南京大学信息管理学院2019-2020学年第一学期期末考试

《数据科学与数据分析》考试试卷（闭卷，A卷）

学号 ___ 姓名 __ 任课教师 裴雷、康乐乐 成绩 ____

一、单项选择题（每题1分，共10分）

1.在统计学中，一般区分样本数据是否属于小样本数据，确定的标准为（B）

A. 20　B. 30　C. 40　D. 50

2. 在数据预测或统计假设检验中，如事实数据为真、测试数据为假，则这类统计错误为（ A）

A. α错误　B. 取伪错误　C. II型错误　D. β错误

3. 安斯库姆四重奏说明了（ A ）

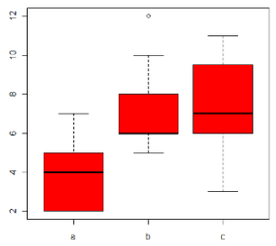
A. 数据统计值并不能完全代表真实的数据分布

B. 数据误差在统计上无法被描述

C. 数据分布差异在统计上可以显著地描述

D. 数据规模会影响数据分布效果

4. 为呈现下图所示的图形（颜色为红色）的效果，应该采用的函数是（D）



A. **plot(day~type,data=bac,col="red")**

B. **boxplot(day~type,data=bac,col="red")**

C. **ggplot(day~type,data=bac,col="red")**

D. **barplot(day~type,data=bac,col="red")**

5.使用R对内置鸢尾花数据集iris进行回归分析，选择Sepal.Length与Petal.Length进行线性回归测试，构建线性模型lm1 <- lm(Sepal.Length~Petal.Length)；运行summary（lm1）发现结果如左栏所示；而plot(lm1)后呈现右栏图形。请问下列判断成立的是（ ）

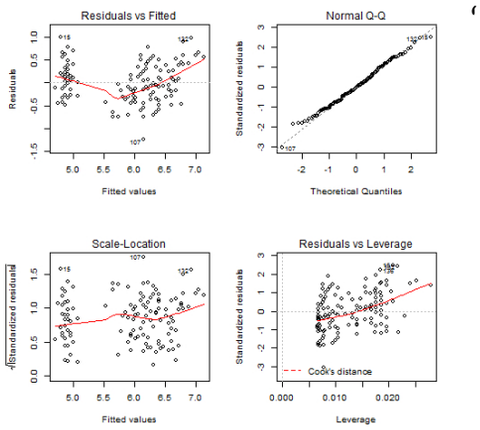
A. Sepal.Length与Pepal.Length 的线性回归系数为4.3066；

B. 该线性回归曲线通过坐标轴原点；

C. Sepal.Length与Pepal.Length 的相关关系在p=0.05的显著性水平上通过检验；

D. 在-2~2的杠杠区间外存在较多的离散点不支持该回归关系。

```
## lm(formula = Sepal.Length ~ Petal.Length)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2468 -0.2966 -0.0152  0.2768  1.0027
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.3066     0.0784    54.9   <2e-16 ***
## Petal.Length   0.4089     0.0189    21.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.407 on 148 degrees of freedom
## Multiple R-squared:  0.76,   Adjusted R-squared:  0.758
```



6.在广义回归模型中，为了判断和优化回归模型的变量，一般采用逐步寻优法则，在下面模型中，应该优先剔除的变量是（ ）

```
> # 逐步寻优法
> logit.step <- step(glm, direction = "both")
Start:  AIC=569
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8

       Df Deviance    AIC
- x2    1   551.54 567.54
- x8    1   551.79 567.79
- x5    1   552.25 568.25
<none>      551.00 569.00
- x1    1   554.95 570.95
- x6    1   555.93 571.93
- x4    1   573.03 589.03
- x7    1   594.37 610.37
- x3    1   635.21 651.21
```

A. X2

B. X1

C. X5

D. X3

7. ID3分类算法所使用的分类判断依据是（ ）

A. 信息增益　B. 信息增益比　C. GINI系数　D. 信息熵

8. 朴素贝叶斯预测需要每个条件概率非零，否则整个预测概率为零。如果出现了零概率的条件属性，一般会（ ）

A.条件属性数量较多时删除该条件　　B.会对概率进行拉普拉斯校正

C.会用平均概率值对缺失概率进行平滑处理　D.不考虑该条件对最终结果的影响

9. 在聚类过程中，关于complete类型的聚类算法（全链算法），描述不准确的是（ ）

A. 对噪音和离群不敏感　B. 对噪音和离群点很敏感

C 可能使大的簇破裂　D. 更偏好球形簇

10. 9个农业产区在7个维度统计中，通过计算Euclidean距离得到如下矩阵（9个农业区依次排列）。在AGNES算法中，当聚类方法采用single时，首先被聚类地区是（ ）

B8U7F]JUG1(C10)B8I4C8.png"　　*　　MERGEFORMATINET

$$D=(d_{ij})_{9\times9}=\begin{bmatrix}0\\1.52&0\\3.10&2.70&0\\2.19&1.47&1.23&0\\5.86&6.02&3.64&4.77&0\\4.72&4.46&1.86&2.99&1.78&0\\5.79&5.53&2.93&4.06&0.83&1.07&0\\1.32&0.88&2.24&1.29&5.14&3.96&5.03&0\\2.62&1.66&1.20&0.51&4.84&3.06&3.32&1.40&0\end{bmatrix}$$

A 1号和4号地区　B. 1号和2号地区　C. 4号和9号地区　D. 5号和7号地区

二、填空题（每空1分，共20分）

11. Aprior算法先验性原理认为：频繁项集的子集必是_____; 非频繁项集的超集也是_____。

12. 线性回归分析通常采用stats包中的____函数；而广义线性分析则是____函数。

13. CART算法通常调用____数据包中的____函数。

14. C4.5算法通常调用____函数包中的____函数。

15.人工神经网络可以调用的函数为：_____程序包_____函数。

16. 朴素贝叶斯算法通常调用两个函数包：e1071包中的_____函数和klaR 包 中的_____函数。

17. K-均值算法在R语言中实现的核心函数为____数据包中的____函数。

18. 基于中心点聚类的k-medoids算法最常调用的函数为____数据包中的____函数。

19. 支持向量机(SVM) 算法通常调用____函数包中的____函数。

20. 在关联规则发现中，如果需要进行apriori算法的应用，通常会调用_____数据包中的____函数。

3、 计算分析题（共5题，40分）

21.分组检验是统计中最常见的方法，请说明以下情景分别需要使用哪种推断统计方法。

（1）35岁年龄人口中，男性的智商比较高，还是女性的智商比较高？用哪种统计检验方法进行对比？（2分）

（2）35岁年龄人口中，北上广、江浙沪、珠三角、其他地区，这四组中，哪个地区的人智商比较高？用哪种统计检验方法进行对比？（2分）

（3）对于上述问题2，进行假设检验时，原假设和备择假设分别是什么？组间差异大更容易发现各组有显著差异，还是组内差异大更容易发现各组有显著差异？为什么？（4分）

22.

dplyr是最常见的R数据预处理包。（1）请写出此包中包含的主要函数和用途。（4分）（2）如下数据集是几个用户1周内在某网站的交易记录，**请写出相关R或者Python的代码**，计算出**每个用户这个月购物的次数、购物的总金额**。如使用R，请用dplyr中的相关函数。(4分)

用户ID	交易时间	交易金额	物品
0001	1	25.00	衣服
0002	2	38.00	衣服
0003	1	29.00	玩具
0001	2	65.00	衣服
0002	1	256.00	电器
0002	1	57.00	玩具
……			

23.

完整ggplot有7层结构。下述代码中，已经展示了ggplot最基本的3层，请问这三层分别是什么？（3分）mindwest对应什么层？（1分）

geom_point()对应什么层？（1分）与geom_point()所在层，还有其他三十多种函数，请写出其中3个。（3分）

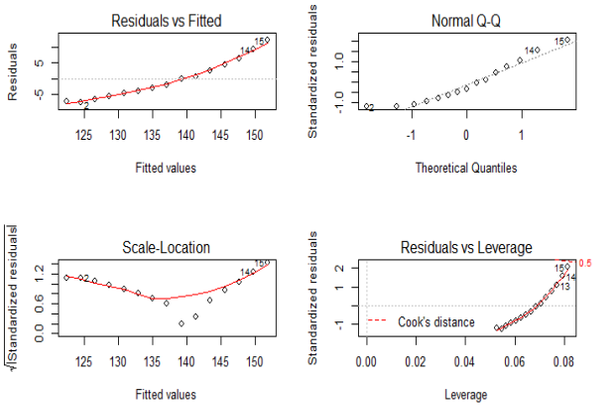
```
library(ggplot2)
ggplot(midwest, aes(x=area, y=poptotal)) + geom_point()
```

24.

在信息管理学院的本科学习阶段，学生会接触不同的程序设计语言和计算工具。下表是某班8名学生的程序设计语言的学习情况，请分别计算R语言学习对MySQL学习的支持度和置信度（4分）。同时，也请计算向学习R和Python的同学推荐学习Gephi语言策略的支持度和置信度。（4分）

学生列表	程序设计语言与数据库工具学习情况
牛二	C、C++、C#、Java、R、SQLSever
张三	C++、R、Python、MySQL、Gephi
李四	R、Python、PHP、MySQL、MogoDB
王五	C、ASP、MySQL、FoxPro、DB2、Gephi
赵六	Python、MySQL、MogoDB
褚七	R、Java、MySQL、Oracle
郑八	Python、MySQL、Gephi
华九	R、Python、MySQL、Gephi、MogoDB

25. 在women数据集中，通过建立lm.model <- lm(weight ~ height - 1, data = women)的线性回归模型，得到如下诊断图像。对此你有哪些基本判断？应该怎样改进该回归模型？（8分）



四、分析诊断题（共2题，28分）

26.

有一个二分类问题的数据集如下：一共有20位顾客的分类数据信息如下，下面以基尼系数（GINI系数）为计算方法，分别计算性别属性、衬衣尺码属性、车型属性的GINI指标值（9分），并比较性别、衬衣、车型哪个属性作为首次条件划分的结果要好（5分）。

顾客ID	性别	车型	衬衣尺码	类	顾客ID	性别	车型	衬衣尺码	类
1	男	家用	小	C0	11	男	家用	大	C1
2	男	运动	中	C0	12	男	家用	加大	C1
3	男	运动	中	C0	13	男	家用	中	C1
4	男	运动	大	C0	14	男	豪华	加大	C1
5	男	运动	加大	C0	15	女	豪华	小	C1
6	男	运动	加大	C0	16	女	豪华	小	C1
7	女	运动	小	C0	17	女	豪华	中	C1
8	女	运动	小	C0	18	女	豪华	中	C1
9	女	运动	中	C0	19	女	豪华	中	C1
10	女	豪华	大	C0	20	女	豪华	大	C1

提示：

	男	女		小	中	大	加大		家用	运动	豪华
C0	6	4	C0	3	3	2	2	C0	1	8	1
C1	4	6	C1	2	4	2	2	C1	3	0	7

27.协同过滤算法(collaborative filtering)本质是基于相似度计算的，其中最常用的一种相似度计算方法是Cosine Similarity，它具体的公式如下：（总分14分）

$$Cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\ \|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2}\ \sqrt{\sum_1^n b_i^2}}$$

where, $\vec{a}\cdot\vec{b} = \sum_1^n a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ is the dot product of the two vectors.

(1) 有如下向量a和b，请计算其Cosine Similarity的值。a = [1,2,3],b=[4,-5,6] (开平方根直接保留，无需计算)（3分）

(2) 有如下4位个人数据，分别利用简单匹配系数(Simple Matching Coefficients, SMC)、Jaccard系数(Jaccard Coefficients, JC)、Cosine S

imilarity计算出前三位与宋江的距离，并说明，谁最像宋江。（9分）

姓名	年龄	性别	月收入	工作年限	户籍性质
张三	35	男	10	3	农村
李四	20	女	15	2	农村
王二	25	男	10	1	城市
宋江	**30**	**男**	**20**	**1**	**城市**

(3)

推荐系统依赖于协同过滤。请简要说明，为什么较多推荐系统进行的是物品与物品之间的协同过滤，而非人与人之间的。（2分）

五、课程考勤题（共1题，2分）

28.《数据科学与数据分析》课程共有几位助教，他们名字是什么?

有一个二分类问题的数据集如下：一共有20位顾客的分类数据信息如下，下面以基尼系数（GINI系数）为计算方法，分别计算性别属性、衬衣尺码属性、车型属性的GINI指标值（9分），并比较性别、衬衣、车型哪个属性作为首次条件划分的结果要好（5分）。

顾客ID	性别	车型	衬衣尺码	类	顾客ID	性别	车型	衬衣尺码	类
1	男	家用	小	C0	11	男	家用	大	C1
2	男	运动	中	C0	12	男	家用	加大	C1
3	男	运动	中	C0	13	男	家用	中	C1
4	男	运动	大	C0	14	男	豪华	加大	C1
5	男	运动	加大	C0	15	女	豪华	小	C1
6	男	运动	加大	C0	16	女	豪华	小	C1
7	女	运动	小	C0	17	女	豪华	中	C1
8	女	运动	小	C0	18	女	豪华	中	C1
9	女	运动	中	C0	19	女	豪华	中	C1
10	女	豪华	大	C0	20	女	豪华	大	C1

提示：

	男	女		小	中	大	加大		家用	运动	豪华
C0	6	4	C0	3	3	2	2	C0	1	8	1
C1	4	6	C1	2	4	2	2	C1	3	0	7