# 实训课题1

“拍照赚钱”是移动互联网下的一种自助式服务模式。用户下载APP，注册成为APP的会员，然后从APP上领取需要拍照的任务（比如上超市去检查某种商品的上架情况），赚取APP对任务所标定的酬金。这种基于移动互联网的自助式劳务众包平台，为企业提供各种商业检查和信息搜集，相比传统的市场调查方式可以大大节省调查成本，而且有效地保证了调查数据真实性，缩短了调查的周期。因此APP成为该平台运行的核心，而APP中的任务定价又是其核心要素。如果定价不合理，有的任务就会无人问津，而导致商品检查的失败。

附件一是一个已结束项目的任务数据，包含了每个任务的位置、定价和完成情况（“1”表示完成，“0”表示未完成）；附件二是会员信息数据，包含了会员的位置、信誉值、参考其信誉给出的任务开始预订时间和预订限额，原则上会员信誉越高，越优先开始挑选任务，其配额也就越大（任务分配时实际上是根据预订限额所占比例进行配发）。附件一和附件二的表结构如表1和表2所示。

表1 附件一：已结束项目任务数据

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 任务号码 | 任务gps 纬度 | 任务gps经度 | 任务标价 | 任务执行情况 |
| A0001 | 22.56614225 | 113.9808368 | 66 | 0 |
| A0002 | 22.68620526 | 113.9405252 | 65.5 | 0 |
| A0003 | 22.57651183 | 113.957198 | 65.5 | 1 |
| A0004 | 22.56484081 | 114.2445711 | 75 | 0 |
| A0005 | 22.55888775 | 113.9507227 | 65.5 | 0 |
| A0006 | 22.55899906 | 114.2413174 | 75 | 0 |
| A0007 | 22.54900371 | 113.9722597 | 65.5 | 1 |
| A0008 | 22.56277351 | 113.9565735 | 65.5 | 0 |
| …… | …… | …… | …… | …… |

表2 附件二：会员信息数据

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 会员编号 | 会员位置(GPS) | 预订任务限额 | 预订任务开始时间 | 信誉值 |
| B0001 | 22.947097 113.679983 | 114 | 6:30:00 | 67997.3868 |
| B0002 | 22.577792 113.966524 | 163 | 6:30:00 | 37926.5416 |
| B0003 | 23.192458 113.347272 | 139 | 6:30:00 | 27953.0363 |
| B0004 | 23.255965 113.31875 | 98 | 6:30:00 | 25085.6986 |
| …… | …… | …… | …… | …… |

注：案例内容及数据来源于2017年全国大学生数学建模竞赛B题

建一个Python脚本文件，命名为test1.py，完成以下任务：

1. 计算A0001、A0002任务位置到所有会员位置的距离，结果采用两个序列来保存，分别记为S\_A0001、S\_A0002，其中index为会员编号，值为其对应的距离。（
2. 计算A0001任务在5公里范围内的会员个数，记为A0001\_Bnum。
3. 计算A0002任务在5公里范围内所有会员信誉值总和，记为A0002\_Bavg。

附注：设定A点（纬度，经度）和B点（纬度，经度），则两点之间的距离可以用以下公式进行计算：



其中距离的单位为：公里。

# 实训课题2

今有某银行的贷款审批数据，采集的特征一共有15个，依次为x1~x15。决策变量为y,取值为1（同意贷款）和0（不同意贷款），数据共690条记录，具体见文件《银行贷款审批数据.xlsx》。

建一个Python脚本文件，命名为test2.py，完成以下任务：

1）特征数据x1~x15存在缺失数据，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用中位数策略填充。

2）对x1~x6变量数据作均值-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。

3）将经过前面两步处理后的数据集，取前600条记录作为训练数据，后90条记录作为测试数据，利用支持向量机、逻辑回归、神经网络三种不同的模型依次做出预测，并计算其预测精度。

4）将预测精度通过柱状图表示出来，其中横轴为模型名称（支持向量机、逻辑回归、神经网络），纵轴为其对应的预测精度。

# 实训课题3

今有我国各地区普通高等教育发展状况数据，具体见《高教数据.xlsx》，其中x1为每百万人口高等院校数；x2为每十万人口高等院校毕业生数；x3为每十万人口高等院校招生数；x4为每十万人口高等院校在校生数；x5为每十万人口高等院校教职工数；x6为每十万人口高等院校专职教师数；x7为高级职称占专职教师比例；x8为平均每所高等院校的在校生数；x9为国家财政预算内普通高教经费占国内生产总值比重；x10为生均教育经费）。

建一个Python脚本文件，命名为test3.py，完成以下任务：

1. 对以上指标数据做主成分分析，并提取其主成分，要求累计贡献率达到90%以上。
2. 基于提取的主成分，对以上30个地区做K-均值聚类分析（K=4），并将地区名称和所属的类别在命令窗口中输出聚类结果。

# 实训课题4

假设有以下数据集，每行代表一个顾客在超市的购买记录：

I1: 西红柿、排骨、鸡蛋、毛巾、水果刀

I2: 西红柿、茄子、水果刀、香蕉

I3: 鸡蛋、袜子、毛巾、肥皂、水果刀

I4: 西红柿、排骨、茄子、毛巾、水果刀

I5: 西红柿、排骨、酸奶

I6: 鸡蛋、茄子、酸奶、肥皂、香蕉

I7: 排骨、鸡蛋、茄子、水果刀

I8: 土豆、鸡蛋、袜子、香蕉、水果刀

I9: 西红柿、排骨、鞋子、土豆、香蕉

建一个Python脚本，命名为test4.py，完成以下任务：

1. 将以上购买记录转换为布尔类型数据，其数据结构为DataFrame
2. 利用Apriori关联规则挖掘算法函数进行关联规则挖掘，最小支持度和最小置信度分别为0.2和0.4，并将挖掘结果导出到Excel表格中

# 实训课题5

今有沪深300指数2014年的交易数据，其数据结构表3所示。

表3 沪深300指数2014年交易数据

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |
| 000300 | 2014-01-08 | 2240.64 | 2262.58 | 2228.42 | 2241.91 | 513488.5 |
| 000300 | 2014-01-09 | 2236.97 | 2258.89 | 2220.8 | 2222.22 | 559870.4 |
| 000300 | 2014-01-10 | 2216.52 | 2224.49 | 2200.22 | 2204.85 | 541692.9 |
| 000300 | 2014-01-13 | 2207 | 2222.07 | 2183.6 | 2193.68 | 501227.7 |
| 000300 | 2014-01-14 | 2192.84 | 2214.12 | 2179.91 | 2212.85 | 540499.5 |
| 000300 | 2014-01-15 | 2210.02 | 2215.9 | 2193.8 | 2208.94 | 489624 |
| …… | …… | …… | …… | …… | …… | …… |

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。

建一个Python脚本，命名test4.py，完成以下任务：

1）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

……

2）取后30条记录作为测试样本，剩下的数据记录为训练样本，利用支持向量机模型进行训练及测试，获得模型的准确率和预测准确率，分别记为score和Rv，并在命令窗口中输出score和Rv。

# 实训课题6

任务如下：

1. 对hepatitis数据集x1~x6数据采用均值策略进行填充，对x7~x19采用最频繁值策略进行填充。
2. 基于填充后的数据，对x1~x6采用0-1极差化法进行标准化处理，x7~x19不做任何标准化处理。
3. 在2）基础上，对数据做主成分分析，并提取主成分（要求累计贡献率在95%以上）
4. 基于提取的主成分数据，以1~130行数据为训练数据，训练决策树模型，并利用训练好的决策树模型，对131~155行数据进行分类预测，同时计算预测准确率。

注：hepatitis数据集自变量有19个，即x1~x19，其中缺失数据值用“？”表示。因变量有1个，即y。y为分类变量，一共有两类，1表示一类，2表示另一类。

# 实训课题7

“拍照赚钱”是移动互联网下的一种自助式服务模式。用户下载APP，注册成为APP的会员，然后从APP上领取需要拍照的任务（比如上超市去检查某种商品的上架情况），赚取APP对任务所标定的酬金。这种基于移动互联网的自助式劳务众包平台，为企业提供各种商业检查和信息搜集，相比传统的市场调查方式可以大大节省调查成本，而且有效地保证了调查数据真实性，缩短了调查的周期。因此APP成为该平台运行的核心，而APP中的任务定价又是其核心要素。如果定价不合理，有的任务就会无人问津，而导致商品检查的失败。

附件一是一个已结束项目的任务数据，包含了每个任务的位置、定价和完成情况（“1”表示完成，“0”表示未完成）；附件二（附件二：会员信息数据2）是会员信息数据，包含了会员的位置、信誉值、参考其信誉给出的任务开始预订时间和预订限额，原则上会员信誉越高，越优先开始挑选任务，其配额也就越大（任务分配时实际上是根据预订限额所占比例进行配发）；

问题如下：

1. 对附件一中的每一个任务，计算以下指标（12个）：

Z1=5公里范围内的其他任务量

Z2=5公里范围内的任务平均价格

Z3=5公里范围内会员可预订任务量

Z4=5公里范围内会员平均信誉值

Z5=5公里范围内会员个数

Z6=5公里范围内6:30发布的可预订任务量

Z7=5公里范围内6:33-6:45发布的可预订任务量

Z8=5公里范围内6:48-7:03发布的可预订任务量

Z9=5公里范围内7:06-7:21发布的可预订任务量

Z10=5公里范围内7:24-7:39发布的可预订任务量

Z11=5公里范围内7:42-7:57发布的可预订任务量

Z12=5公里范围内8:00发布的可预订任务量

1. 对以上计算的12个指标进行相关性分析。如果多个指标之间存在较强的相关性，请对这12个指标进行主成分分析，并提取主成分（要求累计贡献率在90%以上），同时写出每个主成分的表达式。
2. 针对被执行的任务，以提取的主成分作为自变量，任务定价作为因变量，研究自变量与因变量之间的关系，线性还是非线性？如果是线性的，则构建多元线性回归模型；如果是非线性的，则构造BP神经网络回归模型。同时对模型进行训练。
3. 针对未被执行的任务，以提取的主成分作为自变量，利用上一步选择的模型，重新预测其任务定价
4. 针对所有任务，以Z1~Z12和原来的定价作为自变量（共13个），执行情况为因变量。训练支持向量机分类模型
5. 针对未执行任务，以Z1~Z12和重新预测的任务定价作为自变量（共13个），利用上一步训练好的支持向量机分类模型进行预测，获得其执行情况。
6. 模型评价，即针对未执行任务，重新预测的任务定价比原来的任务定价增加了多少？被支持向量机预测为执行的任务有多少，即增加了多少个任务被执行了？

# 实训课题8

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，它能有效的预知部分财务风险。本课题将风险公司记为ST，非风险公司记为非ST，ST判断标准如下：

1.）连续两年年报显示净利润为负值

2）净资产收益率、总资产净利润率为负值。

其影响特征变量如下：



注: 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|;

请利用2018年的财务指标相关数据，构建基于BP神经网络的财务风险预警模型。