(203)上海大学2020-2021年春季学期模拟试卷　2021.5

课程名：数据分析与智能计算 课程号：00864118 学分：3

基础题

应试人声明:

我保证遵守《上海大学学生手册》中的《上海大学考场规则》，如有考试违纪、作弊行为，愿意接受《上海大学学生考试违纪、作弊行为界定及处分规定》的纪律处分。

学号：（见登录信息）　　姓名：（见登录信息）

题目	选择题	程序题			总分
题号	1~15	1	2	3	
题分	30	20	20	30	100
得分					

本试卷由选择题（30分）、程序填空题（40分）和编程题（30分）三部分组成，

选择题共包括15个单选题，由计算机自动完成组卷和阅卷。

（本试卷考试时间　90　分钟）

一、单选题（本大题 15 道小题，每小题 2 分，共 30 分），从下面题目给出的 A、B、C、D 四个可供选择的答案中选择一个正确答案。

1. 下面关于数据科学与大数据之间的关系描述，错误的是_____

A. 大数据属于数据科学的范畴

B. 大数据分析遵循数据科学处理问题的基本工作流程

C. 大数据分析采用的技术完全不同于数据科学技术

D. 大数据技术是指数据量达到某种规模时引入的分布式存储、计算和传输等方法

2. 下列不属于 Python 优势的是_____

A. 语法简洁，程序开发速度快

B. 拥有大量的第三方库，能够调用 C、C++、Java 语言

C. 程序的运行速度在所有计算机语言中最快

D. 开源免费

3. 下列不属于 Numpy 库的 ndarray 数组属性的是_____

A. ndim　　B. shape

C. size　　D. add

4. 记录同学成绩的 scores 数组如下，scores[1:3, [2,5]] 取得的数据是_____

scores:array([[70, 85, 77, 90, 82, 84, 89], [60, 64, 80, 75, 80, 92, 90],
[90, 93, 88, 87, 86, 90, 91], [80, 82, 91, 88, 83, 86, 80], [88, 72, 78, 90, 91, 73, 80]])

A. array([[80,92], [88,90]])
B. array([[64,80], [93,86]])
C. array([[85,84],[64,92], [93,90]])
D. array([[64,92], [93,90],[82,86]])

5.下列不能实现将 shape 为 dtype[5,7]的 scores 数组所有元素都加 10 的语句是_____

A. scores + 10
B. np.add(scores, 10)
C. scores[10].add(10)
D. scores + [10,10,10,10,10,10,10]

6. DataFrame 对象 df 中基于位置序号选取第 2 行第 3 列数据的方式是_____（序号从 0 开始）

A. df.find(1,2)
B. df.iloc[1,2]
C. df.loc[1,2]
D. df.rloc[1,2]

7. 关于 DataFrame 和 Series 对象，下列叙述正确的是_____

A. DataFrame 对象只能用于处理两维数据
B. DataFrame 对象不能转化为 Series 对象
C. Series 对象主要用于处理一维数据
D. Series 对象可以用来处理多维数据

8. 关于 DataFrame 数据对象的添加和删除操作，_____是正确的描述

A. DataFrame 对象不能直接添加新的列数据
B. DataFrame 对象不能直接添加新的行数据
C. 可以设置 axis 的值删除 DataFrame 指定行或列的数据
D. DataFrame 中数据元素的值不能修改

9. 假定 DataFrame 对象 temp 中共有 12 列， 语句_____删除空值（NaN）个数大于 3 的行

A. temp.dropna(threshold = 8)
B. temp.dropna(threshold = 9)
C. temp.dropna(threshold = 7)
D. temp.dropna(threshold = 3)

10. 某人做数据分析测得个人健康和年龄的相关系数是 -1.09。根据这个你可以得出哪个结论_____

A. 年龄是健康程度很好的预测器
B. 年龄是健康程度很糟的预测器

C. 这个相关系数有错误
D. 以上说法都不对

11. 绘制多个子图的正确方法是_____。
A. 导入 matplotlib.pyplot 库，创建 figure 对象，调用 figure.subplot 函数
B. 导入 pandas.pyplot 库，创建 figure 对象，调用 figure.subplot 函数
C. 导入 pandas.pyplot 库，创建 figure 对象，调用 figure.add_subplot 函数
D. 导入 matplotlib.pyplot 库，创建 figure 对象，调用 figure.add_subplot 函数

12. _____可用于展示离散数据
A. 折线图
B. 柱状图
C. 统计地图
D. 曲面图

13. _____属于机器学习中的回归问题。
A. 预测短信是否为垃圾短信
B. 根据房屋特性预测房价
C. 识别车牌
D. 机场安检人脸识别

14. F1_score 可用于衡量分类模型性能，根据混淆矩阵，F1 = _____
A. 2a/(2a+b+c)
B. (a+d)/(a+b+c+d)
C. a/(a+c)
D. a/(a+b)

15. 关于聚类分析，正确的是_____
A. "簇"越少说明聚类效果越好
B. 聚类是有监督学习方法
C. 聚类可作为分类等其他任务的预处理过程
D. 同一个数据集，不同的聚类算法得到的结果是一样的

二、程序题 (本大题 3 道小题，共 70 分)

编程题 1（20 分）

台风记录数据集（winds.csv）记录了 2014 年某区域发生的台风信息，包含台风名、台风等级、气压（百帕）、移动速度（公里/时）、纬度、经度、记录数、顺序、风速（米/秒）等 9 个属性，具体说明见"数据集说明"文件。（源程序 fill_1.py）

1) 从文件中读出台风数据；
2） 查看是否存在缺失数据，删除包含缺失数据的样本；
3） 输出达到超强台风等级的台风名字。

源程序文件（fill_1.py）

请填空完成程序，需要填空的程序为素材文件夹【E:\KS\sc182003001】中的 source.py,其中【1】【2】【3】【4】为需要填空的部分,将填空后完整的程序以文件名 182003001.py 保存到考试文件夹【E:\KS\】文件夹下。

源程序文件（fill_1.py）源代码：

```
import pandas as pd
import numpy as np
#1) 从文件中读出台风数据
filename = 'winds.csv'
winds = pd.【1】(filename)
#print(winds[0:5])
#2）查看是否存在缺失数据，删除包含缺失数据的样本 print(winds.【2】())
【3】(inplace = True)
#3）输出达到超强台风等级的台风名
names = winds.loc[【4】,"windname"  ].unique() print("\n 达到超强台风等级：\n", names  )
```

编程题 2（20 分）

根据 IDC 的统计数据，各品牌手机在中国的年销量如表 1 所示（源程序 fill_2.py）。

1) 根据表 1 的数据，绘制折线图分析各品牌销量发展趋势，如图 1 所示；

2) 计算 2018 年各品牌手机的同比增幅（(Y2018-Y2017)/Y2017），并在原数据中增加新列 "INC2018"，如图 2 所示；

3) 显示增幅为正的品牌 2015-2018 年的销售量。源程序文件（fill_2.py）

表1 品牌手机年销量（单位：百万台）

	Y2015	Y2016	Y2017	Y2018
Huawei	62.9	76.6	90.9	104.97
Apple	58.4	44.9	41.1	36.32
OPPO	35.3	78.4	80.5	78.94
vivo	35.1	69.2	68.6	75.97
Mi	64.9	41.5	55.1	51.99

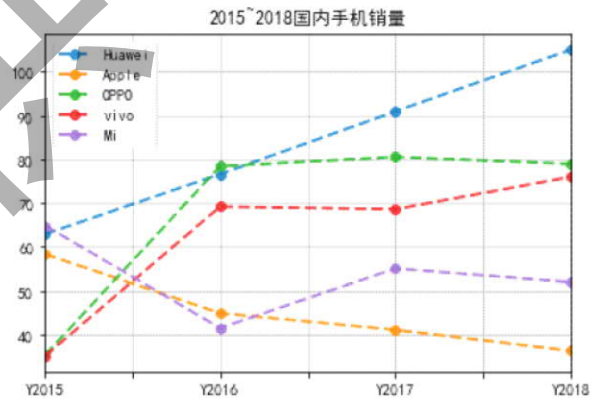


图1 手机销量折线图

	Y2015	Y2016	Y2017	Y2018	INC2018
Huawei	62.9	76.6	90.9	104.97	0.154785
Apple	58.4	44.9	41.1	36.32	-0.116302
OPPO	35.3	78.4	80.5	78.94	-0.019379
vivo	35.1	69.2	68.6	75.97	0.107434
Mi	64.9	41.5	55.1	51.99	-0.056443

图2 增加列：2018年各品牌手机的同比增幅INC2018

请填空完成程序，需要填空的程序为素材文件夹【E:\KS\sc182004001】中的 source.py,其中【1】【2】【3】【4】为需要填空的部分,将填空后完整的程序以文件名 182004001.py 保存到考试文件夹【E:\KS\】文件夹下。

源程序文件（fill_2.py）源代码：

```
#1)记录表 1 的数据，绘制折线图分析各品牌销量发展趋势; index =
['Huawei','Apple','OPPO','vivo','Mi'];
columns = ['Y2015','Y2016','Y2017','Y2018']
data = np.array( [ [62.9,76.6,90.9,104.97], [58.4,44.9,41.1,36.32],
[35.3,78.4,80.5,78.94],[35.1,69.2,68.6,75.97],
   [64.9,41.5,55.1,51.99] ] )
sales = DataFrame(【1】)
print(sales)
#绘制折线图
psales = DataFrame(data.T, columns, index)
print(psales)
plt.rcParams['font.sans-serif'] = ['SimHei']
【2】(title='2015~2018 国内手机销量',LineWidth=2, marker='o',
linestyle='dashed',grid=True,alpha=0.9)
plt.【3】()
#2)计算 2018 年各品牌手机的同比增幅，并在原数据中增加新列"2018 同比增幅";
sales['INC2018'] =【4】
print(sales)
```

编程题 3（30 分）

数据集（素材文件夹【E:\KS\sc182005001】中的 heart-disease.xlsx）记录了某项实验中测试者的身体数据，请编写程序完成下述功能要求，并将完成的程序（.py 或.ipynb 文件）以文件名 182005001 保存到考试文件夹【E:\KS\】文件夹下

根据不同的身体指标可以预测患心脏病的风险等级（Risk）。heart-disease.xlsx 记录了测试者的年龄（Age）、性别（Sex）、胸痛类型（Cp）、血压（Trestbps）、胆固醇（Chol）等 13 种身体指标（具体说明见“数据集说明”文件），以及患心脏病的风险等级。风险等级分为五种：无风险（no）、低风险（low）、中风险（medium）、高风险（high）、极高风险（very high）。

请根据数据集（heart-disease.xlsx）文件格式，正确获取数据样本进行预处理，统计分析，建立分类模型；尝试多种算法，比较分类的性能。

具体要求如下：

根据不同的身体指标可以预测患心脏病的风险等级（Risk）。heart-disease.xlsx 记录了测试者的年龄（Age）、性别（Sex）、胸痛类型（Cp）、血压（Trestbps）、胆固醇（Chol）等 13 种身体指标（具体说明见“数据集说明”文件），以及患心脏病的风险等级。风险等级分为五种：无风险（no）、低风险（low）、中风险（medium）、高风险（high）、极高风险（very high）。

请根据数据集（heart-disease.xlsx）文件格式，正确获取数据样本进行预处理，统计分析，建立分类模型；请尝试多种算法，比较分类的性能。

具体要求如下：

1) 从文件中读出所需的数据，根据分析需求将所需的数据保存到 DataFrame 中。

2) 数据清洗。判断数据集中是否有缺失数据，并采取以下方式处理：

若‘Thal’有缺失则使用同列前一行数据填充；

若‘Ca’有缺失则使用同列中出现最多的数据填充；

若‘Age’有缺失则删掉该行数据。

3) 统计各类‘Risk’的数量，并列出每类‘Risk’的名字。

4) 使用散点图矩阵分析‘Risk’类型与‘Age’、‘Trestbps’特征间的相关性，并计算他们之间的相关系数。

5) 数据预处理，将数据中非数值型的数据转换为数值类型。

6) 选择合适的数据列作为特征和分类标签形成数据集用于训练分类模型，并将数据集分为训练集和测试集。

7) 试用决策树算法建立模型，并在训练集上建立分类模型，在测试集上测试模型预测的准确性。

8) 根据第 7) 步的运行结果，说明此算法在‘Risk’分类数据上的性能。请将结果用文字描述在程序文件给出的注释行中。

选择题 答案

1-5 CCDAC

6-10 BCBBC

11-15 DBBAC

程序题 1 答案

【1】read_csv

【2】isnull

【3】winds.dropna

【4】[‘level’].values==’超强台风’

程序题 2 答案

【1】data=data,index=index,columns=columns

【2】psales.plot

【3】show

【4】(sales['Y2018']-sales['Y2017'])/sales['Y2017']

程序题 3 答案

详细见附页

上海大学2020-2021年春季学期

数据分析与智能计算 模拟试卷

程序题3参考答案

By YANthinkn 2021.5

In [69]:
```
#导入库文件
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

1.从文件中读出所需的数据

In [70]:
```
filename='heart-disease.xlsx'
data=pd.read_excel(filename)
data
```

Out[70]:

	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Risk
0	44.0	male	2	130	219	0	2	188	0	0.0	1	0.0	3.0	no
1	60.0	male	4	130	253	0	0	144	1	1.4	1	1.0	7.0	low
2	54.0	male	4	124	266	0	2	109	1	2.2	2	1.0	7.0	low
3	50.0	male	3	140	233	0	0	163	0	0.6	2	1.0	7.0	low
4	41.0	male	4	110	172	0	2	158	0	0.0	1	0.0	7.0	low
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
904	35.0	male	2	120	192	0	0	174	0	0.0	1	0.0	3.0	no
905	56.0	male	2	118	240	0	0	169	0	0.0	3	0.0	3.0	no
906	63.0	female	4	122	197	0	0	136	1	0.0	2	0.0	3.0	low
907	59.0	male	4	162	176	1	2	90	0	1.0	2	2.0	6.0	high
908	57.0	female	4	138	241	0	0	123	1	0.2	2	0.0	7.0	low

909 rows × 14 columns

2.数据清洗

In [71]:
```
#判断数据集中是否有缺失数据
data.isnull()
```

In [71]:
```
#判断数据集中是否有缺失数据
data.isnull()
```

Out[71]:

	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Risk
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
904	False	False	False	False	False	False	False	False	False	False	False	False	False	False
905	False	False	False	False	False	False	False	False	False	False	False	False	False	False
906	False	False	False	False	False	False	False	False	False	False	False	False	False	False
907	False	False	False	False	False	False	False	False	False	False	False	False	False	False
908	False	False	False	False	False	False	False	False	False	False	False	False	False	False

909 rows × 14 columns

In [88]:
```
#若‘Thal’有缺失则使用同列前一行数据填充;
data['Thal'].fillna(method='ffill',inplace=True)

#若‘Ca’有缺失则使用同列中出现最多的数据填充;
pd.value_counts(data['Thal'])
#出现最多的是3.0，即用3.0填充
data['Thal'].fillna(value='3.0',inplace=True)

#若‘Age’有缺失则删掉该行数据.
data.dropna(how='any',inplace=True)

data
```

Out[88]:

	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Risk
0	44.0	1	2	130	219	0	2	188	0	0.0	1	0.0	3.0	0
1	60.0	1	4	130	253	0	0	144	1	1.4	1	1.0	7.0	1
2	54.0	1	4	124	266	0	2	109	1	2.2	2	1.0	7.0	1
3	50.0	1	3	140	233	0	0	163	0	0.6	2	1.0	7.0	1
4	41.0	1	4	110	172	0	2	158	0	0.0	1	0.0	7.0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
904	35.0	1	2	120	192	0	0	174	0	0.0	1	0.0	3.0	0
905	56.0	1	2	118	240	0	0	169	0	0.0	3	0.0	3.0	0
906	63.0	0	4	122	197	0	0	136	1	0.0	2	0.0	3.0	1
907	59.0	1	4	162	176	1	2	90	0	1.0	2	2.0	6.0	3
908	57.0	0	4	138	241	0	0	123	1	0.2	2	0.0	7.0	1

888 rows × 14 columns

3.统计各类‘Risk’的数量，并列出每类‘Risk’的名字

In [89]:
```
pd.value_counts(data['Risk'])
#每类‘Risk’的名字:'No' 'Low' 'Medium' 'High' 'Very_High'
```

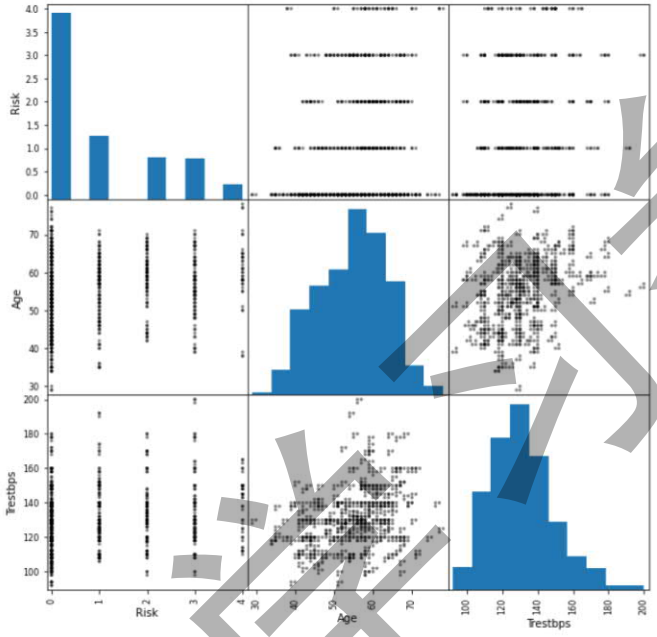
Out[89]:
```
0    474
1    162
2    108
3    105
4     39
Name: Risk, dtype: int64
```

4.使用散点图矩阵分析‘Risk’类型与‘Age’、‘Trestbps’特征间的相关性

In [90]:
```
fig=plt.figure(figsize=(10,10))
ax1=fig.add_subplot(1,1,1)
new_data=data.loc[:,['Risk','Age','Trestbps']]

#此处将'Risk'中非整数的数据转换为数值类型，否则无法用散点图矩阵展示
#此步骤与第五问有重复
new_data.loc[new_data['Risk']=='no','Risk']=0
new_data.loc[new_data['Risk']=='low','Risk']=1
new_data.loc[new_data['Risk']=='medium','Risk']=2
new_data.loc[new_data['Risk']=='high','Risk']=3
new_data.loc[new_data['Risk']=='very_high','Risk']=4
new_data['Risk']=new_data['Risk'].values.astype(int)

pd.plotting.scatter_matrix(new_data,diagonal='hist',color='k',ax=ax1)
plt.show()
```



In [91]:
```
#计算他们之间的相关系数
new_data.corr()
```

Out[91]:

	Risk	Age	Trestbps
Risk	1.000000	0.228721	0.161688
Age	0.228721	1.000000	0.286472
Trestbps	0.161688	0.286472	1.000000

5.数据预处理，将数据中非数值型的数据转换为数值类型

In [93]:
```
#Sex
data.loc[data['Sex']=='female','Sex']=0
data.loc[data['Sex']=='male','Sex']=1
#Risk
data.loc[data['Risk']=='no','Risk']=0
data.loc[data['Risk']=='low','Risk']=1
data.loc[data['Risk']=='medium','Risk']=2
data.loc[data['Risk']=='high','Risk']=3
data.loc[data['Risk']=='very_high','Risk']=4

data
data.to_excel("test.xls")
```

6.选择合适的数据列作为特征和分类标签形成数据集用于训练分类模型，并将数据集分为训练集和测试集

In [94]:
```
#取data前13列数据作为特征属性值,最后一列作为分类值
X = data.iloc[ :, 0:12 ].values.astype(float)
y = data.iloc[ :, 13].values.astype(int)

#将数据集分割为训练集和测试集
from sklearn import model_selection
X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y, test_size=0.35, random_state=1)
```

7.试用决策树算法建立模型，并在训练集上建立分类模型，在测试集上测试模型预测的准确性

In [100]:
```
#训练模型，预测样本分类
from sklearn import tree
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X_train, y_train)
clf.score(X_train,y_train)

#评估分类器性能,计算混淆矩阵，Precision 和 Recall
predicted_y_test = clf.predict(X_test)
from sklearn import metrics
print(metrics.classification_report(y_test, predicted_y_test))
print('Confusion matrix:' )
print( metrics.confusion_matrix(y_test, predicted_y_test) )
```

```
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       157
           1       0.92      0.86      0.89        66
           2       0.92      0.92      0.92        36
           3       0.78      0.82      0.79        38
           4       1.00      1.00      1.00        14

    accuracy                           0.93       311
   macro avg       0.92      0.92      0.92       311
weighted avg       0.93      0.93      0.93       311

Confusion matrix:
[[155   2   0   0   0]
 [  3  57   0   6   0]
 [  0   0  33   3   0]
 [  1   3   3  31   0]
 [  0   0   0   0  14]]
```

8.根据第 7步的运行结果，说明此算法在'Risk'分类数据上的性能

In [ ]:
```
#此算法在'Risk'分类数据上的性能较好，准确率较高，Accuracy,F1数值较好，可以作为预测风险等级的工具
#但是...
```

答案仅供参考