阿里巴巴2016数据挖掘工程师笔试

一. 单项选择题

1. 想要了解上海市小学生的身高,需要抽取500个样本,这项调查中的样本是?

A 从中抽取的500名学生的身高

B 上海市全部小学生的身高

C 从中抽取的500名小学生

D 上海市全部小学生

正确答案：A

2. 以下对k-means聚类算法解释正确的是

A 能自动识别类的个数,随即挑选初始点为中心点计算

B 能自动识别类的个数,不是随即挑选初始点为中心点计算

C 不能自动识别类的个数,随即挑选初始点为中心点计算

D 不能自动识别类的个数,不是随即挑选初始点为中心点计算

正确答案：C

3. 以下哪个是常见的时间序列算法模型

A RSI

B MACD

C ARMA

D KDJ

正确答案：C

4. 有个袋子装有2个红球,2个蓝球,1个黄球,取出球以后不再放回,请问取两次出来的球是相同颜色的概率是多少

A 0.3333

B 0.25

C 0.2

D 0.1667

正确答案：C

5.
65,8,50,15,37,24,()。括号中的数字是()

A 25

B 26

C 22

D 27

正确答案：B

6. 一组数据,均值>中位数>众数,问这组数据

A 左偏

B 右偏

C 钟型

D 对称

正确答案：B

7. SQL语言允许使用通配符进行字符串匹配的操作,其中'%'可以表示

A 零个字符

B 1个字符

C 多个字符

D 以上都可以

正确答案：D

8. 关于正态分布,下列说法错误的是:

A 正态分布具有集中性和对称性

B 正态分布的均值和方差能够决定正态分布的位置和形态

C 正态分布的偏度为0，峰度为1

D 标准正态分布的均值为0，方差为1

正确答案：C

9. 在以下不同的场景中,使用的分析方法不正确的有

A 根据商家最近一年的经营及服务数据,用聚类算法判断出天猫商家在各自主营类目下所属的商 家层级

B 根据商家近几年的成交数据,用聚类算法拟合出用户未来一个月可能的消费金额公式

C 用关联规则算法分析出购买了汽车坐垫的买家,是否适合推荐汽车脚垫

D 根据用户最近购买的商品信息,用决策树算法识别出淘宝买家可能是男还是女

正确答案：B

10. 下列时间序列模型中,哪一个模型可以较好地拟合波动性的分析和预测

A AR模型

B MA模型

C ARMA模型

D GARCH模型

正确答案：D

二. 多选选择题

11. excel工作簿a中有两列id、age,工作簿b中有一列id,需要找到工作薄b中id对应的age,可用的函数包括

A index+match

B vlookup

C hlookup

D find

E if

F like

正确答案：A,B

12. 现在有M个桶,每桶都有N个乒乓球,乒乓球的颜色有K种,并且假设第i个桶第j种颜色的球个数为Cij, 比例为Rij=Cij/N,现在要评估哪个桶的乒乓球颜色纯度最高,下列哪种算法和描述是合理的?

A ∑(N/K-Cij)(N/K-Cij)越小越纯

B -∑Cij*LOG(Rij)越小越纯

C ∑(1-Rij*Rij)越小越纯

D ∑(1-Rij)*(1-Rij)越小越纯

E ∑(1-Rij)^2 越小越纯

F -∑Rij*LOG(Rij)越小越纯

正确答案：C,D,E,F

13. 关于相关系数,下列描述中正确的有:

A 相关系数为0.8时,说明两个变量之间呈正相关关系

B 相关系数等于1相较于相关系数等于-1,前者的相关性更强

C 相关性等于1相较于相关系数等于0,前者的相关性更强

D Pearson相关系数衡量了两个定序变量之间的相关程度

E Spearman相关系数可以衡量两个定序变量之间的相关程度

F 相关系数为0.2相较于-0.8,前者的相关性更强

正确答案：A,C,E

14. 关于线性回归的描述,以下正确的有:

A 基本假设包括随机干扰项是均值为0,方差为1的标准正态分布

B 基本假设包括随机干扰下是均值为0的同方差正态分布

C 在违背基本假设时,普通最小二乘法估计量不再是最佳线性无偏估计量

D 在违背基本假设时,模型不再可以估计

E 可以用DW检验残差是否存在序列相关性

F 多重共线性会使得参数估计值方差减小

正确答案：B,C,E

15. 下列哪些方法可以用来对高维数据进行降维:

A LASSO

B 主成分分析法

C 聚类分析

D 小波分析法

E 线性判别法

F 拉普拉斯特征映射

正确答案：A,B,D,E,F

三. 问答题

16.
查询成交表a中的城市city的成交金额大于0的购买人数(buyer_id)和成交金额(amt)
city buyer_id order_id amt
a 1 1 100
a 1 2 100
b 2 3 100
b 3 4 20
c 4 5 0

正确答案：

```
select buyer_id,sum(amt) as amt from a
where city in
(
select city from
(
select city,sum(amt) as amt from a
group by city
)t
where t.amt>0
)
```

17. 公司要构建淘宝商家健康指数,所以要对最近1年内有交易的淘宝商家进行问卷调研。为不过于打搅商 家,问卷调研采取抽样的方式进行确定商家名单。怎么抽样比较好?

正确答案：可以考虑采用分层抽样的方式。 首先根据销售额或销售量对商家进行分层,这样可能会将商家分为高销售额(量) 商户,中销售额(量)商户,低销售额(量)商户等,然后根据这三者的比例确定 各个层次应抽取的商户数。 对抽取出来的样本,根据相应的指标,如访问量、购买量、买家评级,评论数,发 货速度等指标来综合考虑商家的健康指数。