**2020年上海市高等学校信息技术水平考试试卷**

**三级 数据科学技术及应用（A 场）**

（本试卷考试时间 150 分钟）

**一、单选题 ( 本大题 15 道小题 ，每小题 1 分，共 15 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的基本工作流程

C.大数据分析采用的方法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、计算和传输方法

2.下面关于DataFrame存储表结构数据的说法，错误的是\_\_\_\_\_\_\_\_。

A.通常使用行存储一条数据，列存储该数据的各个特征项

B.DataFrame对象只能使用行、列索引对进行数据切片，不能使用位置序号

C.从DataFrame对象中取出一列，得到Series对象

D.Series对象可以使用Numpy的函数进行统计分析

3.统计量“方差”描述了\_\_\_\_\_\_\_\_。

A.样本的平均值

B.样本的离散程度

C.样本中不同的值占样本容量的比例

D.样本中出现次数最多的值

4.CSV文件是常用的数据文件格式，可以使用\_\_\_\_\_\_\_查看。

A.文本编辑器、Excel

B.photoshop

C.powerpoint

D.画图工具

5.students对象数据如下：

age height weight

1 19 170 68

2 20 165 65

3 18 175 65

下面语句筛选出\_\_\_\_\_\_\_。

>>>students.loc[students['height']<170, 'weight']

A.身高小于170同学的体重

B.身高小于170同学的信息

C.身高大于170同学的体重

D.身高小于170同学的身高

6.为描述高校教师学历占比情况，适合的图形是\_\_\_\_\_\_\_。

A.散点图

B.曲面图

C.直方图

D.饼图

7.目前人工智能技术，特别是机器学习，主要模拟了人的\_\_\_\_\_\_\_\_\_过程。

A.学习

B.推理

C.思考

D.规划

8.\_\_\_\_\_\_\_\_属于机器学习中的有监督学习问题。

A.分类和聚类

B.回归和聚类

C.分类和回归

D.聚类和数据降维

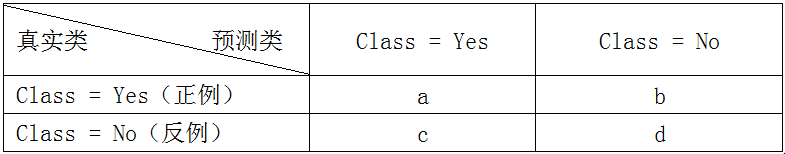
9.建模分析时，通常用于训练的样本数量\_\_\_\_\_\_\_\_\_测试的样本数量。

A.大于

B.小于

C.等于

D.小于等于

10. F1\_score可用于衡量分类模型性能，根据以下混淆矩阵，F1 = \_\_\_\_\_\_\_\_\_。

A.2a/(2a+b+c)

B.(a+d)/(a+b+c+d)

C.a/(a+c)

D.a/(a+b)

11.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

12.\_\_\_\_\_\_\_\_属于机器学习中的回归问题。

A.垃圾短信预测

B.房价预测

C.车牌识别

D.人脸识别

13.识别文本中的情感通常使用\_\_\_\_\_\_\_\_方法处理。

A.文本分类

B.文本聚类

C.自动问答

D.机器翻译

14.关于计算机数字图像的说法，错误的是\_\_\_\_\_\_\_\_。

A.数字图像存储每个像素点的颜色值

B.数字图像存储的是组成图像的几何形状、大小、颜色等信息

C.同样大小的图，存储使用的像素点越多，图像越清晰

D.JPEG是一种有损的图像压缩方式

15.天气预报主要采用\_\_\_\_\_\_\_\_数据处理技术。

A.Web

B.文本

C.图结构

D.时间序列

**二、多选题 ( 本大题 5 道小题 ，每小题 1 分，共 5 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所有正确答案。**

1.大数据的特征有\_\_\_\_\_\_\_\_。

A.规模性

B.高速性

C.多样性

D.低价值性

2.\_\_\_\_\_\_\_\_属于聚类问题。

A.根据企业校招历史数据，建立应聘者是否被录用的分类器

B.给定房屋特征数据，构建出估计房屋价格的模型

C.给定文档集，将相似的文档分到同一组

D.给定用户的消费数据，将用户分为不同消费特征的群体

3.\_\_\_\_\_\_\_\_\_可用于展示离散数据。

A.柱状图

B.饼图

C.折线图

D.曲面图

4.神经网络可用于\_\_\_\_\_\_\_\_\_等问题的建模分析。

A.电信用户分类

B.根据房屋特性预测房价

C.机动车识别

D.数据降维

5.智能语音对话系统，主要通过人工智能技术处理\_\_\_\_\_\_\_\_\_等数据实现。

A.语音

B.文本

C.图形

D.图像

**三、程序填空题 ( 本大题 4 道小题 ，每空 3 分，共 36 分）。**

1. 提示：

a) 题目源程序存放在"C:\KS"文件夹下，供程序调试；

b) Python科学计算库函数使用说明存放在"C:\KS"文件夹下，注意不同类库的函数存放在相应的sheet下。

某商品的成本（cost）可以根据产量（output）进行计算： cost=0.14\*output+42.7，编写程序模拟商品的生产数据，估计商品的成本（源程序fill\_1.py）。

1) 使用数组记录6次生产的商品产量（千件），分别为10、5、7、9、11、8；

2) 根据公式计算每次生产商品的成本；

3) 假设实际成本围绕计算的成本值上下波动，波动值服从均值为0、方差为2的正态分布，随机生成6个数据，模拟每次的波动；

4）加上波动值，计算6次生产商品的实际成本。

源程序文件（fill\_1.py）

#1）使用数组记录6次生产的商品产量（千件），分别为10、5、7、9、11、8；

output = 【1】

#2)根据公式计算每次生产商品的成本；

cost = 0.14\*output + 42.7

print( '1:cost: ',cost)

#3)实际成本围绕计算成本上下波动，波动值服从均值为0，方差为2的正态分布。

#随机生成6个数据，模拟每次的波动；

varcost = np.【2】(0,2,6)

print( '2:variance: ',varcost)

#4）加上波动值，计算6次生产商品的实际成本。

cost =【3】

print( '3:cost: ',cost)

2. 提示：

a) 题目源程序存放在"C:\KS"文件夹下，供程序调试；

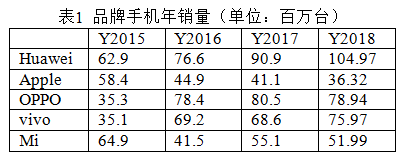
b) Python科学计算库函数使用说明存放在"C:\KS"文件夹下，注意不同类库的函数存放在相应的sheet下。

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。



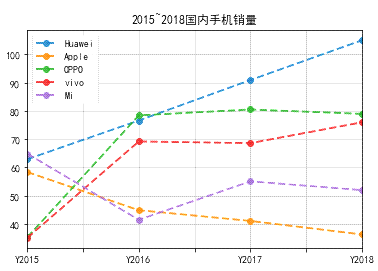


图1 手机销量折线图

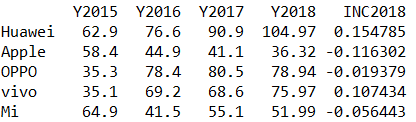


图2 增加列：2018年各品牌手机的同比增幅INC2018

源程序文件（fill\_2.py）

#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.show()

#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'] =【3】

print(sales)

3. 提示：

a) 题目源程序存放在"C:\KS"文件夹下，供程序调试；

b) Python科学计算库函数使用说明存放在"C:\KS"文件夹下，注意不同类库的函数存放在相应的sheet下。

表2和表3分别记录了部分"人工智能"类图书的一周销售数据（源程序fill\_3.py）。

1) 根据表2和表3分别创建数据对象，然后将两个数据对象合并，如表4所示；

2) 统计每家出版社出版的图书数，如图3所示；

3) 显示一周各出版社销售额，如图4所示。



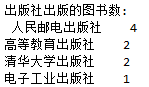
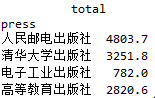
 

图3 每家出版社出版的图书数 图4 一周各出版社销售额

源程序文件（fill\_3.py）

import numpy as np

import pandas as pd

from pandas import DataFrame

#1)分别记录根据表2和表3中数据，然后合并

books1={"bookname": ['Python数据分析基础', '数据科学与大数据分析','机器学习','人工智能简史'],"press": ['人民邮电出版社', '高等教育出版社', '清华大学出版社', '人民邮电出版社'], "price": [38.9, 56.4, 45.2, 23.5], "sales": [25, 39, 44, 24]}

col\_name=['bookname','press','price','sales']

df1=DataFrame(books1,index=['A01','A02','A03','A04'], columns = col\_name)

print(df1)

books2={"bookname": ['Python程序设计', '数据科学导引','深度学习', '机器学习实战','TensorFlow框架',],"press": ['清华大学出版社', '高等教育出版社','人民邮电出版社', '人民邮电出版社', '电子工业出版社'], "price":[42.1, 34.5, 67.1, 56.0,78.2],"sales": [30,18,32,20,10]}

df2=DataFrame(books2,index=['B01','B02','B03','B04','B05'], columns = col\_name)

print(df2)

#合并df1和df2

df3=pd.【1】([df1,df2])

print("数据集合并后:\n",df3)

#2）统计每家出版社出版的图书数量

print("\n出版社出版的图书数:\n",df3['press'].【2】, "\n")

#3)显示一周各出版社销售额

df3['total'] = df3['price']\*df3['sales']

grouped = df3.groupby('press')

print( grouped.aggregate(【3】) )

4. 提示：

a) 题目源程序存放在"C:\KS"文件夹下，供程序调试；

b) Python科学计算库函数使用说明存放在"C:\KS"文件夹下，注意不同类库的函数存放在相应的sheet下。

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\_4.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\_4.py）

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.isnull())

【2】(inplace = True)

#3）输出达到超强台风等级的台风名

names = winds.loc[ 【3】,"windname" ].unique()

print("\n达到超强台风等级：\n", names )

**四、操作题**

**素材、样张、KS目录均在zip文件中，可双击此图标打开**

（一）、**简答题（共2题，每题5分，共10分）**

提示：打开C:\KS\Answer.doc文件，将简答题答案写在该文件的相应题目下并保存。

1. 请描述所学专业或日常生活中某个具体场景所涉及的数据，给出各项数据名称，说明以及数据的类型（连续数值/可选项/文本/图像/视频/声音/时序）等。

2. 试简述分类方法和聚类方法的区别，请根据实际案例所涉及的数据以及分析目标进行说明。

（二）、**综合应用题（共1题，34分）**

提示：打开"C:\KS"文件下的程序文件"prog.py"，按照程序注释说明，编写代码实现功能要求。

台风记录数据集（winds.csv）记录了2014年某区域发生的台风信息，包括台风名、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)。试分析与台风等级相关的特征，并建立等级判别模型。

具体要求如下：

1) 从文件中读出台风数据（3分）；

2) 数据集中表示台风等级level有六个等级为：热带低压、热带风暴、强热带风暴、台风、强台风、超强台风。将台风等级字符串依次替换为数字1-6（4分）；

3) 计算台风的各个特征与台风等级的相关性，筛选出相关性较高（相关系数>0.6）的特征建立数据集（5分）；

4) 绘制图形展示筛选出的特征与台风等级的相关性（4分）；

5) 按照合适比例将分析数据分为训练集和测试集（3分）；

6) 在训练集上建立分类模型，至少选用两种分类算法建立模型（7分）；

7) 在测试集上测试分类模型的性能（3分）；

8) 根据第7）步的运行结果，说明分类模型在台风等级判别上的性能，请描述在程序文件给出的注释行中（5分）。