模型		燃气灶不具备防干烧功能				
		1.我喜欢	2.理应如此	3.无所谓	4.我能忍受	5.我不喜欢
燃气灶具备防干烧功能	1.我喜欢	Q	A	A	A	O
	2.理应如此	R	I	I	I	M
	3.无所谓	R	I	I	I	M
	4.我能忍受	R	I	I	I	M
	5.我不喜欢	R	R	R	R	Q

魅力因素属性用符号 A 表示；必备因素属性用符号 M 表示；期望因素属性用符号 O 表示；可

有可无无差异因素属性用符号 I 表示；用户讨厌的反向因素属性用 R 表示；有问题的回答用 Q 表示

（4）根据【better 系数=（O+A）/（O+I+M+A）；worse=（O+M）/（O+I+M+A）】得到

人数 \ 功能	O	I	M	A
防风	118	30	29	63
防干烧	47	78	30	85
定时	22	140	8	70
快速而准确地打火	45	20	107	68
不沾油	69	51	29	89

防风 better 系数=（118+63）/(118+30+29+63)=75.4%

防干烧 better 系数=（47+85）/（47+78+30+85）=55%

定时 better 系数=（22+70）/（22+140+8+70）=38.33%

快速而准确地打火 better 系数=（45+68）/（45+20+107+68）=47.08%

不沾油 better 系数=（69+89）/（69+51+29+89）=66.38%

防风 worse 系数=（118+29）/(118+30+29+63)=61.25%

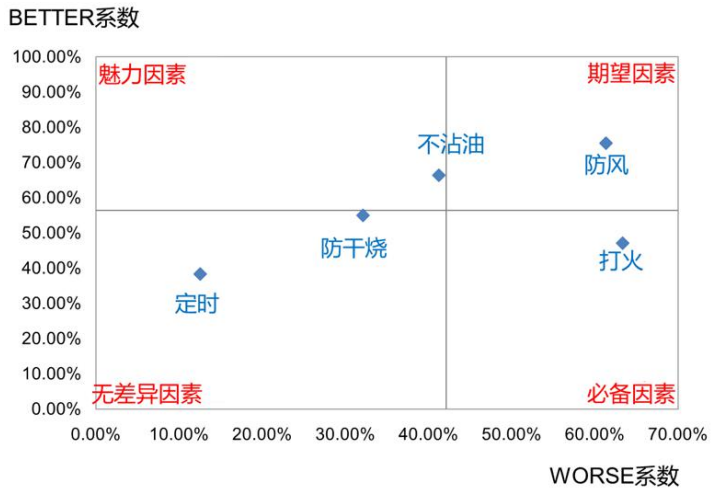
防干烧 worse 系数=（47+30）/（47+78+30+85）=32.08%

定时 worse 系数=（22+8）/（22+140+8+70）=12.5%

快速而准确地打火 worse 系数=（45+107）/（45+20+107+68）=63.33%

不沾油 worse 系数=（69+29）/（69+51+29+89）=41.17%

绘制 KANO 模型图：



根据各功能所处区域可判断出：防风属期望因素，快速而准确地打火属必备因素，不沾

油介于期望和魅力之间，防干烧介于魅力和无差异之间，定时属无差异因素。

（5）产品对几类属性的研发重要性排序为 M>O>A>I。应优先确保产品 100%具备快速而准确地打火的功能，且关注产品质量，降低故障率；防风功能的好坏将直接影响用户的满意度，故要力求产品有较好的防风性能，确保用户满意度；不沾油功能很有可能，直接影响用户的满意度，故要力求产品有较好的不沾油功能，确保用户满意度；在保证前三项功能的前提下，防干烧功能可进一步在细分市场的特殊人群定向推出；暂不需开发定时功能。

三、供应链数据分析算法与模型

1、 考试样题：规划求解算法与模型试题

假设一个地区的总悬浮微粒(TSP)来源于当地的三个工厂。若工厂 1 和工厂 2 燃煤 TSP 排放因子为 95kg/t(煤)，工厂 3 的 TSP 排放因子为 85kg/t(产品)。工厂 3 产量为 250000 t (产品)/a，工厂 1 和工厂 2 的燃煤量分别为 400,000 t/a 和 300,000t/a，为了满足环境质量要求，TSP 最大允许排放量为 17,600 ,000kg/a，各种除尘设备效率和各种除尘方法的费用如下表所示：

说明：各种除尘设备去除TSP 的效率（表1）并不适应于每个污染源，可行的各种除尘方法的相应费用见表2，表中的费用折合成了单位产品所需金额（元/t）。

表1 各种除尘设备的效率

j	设备类型	设备类型除尘效率/%
0	无	0
1	重力沉降室	59
2	惯性除尘器	74
3	旋风除尘器	84
4	喷雾洗涤器	94
5	电除尘器	97

表2 可行的各种除尘方法的费用

说明：决策变量 X_{ij} 表示工厂 i 使用第 j 种除尘方法处理 X 吨煤（产品）

TSP 控制方法	i 污染源					
	1（工厂 1）		2（工厂 2）		3（工厂 3）	
	决策变量	费用（元/T）	决策变量	费用（元/T）	决策变量	费用（元/T）
0	X_{10}	0.0	X_{20}	0.0	X_{30}	0.0
1	X_{11}	1.0	X_{21}	1.4	X_{31}	1.1
2	不可行	不可行	不可行	不可行	X_{32}	1.2
3	不可行	不可行	不可行	不可行	X_{33}	1.5

4	X_{14}	2.0	X_{24}	2.2	X_{34}	3.0
5	X_{15}	2.8	X_{25}	3.0	不可行	不可行

数据文件（考试系统提供）：环境治理数据.xls

要求：（25 分）

求达到环境目标时，所用的最小的治理费用。

参考答案：

a 变量设置：工厂 1 四种方案 X10，X11，X14，X15 费用分别为 C10，0.0，C11，1.0，C14，2.0，C15，2.8，工厂 2 四种方案 X20，X21，X24，X25 费用分别为 C20，0.0，C22，1.4，C24，2.2，C25，3.0，工厂 3 五种方案 X30，X31，X32，X33，X34 费用分别为 C30，0.0，C31，1.1，C32，1.2，C33，1.5，C34，3.0

b.目标函数：

min=1.0*x11+2.0*X14+2.8*X15+1.4*X21+2.2*X24+3.0*X25+1.1*X31+1.2*X32+1.5*X33+3.0*X34

c.约束条件:

约束条件 1:TSP 最大允许排放量

（1）污染源工厂 1 的年排放 TSP 量为 95*[(1-0%)*X10+(1-59%)*X11+(1-94%)*X14+(1-97%)*X15]，

（2）污染源工厂 2 的年排放 TSP 量为 95*[(1-0%)X20+(1-59%)X21+(1-94%)X24+(1-97%)X25]

（3）污染源工厂 3 的年排放 TSP 量为 85*[1-0%)X30+(1-59%)X31+(1-74%)X32+(1-84%)X33+(1-94%)X34]

（1）+（2）+（3）≤17,600,000

约束条件 2：燃煤量

（1）工厂 1：X10+X11+X14+X15=400，000

（2）X20+X21+X24+X25=300,000

（3）X30+X31+X32+X33+X34=250,000

约束条件 3：每种方案的燃煤量得大于等于 0

Xij≥0.

d.结果及描述

1516842 或 1517207(由于小数保留个数不同会导致结果细微差别,不影响得分)

达到环境目标最小的治理费用为 1516842 元。

2、 考试样题：回归分析算法与模型试题

美国各航空公司业绩的统计数据公布在《华尔街日报 1999 年年鉴》（The Wall Street Journal Almanac 1999）上。航班正点到达的比率和每 10 万名乘客投诉的次数的数据预览如下：

航空公司名称	航班正点率	投诉率（次/10 万名乘客）
西南（Southwest）航空公司	81.8	0.21
大陆（Continental）航空公司	76.6	0.58
西北（Northwest）航空公司	76.6	0.85
美国（US Airways）航空公司	75.7	0.68
联合（United）航空公司	73.8	0.74
美洲（American）航空公司	72.2	0.93
德尔塔（Delta）航空公司	71.2	0.72
美国西部（Americawest）航空公司	70.8	1.22
环球（TWA）航空公司	68.5	1.25

数据文件（考试系统提供）：航班正点到达的比率和每 10 万名乘客投诉的次数数据.csv

乘客投诉次数预测.csv

要求（25 分）：

（1）画出这些数据的散点图。（2 分）

（2）根据散点图，表明二变量之间存在什么关系？（5 分）

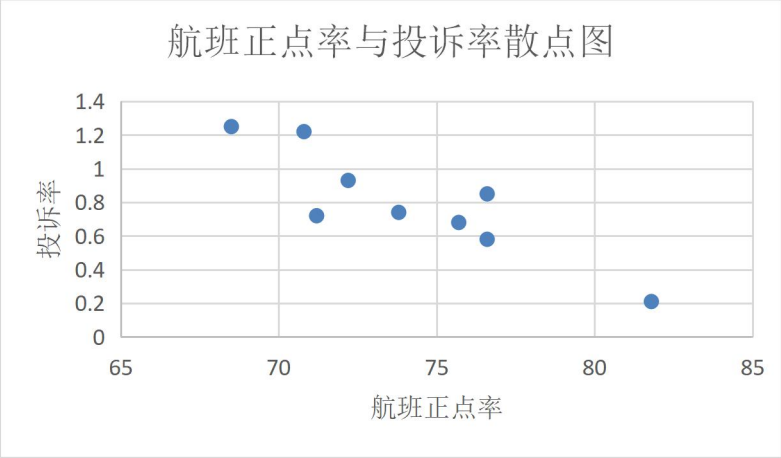
（3）求出描述投诉率是如何依赖航班按时到达正点率的估计的回归方程。（8 分）

（4）估计的回归方程的斜率做出解释；并写出判定系数的值，解释其值代表的实际意义。（5 分）

（5）求航班按时到达的正点率为 80%，82%，85%,87%,90%,估计每 10 万名乘客投诉的次数是多少？并进行实际意义解读。（5 分）

参考答案：

（1）



（使用不同工具得到图形形态差异不影响得分）

（2）根据散点图可以看出，随着航班正点率的提高，投诉率呈现出下降的趋势，说明航班整点率与投诉率两者之间，存在着一定的负相关关系。

（3）设投诉率为 y，航班正点率为 x，建立回归方程 y=ax+b

方法一：利用 Excel 计算，得到数据如下

SUMMARY OUTPUT

回归统计	
Multiple R	0.882607
R Square	0.778996
Adjusted R Square	0.747424
标准误差	0.160818
观测值	9

方差分析

	df	SS	MS	F	Significance F
回归分析	1	0.638119	0.638119	24.67361	0.001624
残差	7	0.181037	0.025862		
总计	8	0.819156			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	6.017832	1.05226	5.718961	0.000721	3.529633	8.506031	3.529633	8.506031
航班正点率	-0.07041	0.014176	-4.96725	0.001624	-0.10393	-0.03689	-0.10393	-0.03689

根据以上数据得到回归方程 y=6.02-0.07x

方法二：利用 Datahoop2.0，得到数据入下；

模型	系数	P 值（t 检验）
航班正点率	-0.07	0.00162
常数项	6.0178	0.00072

根据以上数据得到回归方程 y=6.017831995-0.0704144x（可保留两位小数）

（4）回归方程斜率解释：根据第（3）回归方程，斜率为 0.07，航班正点率每提高一个百分点，相应的投诉率（次/10 万名乘客）下降 0.07.

根据 datahoop 数据：

R 方	调整 R 方
0.78	0.75

判定系数：R^2=0.77, 回归关系可以解释因变量 77%的变异，投诉次数可以由 77%由正点率解释。

（5）方法一：根据 y=6.017831995-0.0704144x 依次带入航班按时到达的正点率为 80%，82%，85%,87%,90%的结果为 0.384679995，0.24385119，0.032607995，0，0。

方法二:使用 Datahoop2.0 直接用预测，得到数据如下：

航班正点率	预测
80	0.3847
82	0.2439
85	0.0326
87	-0.1082
90	-0.3195

在预测的每 10 万名乘客投诉次数为负的情况下要考虑实际意义，不会为负值，而应该为 0。

实际意义在于说明乘客投诉率不仅仅受航班正点率影响，也就是说不完全是正点率越高，投诉率越低。

四、投资数据分析算法

1、考试样题：实业投资数据分析算法与模型

某公司正在考虑是否购买一台生产用于粘接瓷砖用胶水的机器。机器价值 100 万元，预期使用寿命为三年。然而，对于机器带来的年净营业额的估计带有不确定性，要视家装行业的状态而定。公司的管理层做出了如下的估算：基准贴现率为 10%

行业状态	0	1	2	3
繁荣	－100	50	70	98
正常	－100	50	60	70
萧条	－100	30	30	25

数据文件（考试系统提供）：投资数据.xls

要求：（本题 25 分）

（1）请根据三种行业状态，分别计算对应的 NPV 是多少？（10 分）

（2）根据我国建筑技术研究院的最新数据显示，行业未来的前景如下：繁荣的概率为 20%、正常的概率为 60%、萧条的概率为 20%，如果根据这种估计，请计算项目的期望 NPV（10 分）

（3）根据期望 NPV，计算项目的标准差。（5 分）

正确答案:

（1）根据 NPV 计算公式得到结果如下：

行业状态	0	1	2	3	NPV
繁荣	-100	50	70	98	76.93
正常	-100	50	60	70	47.63
萧条	-100	30	30	25	29.15

（2）

行业状态	NPV	概率	iNPV	ENPV
繁荣	76.93	20%	=76.93*20%=15.387	=15.387+28.58-5.83=38.137
正常	47.63	60%	=47.63*60%=28.580	

萧条	29.15	20%	=-29.15*20%=-5.830	

（3）根据期望NPV，项目的标准差为：[(76.93-15.387)^2*20%+(47.63-28.58)^2*60%+(-29.15+5.83)^2*20%]^(1/2)=35.51

2、考试样题：金融投资数据分析算法与模型

某公司拟进行股票投资，计划购买A、B、C三种股票，并分别设计了甲、乙两种投资组合。已知三种股票的β系数分别为1.5、1.0和0.5，它们在甲种投资组合下的投资比重为50%、30%和20%；乙种投资组合的风险收益率为3.4%。目前无风险利率是8%，市场组合收益率是12%。

要求（25分）：

（1）根据A、B、C股票的β系数，分别评价这三种股票相对于市场投资组合而言的投资风险大小（5分）。

（2）按照资本资产定价模型计算A股票的必要收益率（5分）。

（3）计算甲种投资组合的β系数和风险收益率（5分）。

（4）计算乙种投资组合的β系数和必要收益率（5分）。

（5）比较甲、乙两种投资组合的β系数，评价它们的投资风险大小（5分）。

参考答案：

（1）A股票的β＞1，说明该股票所承担的系统风险大于市场投资组合的风险（或A股票所承担的系统风险等于市场投资组合风险的1.5倍）

B股票的β=1，说明该股票所承担的系统风险与市场投资组合的风险一致（或B股票所承担的系统风险等于市场投资组合的风险）

C股票的β＜1，说明该股票所承担的系统风险小于市场投资组合的风险（或C股票所承担的系统风险等于市场投资组合风险的0.5倍）

（2）A股票的必要收益率=8%+1.5×（12%-8%）=14%

（3）甲种投资组合的β系数=1.5×50%+1.0×30%+0.5×20%=1.15

甲种投资组合的风险收益率=1.15×（12%-8%）=4.6%

（4）乙种投资组合的β系数=3.4%/（12%-8%）=0.85

乙种投资组合的必要收益率=8%+3.4%=11.4%

或者：乙种投资组合的必要收益率=8%+0.85×（12%-8%）=11.4%

（5）甲种投资组合的β系数（1.15）大于乙种投资组合的β系数（0.85），说明甲投资组合的系统风险大于乙投资组合的系统风险。

科目三 数据分析应用

数据分析应用，考查考生是否能够熟练应用数据分析技术，根据企业决策需求与业务场景，借助数据分析工具与算法，以数据驱动的思维模式解决实际问题的能力。考试题型主要是通过数据分析流程、分析业务背景辨别适合应用的分析算法模型，并综合评估分析结果，对实际问题进行分析、预测并提出解决方案。

试卷结构：2 道数据分析应用题，每题 50 分，共 100 分。

数据分析应用考试内容：

一、 考试样题：通过游戏用户相关行为数据预测用户是否会付费。

某游戏公司，想通过半年用户行为数据，对用户是否会付费进行预测，并根据预测结果对可能付费用户进行精准营销。

数据字段如下：

user_id	install_date	last_login_date	level_end	os	is_payer	active_days	duration	avg_session_cnt
用户编号	游戏安装时间	最后一次登录游戏时间	用户退出时的游戏等级	登录手机系统	是否付费	活跃天数	用户留存天数（构造变量）	每天登录频次

数据类型解释

user_id	install_date	last_login_date	level_end	os	is_payer	active_days	duration	avg_session_cnt
字符型	日期型	日期型	数值型	字符型，取值为：Android 和 iOS	是否付费 1 代表付费，0 代表未付费	数值型	数值型	数值型

数据文件（考试系统提供）：游戏用户训练数据.csv

游戏用户预测数据,csv

要求（50 分）：

请根据原始数据，对数据进行预处理（包括对类别型变量进行数值化处理、重新构造新的变量），然后自行选择变量和分析算法进行分析，写出分析过程和思路，并且根据模型进行预测。

参考答案：

1、分析题目背景，确定要解决的问题

2、数据预处理:

(1)从原始数据中可以看出 leve_end（用户退出时的游戏等级）， is_payer（是否付费），active_days（活跃天数），avg_session_cnt（每天登录频次）这几个变量可以直接选入进行分析，对于 install_date（游戏安装时间）和 last_login_date（最后一次登录游戏时间）由于是日期型数据，因此选择用 last_login_date-install_date 得到的相差的天数来代表游戏用户的活跃时长（题目已经构造好）。

(2)Os 是文本型，因此需要提前进行处理，转化为数值型数据，本次处理把取值设为：

Android	ios
1	0

(3)原始数据不存在缺失值，通过描述分析原始数据发现付费玩家和非付费玩家数据比例为 0.87:1，因此样本比较均衡，不需要调整样本。

(4)不需要标准化，异常客户在此场景中属于正常研究情况，不处理异常情况，让其参与模型

3、数据分析算法选取

由于采用分类算法进行分析，样本比较均衡，因此可以采用逻辑回归进行分析，也可以采用其他方法如神经网络进行分析。可自行选取。

4、指标结果，指标分析

根据样本的测试集数据得到模型测试结果为：（参照模型的 Accuracy，precision, recall 的值）

（1）选用逻辑回归方法，测试集 20%，随机数种子 1，得到如下结果

Accuracy：0.8728

AUC：0.93523

	precision	Recall	F1-score	Support

0	0.8838	0.8899	0.8869	436
1	0.8584	0.8508	0.8546	342

（2）选用 SVM 方法，测试集 20%，随机数种子 1，得到如下结果:

Accuracy：0.7904

AUC：0.8862

	precision	Recall	F1-score	Support
0	0.9475	0.6628	0.78	436
1	0.6892	0.9532	0.79	342

将几种方法模型计算对比（对比结果略），通过分类分析结果选出准确率和召回率都较高，模型的准确度也较好的模型进行分析和预测。

5、预测结果

利用较优模型根据给定的预测数据选取预测模型进行预测，判断以预测结果和实际结果比较来判断，选用逻辑回归得到预测结果如下：

1,1,1,1,0,1,1,1,1,0。

二、 考试样题：移动公司精准营销数据化解决方案

假如你是某移动运营商的数据分析师，结合用户通话行为数据，通过数据分析为用户推荐相应套餐，或者结合用户现有套餐优化套餐，提供个性化套餐，从而对客户进行精准营销，增加客户粘性。

运营商收集到的数据包含下列字段：

变量名称	变量标签
Customer_ID	用户编号
Peak_mins	工作日上班时间电话时长
OffPeak_mins	工作日下班时间电话时长
Weekend_mins	周末电话时长
International_mins	国际电话时长
Total_mins	总通话时长

average_mins	平均每次通话时长

数据文件（考试系统提供）：电信数据.csv

要求（50 分）：

根据客户行为数据，进行数据的预处理（可以自行根据现有变量构造新变量进行分析），预处理之后选择合适变量进行分析，分析算法自行选择，写出分析思路和过程，通过数据分析对客户进行细分，并为运营商提供客户精准营销的相关建议。（请写出分析的流程并刻画最后细分之后的客户的特点和相应的营销建议）

参考答案：

1、分析题目背景，确定要解决的问题

分析题目背景根据题目选取 k-means 聚类方法对客户进行细分，其他方法依据结果得分。

2、预处理过程：缺失异常处理

3、构造新变量

对数据进行预处理，查看数据的相关性，通过查看相关系数矩阵看到 peak_mins 和 total_mins 相关性很高，International_mins 和 total_mins 相关性很高。

列名	Peak-mins	Offpeak-mins	Weekend-mins	International-mins	Total-mins	Average-mins
Peak-mins	1	0.121	0.145	0.692	0.942	-0.037
Offpeak-mins	0.121	1	0.025	0.262	0.443	0.009
Weekend-mins	0.145	0.025	1	0.128	0.2	-0.107
International-mins	0.692	0.262	0.128	1	0.712	-0.034
Total-mins	0.942	0.443	0.2	0.712	1	-0.036
Average-mins	-0.037	0.009	-0.107	-0.034	-0.036	1

因此对变量进行处理，只选取其中一个，然后构造 2 个新的变量 peak_mins/total_mins，和 International_mins/ total_mins。

4、数据标准化

对这些变量进行聚类分析，由于新构造的变量取值与其他变量取值范围相差较大，因此，在聚类分析时，选择标准化处理之后的数据进行聚类

5、变量选取

对数据进行聚类分析，选取变量为：peak_mins/total_mins，offpeak_mins，weekend_mins，international_mins/total，tltal_mins, average_mins。

6、选取方法和分析标准

根据题目选取 k-means 聚类方法对客户进行细分（其他方法依据结果得分）。根据 keans 聚类结果分析 5 类效果最好，分析每一类客户在现有变量上的特征，这里选取平均值作为参考依据。

7、得到聚类分析描述结果

标准化前

Peak_mins	OffPeak_mins	Weekend_mins	International_mins	Total_mins	average_mins	Peak_mins/Total_mins	类别
2145.24	392.2971	64.79714	530.9894	2602.334	3.329268	0.82577	0
1416.307	405.9186	57.38217	380.2975	1879.608	3.487637	0.751837	1
1033.352	292.2824	56.31813	317.7485	1381.952	3.541277	0.744067	2
667.619	230.969	51.36746	246.0991	949.9556	4.169006	0.695814	3
52.6952	39.37861	18.23432	15.56027	110.3081	2.785264	0.49393	4

标准化后

Offpeak-mins	Weekend-mins	Total-mins	Average-mins	Peak-mins/ Total-mins	International-mins/ Total-mins	类别
0.718	0.404	0.967	-0.116	0.064	-0.885	0
-0.764	-0.137	0.029	-0.136	0.927	0.292	1
0.519	0.059	-0.371	-0.107	-0.794	0.587	2
-1.37	-0.989	-2.305	-0.426	-1.802	-1.329	3
-0.031	-0.567	-0.487	3.645	-0.116	-0.047	4

不同做法，结果会有差异，但不影响分类和得分

8、业务描述

从图中可以看出，第 0 类用户几乎所有指标都最高，尤其是总时长，下班和周末通话时长特别高，上班时间也较高，只有平均通话时长较低和国际通话在总通话时长占比较低，因此第一类用户属于不分工作日纯办公的通话用户，高端商用用户；

第 2 类用户在所有指标上属于中等偏上，其中国际通话最高，上班通话时长在总时长占比不高，平均时长较高，总时长居中，所以第二类用户属于生活通话用户，中高等商用用户；

第 1 类用户在所有指标上处于中等水平，上班通话最高，下班通话最短，所以第三类用户属于电话客服类用户，属于办公商用客户，中等商用用户；

第 4 类用户只在平均通话时长上高，其他指标均偏低，说明第四类用户属于常聊用户；

第 3 类用户在所有指标上均较低，因此第五类用户属于低端用户。

9、决策建议

移动公司可以针对这五种用户推荐不同的套餐，高端用户推荐各项指标偏高，套餐费用也偏高的套餐；中端用户和中高端用户可以较高端用户偏低一点进行套餐推荐，常聊用户可以推荐符合常聊特点的套餐，比如通话次数优惠类套餐，低端用户可以推荐资费便宜的套餐。