## 大数据分析与挖掘复习题

### 一、单选题

1、请问在Python中缺失值通常用（ ）来记号。

A NaN B ? C NA D na

参考答案：A

2、请问异常点通常出现在盒须图的（ ）。

A 中部 B 上部 C 下部 D 两端

参考答案：D

3、平稳的时间序列：（ ）

A. 均值随时间变化而变化

B. 协方差随时间变化而变化

C. 分布与时间无关

D. 分布与时间有关

参考答案：C

4、回归模型评价时，通常采用（ ）损失。

A 0-1 B 指数 C 负二项 D 平方

参考答案：D

5、1950年1月3日到2008年4月11日的标准普尔指数值有一种无固定规律的交替波动，这种变动被称为（ ）

A. 长期趋势变动

B. 季节变动

C. 循环波动

D. 不规则变动

参考答案：C

6、复杂的模型相较于简单模型（ ）。

A 训练误差更大 B 偏差更大 C 方差更大 D 期望误差更小

参考答案：C

7、典型相关分析解决的问题是（ ）

A 两个变量之间的线性相关关系 B 两个变量之间的非线性相关关系

C 两组变量之间的线性相关关系 D 两组变量之间的非线性相关关系

参考答案：C

8、线性回归模型的拟合优度的判定系数越大，说明（ ）

A. 回归方程解释能力较差 B. 回归方程解释能力较好

C. 回归方程预测能力较差 D. 回归方程预测能力较好

参考答案：B

9、数据离散化可以采用（ ）方式。

A 等宽或者等频 B 插补 C 抽样 D 查询

参考答案：A

10、数据抽样后（ ）。

A 数据量会增加 B 数据量会减少 C 数据的波动性增加 D 数据波动性不变

参考答案：B

11、聚类分析的内部评价指标通常有（ ）。

A DB指数 B 兰德指数 C Jaccard系数 D FM指数

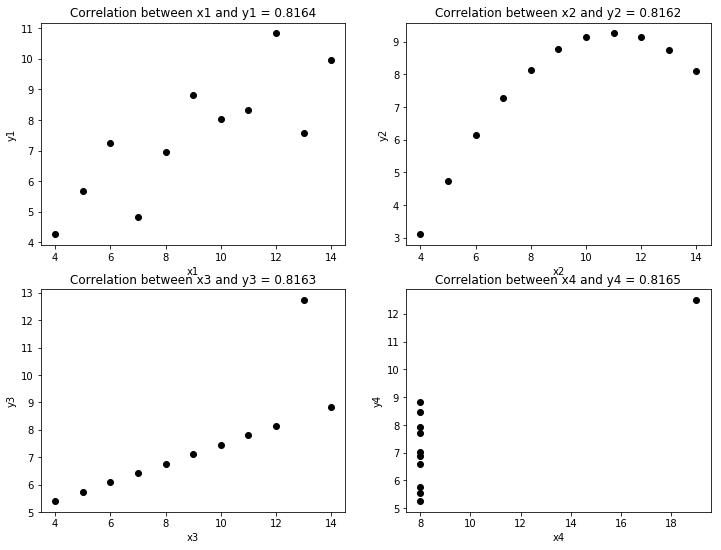
参考答案：A

12、期望预测误差是指（ ）。

A 真实值与预测值在某种损失函数下的平均值 B 真实值与预测值之差 C 真实值与预测值在某种损失函数下的值 D 真实值与预测值之比

参考答案：A

13、下面图中看出两个变量之间真实的线性相关系数为0.8164的变量对是（ ）。



A x1与y1 B x2与y2

C x3与y3 D x4与y4

参考答案：A

14、在关联分析中，下面哪个有可能是频繁3项集（ ）

A {面包，牛奶} B {面包，牛奶，啤酒}

C {面包} D {面包，牛奶，花生，啤酒}

参考答案：B

15、某超市研究销售纪录数据后发现，买啤酒的人很大概率也会购买尿布，这种属于数据挖掘的哪类问题？( )

A.关联规则发现 B.聚类

C.分类 D.自然语言处理

参考答案：A

16、线性回归模型显著性检验的P值越小，说明（ ）

A. 对应的自变量对因变量影响越大 B. 对应的自变量对因变量影响越小

C. 变量之间的非线性关系明显 D. 变量之间的非线性关系不明显

参考答案：A

17、要考虑“储蓄水平”与 “人口水平”之间的相关。“储蓄水平”收集了储蓄比率（sr）、人均可支配收入（dpi）、人均可支配收入变化的百分率（ddpi）3个变量，“人口水平”收集了15岁以下人口的百分比、75岁以上人口百分比2个变量。请问采用下面哪种分析方法更合适（ ）。

A 单向关

B 典型相关

C 偏相关

D 点二列（点双列）相关

参考答案：B

18、在考虑为网页点击率寻找相关变量的问题中说法不正确的是（ ）

A.是否是工作日可以作为自变量

B.地理位置可以作为自变量

C.网页登入方式可以作为因变量

D.网页点击率可以作为因变量

参考答案：C

19、关于随机森林算法的说法错误的是：（ ）

A. 随机森林由很多的决策树组成

B. 随机森林对输入的数据要进行行、列的采样

C. 构成随机森林的决策树越多，随机森林的预测精度越高

D. 随机森林能够处理很高维度（特征很多）的数据

参考答案：C

20、下面关于线性判别与Bayes判别说法错误的是：（ ）

A.距离判别没有考虑到每个分类的观察值不同时，每类出现的机会是不同的，

B.距离判别没有考虑到误判所造成的损失差异

C.Bayes判别依据每个被判入某个类别的后验概率进行归类。

D.后验概率越小的类就是被判别的类

参考答案：D

21、数据整合后（ ）。

A 数据量会增加 B 数据量会不变 C 数据的波动性增加 D 数据波动性减小

参考答案：D

22、分类模型评价时，通常采用（ ）损失。

A 0-1 B 指数 C 绝对值 D 平方

参考答案：A

23、取暖器销售量随着季节不同而发生较大变动，冬季的销售量一般高于冬季的销售量。取暖器销售量数据的这种变动称为（ ）

A. 长期趋势变动

B. 循环波动

C. 季节变动

D. 不规则变动

参考答案：C

24、混淆矩阵中的TP数是指（ ）。

A 真阳数 B 假阳数 C 真阴数 D 假阴数

参考答案：A

25、下面关于单位根检验说法不正确的是（ ）

A. 具有单位根的时间序列是非平稳的

B. 具有单位根的时间序列是指原始序列是平稳的

C. 具有单位根的时间序列是指将原始序列差分后可以平稳化的时间序列

D. 单位根检验用于检验时间序列的平稳性

参考答案：B

26、在研究父母的身高如何影响孩子的身高的问题中，得到如下的回归分析的结果：

Out [3]:

LinregressResult(slope=0.6462905819936418, intercept=23.941530180412784, rvalue=0.45876236829282285, pvalue=1.732509292015251e-49, stderr=0.041135882237933415)

r\_square: 0.21046

请问关于上面的结果叙述不正确的是（ ）

A.拟合优度的判定系数R2（r-Square）为0.21046

B.该回归分析拟合程度不高

C.此次分析采用的线性回归分析

D.孩子身高是自变量

参考答案：D

27、下面关于Pearson相关系数的理解不正确的是（ ）。

A 是两个变量之间线性关系的度量指标

B 反映两变量之间的因果关系

C 不具有传递性

D 是由数据计算得到的

参考答案：B

28、关于抽样后的数据的大小说法正确的是（ ）。

A 抽样比例越低，抽样数据越大 B 抽样样本量越小，抽样数据越大

C 抽样比例越高，抽样数据越大 D 抽样样本量越大，抽样数据越小

参考答案：C

29、使用一份某运营商用户订购套餐情况数据进行分析后发现，开通来电提醒业务的用户也都倾向于开通邮件功能，这种解决问题的方法方法属于数据挖掘的哪类方法？( )

A.关联规则发现 B.聚类

C.分类 D.自然语言处理

参考答案：A

30、下面关于Pearson相关系数的数值叙述不正确的是（ ）。

A Pearson相关系数的数值是由数据计算出来的

B Pearson相关系数的数值大小与计量尺度无关

C Pearson相关系数的数值大小不能完全反映因果关系

D Pearson相关系数的数值不具有对称性

参考答案：D

31、请问下面的哪个模型是移动平均模型（ ）

A.

B.

C.

D.

参考答案：C

32、关于设定虚拟变量时应当遵循如下原则，下列说法错误的是：( )

A.对于有k个表现值的定性变量，只设定（k-1）个虚拟变量；

B.虚拟变量的值通常用“0”或“1”来表示；

C.对于每个样本而言，同一个定性变量对应虚拟变量的值之和不超过1；

D.对于季节变量而言，四个季节需要设定4个虚拟变量

参考答案：D

33、典型相关分析中的综合变量选择问题通常采用（ ）来度量。

A 特征根贡献率

B 特征向量

C 相关系数

D p值

参考答案：A

34、在为员工报酬寻找相关变量的问题中说法不正确的是（ ）

A.员工职位可以作为自变量

B.员工年龄可以作为解释变量

C.员工性别可以作为相依变量

D.员工工资可以作为因变量

参考答案：C

35、关于下面的决策树说法不正确的是（ ）。

图片包含 图示

描述已自动生成

A. 此决策树的构建基于样本量为100的样本

B. 此样本是三分类的目标变量

C. 第一层的节点的最优划分是Petal.Length≤2.45

D. Petal.Length≤2.45的节点还可以继续划分

参考答案：D

36、关于支持向量机算法的说法错误的是：（ ）

A. 超平面到所有两类数据点的距离之和称为分类器的边缘

B. 具有最大边缘的超平面即是我们需要的最大边缘超平面。

C. 以超平面以及相应的分类决策函数来分类的支持向量机称为线性可分支持向量机。

D. 非线性支持向量机通过引进核函数解决线性不可分问题。

参考答案：A

37、下面关于快速聚类说法错误的是：（ ）

A. 快速聚类又叫K-Means聚类

B. 快速聚类需要事先已知类别的个数k

C. 快速聚类过程通过中心的迭代完成

D. 快速聚类也就是层次聚类

参考答案：D

38、下面关于时间序列的差分方法的说法不正确的是（ ）

A. 差分可以用于将非平稳时间序列平稳化

B. 差分是指用每一个数值减去其前面的一个数值，即

C. 差分方法不可以用每一个数值减去其前面的、任何间隔为的一个数值，即

D. 对原始数据进行一次差分的过程被称为一阶差分

参考答案：C

39、DBSCAN算法中的核心对象是（ ）

A. 对给定对象ε邻域内的样本点数小于MinPts的对象

B. 对给定对象ε邻域内的样本点数大于等于MinPts的对象

C. 对给定对象ε邻域内的样本点数小于等于MinPts的对象

D. 对给定对象ε邻域内的样本点数大于MinPts的对象

参考答案： B

40、近邻分类中的近邻个数说法正确的是：（ ）

A.1-近邻最容易受敏感点的影响

B.随着近邻个数的增加，模型的复杂度增加

C.近邻个数不会影响模型的预测能力

D.近邻个数越大，模型越不稳定

参考答案：A

41、下面关于主成分分析说法错误的是（ ）。

A 主成分是相互正交（垂直）的 B 主成分捕获了数据差异最大的方向

C 是线性降维方法 D 是非线性降维方法

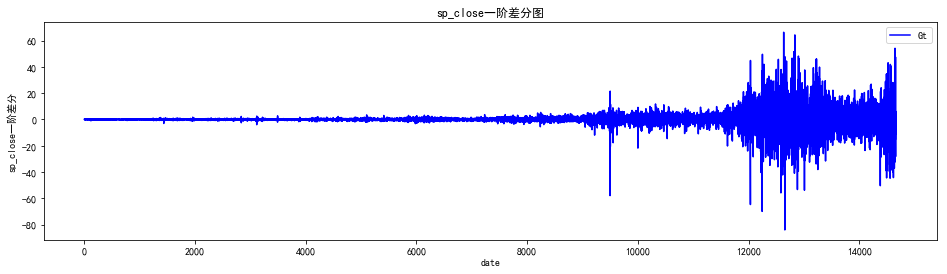
参考答案：D

42、回归模型评价时，通常采用（ ）损失。

A 0-1 B 指数 C 负二项 D 平方

参考答案：D

43、从下面的时序图一定可以得到的结论是（ ）



A. 该序列具有零均值

B. 该序列具有同方差性

C. 这是一个白噪声序列

D. 这是一个平稳时间序列

参考答案：A

44、如果支持度阈值设定为0.6，那么关于下面的购物篮事务数据叙述正确的是（ ）

|  |  |  |
| --- | --- | --- |
|  | **TID** | **项集** |
| **1** | 1 | {面包，牛奶} |
| **2** | 2 | {面包，尿布，啤酒，鸡蛋} |
| **3** | 3 | {牛奶，尿布，啤酒，可乐} |
| **4** | 4 | {面包，牛奶，尿布，啤酒} |
| **5** | 5 | {面包，牛奶，尿布，可乐} |

A {面包，牛奶}不是一个频繁项集

B {啤酒，尿布，牛奶}是一个频繁项集

C {面包}是一个频繁项集

D {面包，尿布}不是一个频繁项集

参考答案：C

45、关于非线性回归，下面说法正确的是( )

A.非线性回归总是比线性回归结果好

B.非线性回归没有回归模型结果好

C.有的非线性回归可以转化为线性回归

D.非线性回归不需要做模型检验

参考答案：C

46、数据离散化可以采用（ ）方式。

A 等宽或者等频 B 插补 C 抽样 D 查询

参考答案：A

47、数据抽样后（ ）。

A 数据量会增加 B 数据量会减少 C 数据的波动性增加 D 数据波动性不变

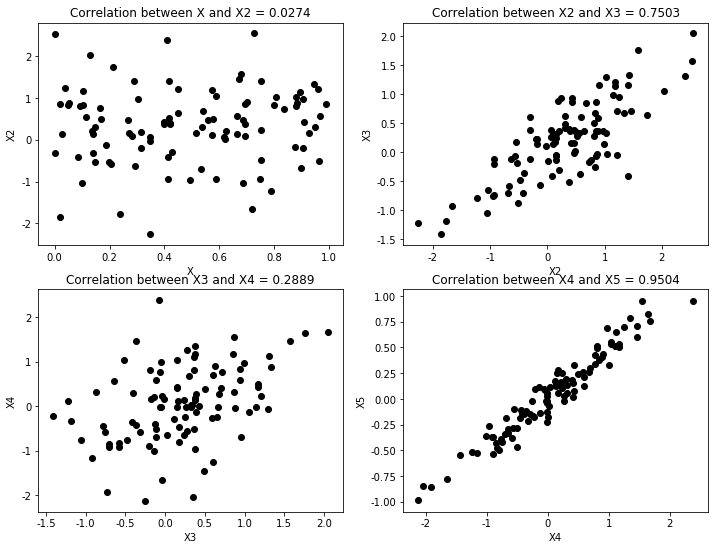
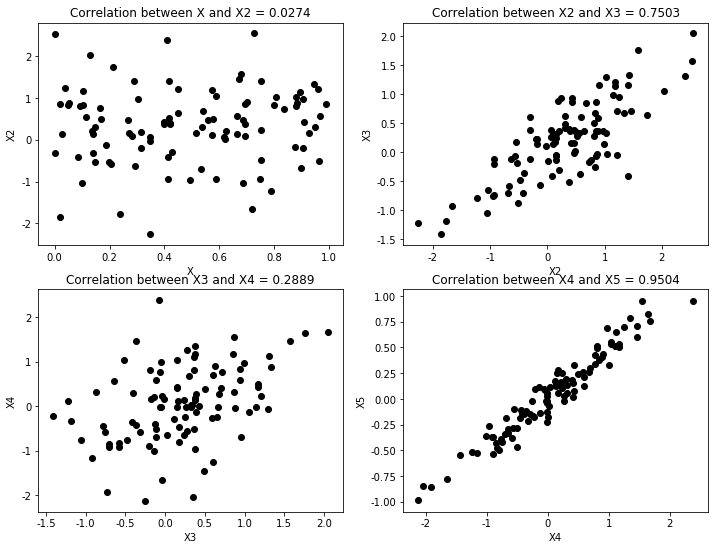
参考答案：B

48、期望预测误差是指（ ）。

A 真实值与预测值在某种损失函数下的平均值 B 真实值与预测值之差 C 真实值与预测值在某种损失函数下的值 D 真实值与预测值之比

参考答案：A

49、从下面两个变量的相关系数图可以看出（ ）



A 左图相关系数大于0 B 右图相关系数小于0

C 左图相关系数等于0 D 右图相关系数等于0

参考答案：A

50、下面关于Pearson相关系数的理解不正确的是（ ）。

A 是两个变量之间线性关系的度量指标

B 反映两变量之间的因果关系

C 容易受到异常点或者偏离点影响的度量

D 不能够度量变量之间的非线性相关程度

参考答案：B

51、ARMA模型也被叫做( )

A. 自回归移动平均模型

B. 自回归模型

C. 移动平均模型

D. 整合自回归移动平均模型

参考答案：A

52、关于设定虚拟变量时应当遵循如下原则，下列说法错误的是：( )

A.对于有k个表现值的定性变量，只设定（k-1）个虚拟变量；

B.虚拟变量的值通常用“0”或“1”来表示；

C.对于每个样本而言，同一个定性变量对应虚拟变量的值之和不超过1；

D.对于季节变量而言，四个季节需要设定4个虚拟变量

参考答案：D

53、依据因变量的条件分位数对自变量进行回归的模型是（ ）。

A. 线性回归模型

B. Logistic回归模型

C. 分位数回归

D. 非线性回归

参考答案：C

54、在为员工报酬寻找相关变量的问题中说法不正确的是（ ）

A.员工职位可以作为自变量

B.员工年龄可以作为解释变量

C.员工性别可以作为相依变量

D.员工工资可以作为因变量

参考答案：C

55、下面关于线性可分与不可分的理解不正确的是（ ）。

A. 线性可分的是指可以找到一个超平面将一类数据与另一类完全分开

B. 线性不可分的则是指不能找到一个超平面将一类数据与另一类完全分开

C. 线性不可分问题可以通过空间变换实现线性可分

D. 线性不可分问题不能转成线性可分问题

参考答案：D

56、下面关于随机森林的理解不正确的是（ ）。

A. 能够处理很高维度（特征很多）的数据

B. 能够检测到特征间的交互作用

C. 容易做成并行化方法

D. 需要事先做特征选择

参考答案：D

57、关于下面的贝叶斯判别结果说法不正确的是（ ）

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Setosa** | **versicolor** | **virginica** |
| **0** | 1.000000e+00 | 5.048809e-17 | 1.869461e-25 |
| **1** | 6.250263e-171 | 2.635294e-02 | 9.736471e-01 |
| **2** | 1.937522e-315 | 2.179357e-12 | 1.000000e+00 |
| **3** | 3.729519e-116 | 9.868228e-01 | 1.317724e-02 |
| **4** | 1.000000e+00 | 5.634438e-18 | 4.545093e-26 |

A. 编号0的观测被判为**Setosa**的概率最大

B. 编号4的观测将会被判为**Setosa**一类

C. 编号3的观测将会被判为**versicolor**一类

D. 编号1的观测将会被判为**versicolor**一类

参考答案：D

58、下面关于决策树的剪枝操作理解不正确的是（ ）。

A. 对决策树进行剪枝的目的是提高模型的预测误差

B. 对决策树进行剪枝是为了提高模型的解释能力

C. 只能在生成决策树后再剪枝

D. 剪枝就是从下往上的剪掉一些节点

参考答案：C

59、DBSCAN算法中的直接密度可达、密度可达与密度相连理解不正确的是（ ）

A. 只有核心对象之间相互密度可达

B. 密度相连是对称关系

C. DBSCAN目的是找到密度相连对象的最大集合

D. 密度可达也是对称关系

参考答案： D

60、聚类分析中有多种类间距的定义，聚类结果会由所不同。那么下面不是聚类方法取舍时应遵循的原则是：（ ）

A. 各类重心（常用平均数衡量）之间应该有最小的距离；

B. 确定的类中，各类所包含的元素都不宜过分多

C. 聚类数目应符合实际

D. 当用许多方法进行分类时，应选出现次数最多的那种分类结果

参考答案：A

61、请问不是缺失值通常用的处理方式是（ ）。

A 填补 B 删除 C 插补 D 提取

参考答案：D

62、重复行的数据是指（ ）。

A 某行数据与另一行数据完全相同

B 某行数据与另一行数据部分相同

C 某列数据与另一列数据完全相同

D 某列数据与另一列数据部分相同

参考答案：A

63、苹果公司2010年1月4日至2011年12月7日股票数据中有呈直线上升的趋势，这种变动是（ ）

A. 不规则变动

B. 季节变动

C. 循环波动

D. 长期趋势变动

参考答案：D

64、混淆矩阵中的FN数是指（ ）。

A 真阳数 B 假阳数 C 真阴数 D 假阴数

参考答案：D

65、典型相关分析解决的问题是（ ）

A 两个变量之间的线性相关关系 B 两个变量之间的非线性相关关系

C 两组变量之间的线性相关关系 D 两组变量之间的非线性相关关系

参考答案：C

66、下面关于单位根检验说法不正确的是（ ）

A. 具有单位根的时间序列是非平稳的

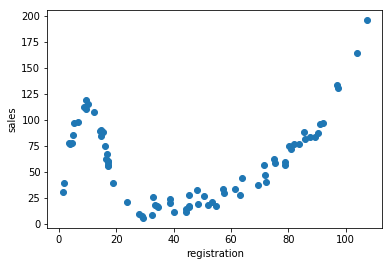
B. 具有单位根的时间序列是指原始序列是平稳的

C. 具有单位根的时间序列是指将原始序列差分后可以平稳化的时间序列

D. 单位根检验用于检验时间序列的平稳性

参考答案：B

67、为研究电商注册用户数量与其销售收入之间的关系，收集数据得到下面的散点图。请问这样的散点图适用建立下面哪种回归模型（ ）。



A. 线性回归模型

B. 非线性回归模型

C. 对数线性模型

D. Logistic回归模型

参考答案：B

68、数据经过标准化处理后（ ）。

A 数据的波动性将会增加 B 数据的波动单位为1

C 数据的波动性将会减小 D 数据值将被放大

参考答案：B

69、下面关于系统聚类说法错误的是（ ）

A. 系统聚类也叫层次聚类

B. 系统聚类主要依据个体之间的距离来聚类

C. 系统聚类中可以使用各种距离度量

D. 系统聚类不能使用绝对值距离

参考答案：D

70、期望预测误差是指（ ）。

A 真实值与预测值在某种损失函数下的平均值 B 真实值与预测值之差 C 真实值与预测值在某种损失函数下的值 D 真实值与预测值之比

参考答案：A

71、下面关于Pearson相关系数的叙述不正确的是（ ）。

A Pearson相关系数的数值是由数据计算得到的

B Pearson相关系数的数值大小容易受到异常点的影响

C Pearson相关系数不能完全反映因果关系

D 相关系数取值的大小与x和y的计量尺度有关

参考答案：D

72、某超市研究销售纪录数据后发现，买啤酒的人很大概率也会购买尿布，这种属于数据挖掘的哪类问题？( )

A.关联规则发现 B.聚类

C.分类 D.自然语言处理

参考答案：A

73、请问下面的哪个模型是自回归模型（ ）

A.

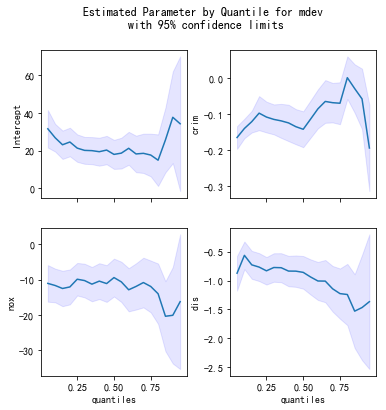
B.

C.

D.

参考答案：B

74、关于下面的拟合图叙述不正确的是（ ）。



A.图中是95%的置信区间估计

B.图中是分位数回归的估计结果

C.当mdev越大，受crim的影响也越来越大

D.当mdev越大，受dis的影响越来越小

参考答案：D

75、下面关于时间序列的差分方法的说法不正确的是（ ）

A. 差分可以用于将非平稳时间序列平稳化

B. 差分是指用每一个数值减去其前面的一个数值，即

C. 差分方法不可以用每一个数值减去其前面的、任何间隔为的一个数值，即

D. 对原始数据进行一次差分的过程被称为一阶差分

参考答案：C

76、下面的距离度量中，欧氏距离是（ ）

A.

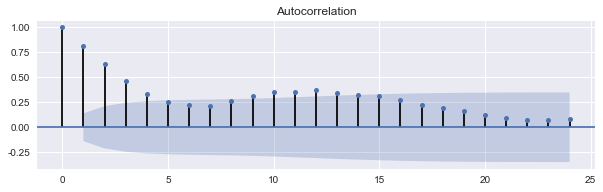
B.

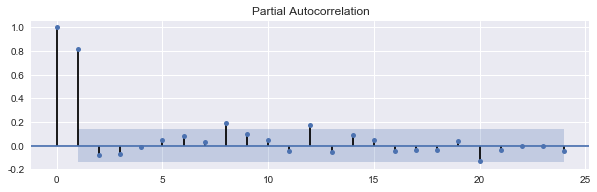
C.

D.

参考答案： B

77、自相关系数图与偏自相关图如下。下面的说法不正确的是（ ）

****

****

A. 自相关系数具有拖尾性

B. 偏自相关系数具有截尾性

C. 此序列可以认为是AR(1)模型

D. 此序列可以认为是MA(1)模型

参考答案：D

78、关于抽样后的数据的大小说法正确的是（ ）。

A 抽样比例越低，抽样数据越大 B 抽样样本量越小，抽样数据越大

C 抽样比例越高，抽样数据越大 D 抽样样本量越大，抽样数据越小

参考答案：C

79、下面关于DBSCAN聚类说法正确的是：（ ）

A. 当数据量增大时，内存消耗很小；

B. 当空间聚类的密度不均匀时，聚类结果不受影响

C. 聚类间距差相差很大时，聚类结果不受影响

D. 使用了一个全局性的表征密度的参数

参考答案：D

80、采用AIC准则找最优模型得到如下的结果：

The AIC of ARMA(0,0) is 13679.401951273543

The AIC of ARMA(0,1) is 13332.135879407648

The AIC of ARMA(0,2) is 13289.758498467238

The AIC of ARMA(1,0) is 13458.95993448334

The AIC of ARMA(1,1) is 13266.651183460379

The AIC of ARMA(1,2) is 13241.110101949496

The AIC of ARMA(2,0) is 13396.594961413079

The AIC of ARMA(2,1) is 13245.885887026114

The AIC of ARMA(2,2) is 13242.201821971108

请问哪个模型最优（ ）

A. ARMA(0,2)

B. ARMA(2,2)

C. ARMA(1,0)

D. ARMA(1,2)

参考答案：D

### 二、判断题

1、缺失值通常用的处理方式有填补或者删除。（ ）

参考答案：√

2、离散选择模型主要研究选择结果的概率与影响因素之间的关系。（ ）

参考答案：√

3、决策树是一种由结点和有向边组成的层次结构。（ ）

参考答案：√

4、快速聚类又叫K-Means聚类。（ ）

参考答案：√

5、平稳的时间序列可能在整体上有明显的上升的趋势。( )

参考答案：◊

6、数据经过标准化处理后，数据的波动性将会减小。（ ）

参考答案：◊

7、cp值越小的决策树越复杂。（ ）

参考答案：√

8、Pearson相关系数取值为0时，表现为变量无相关。（ ）

参考答案：◊

9、单位根检验用于检验一个时间序列的平稳性。（ ）

参考答案：√

10、分类模型评价时，通常采用平方损失。（ ）

参考答案：◊

11、主成分分析中可以得到的主成分的个数与原始变量的个数一样。（ ）

参考答案：√

12、回归模型评价时，通常采用0-1损失。（ ）

参考答案：◊

13、线性回归模型的拟合优度的判定系数越大，说明回归方程预测能力较好。（ ）

参考答案：◊

14、Bayes判别依据每个被判入某个类别的后验概率进行归类。（ ）

参考答案：√

15、决策树节点中最纯的节点的Gini指数最大。（ ）

参考答案：◊

16、要解决居民出行方式：公交、私家车、自行车、地铁，与哪些变量相关的问题可以采用多重选择模型。（ ）

参考答案：√

17、数据抽样后数据波动性可能减小。（ ）

参考答案：√

18、cp值越小的决策树越复杂。（ ）

参考答案：√

19、白噪声序列具有零均值与同方差性。（ ）

参考答案：◊

20、在Python中缺失值通常用NA来记号。（ ）

参考答案：◊

21、线性回归模型的拟合优度的判定系数越大，说明回归方程预测能力较好。（ ）

参考答案：◊

22、欧氏距离不受总体空间大小的影响，也不受计量单位的影响。( )

参考答案：◊

23、随机森林算法由很多的决策树组成。（ ）

参考答案：√

24、期望预测误差中的方差部分是由估计值自身的波动带来的。（ ）

参考答案：√

25、平稳的时间序列可能在整体上有明显的上升的趋势。( )

参考答案：◊

26、数据经过标准化处理后，数据的波动性将会减小。（ ）

参考答案：◊

27、cp值越小的决策树越复杂。（ ）

参考答案：√

28、偏相关系数总是比相关系数高。（ ）

参考答案：◊

29、单位根检验用于检验一个时间序列的平稳性。（ ）

参考答案：√

30、分类模型评价时，通常采用平方损失。（ ）

参考答案：◊

31、特征值越大的主成分越代表的原始数据的信息越多。（ ）

参考答案：√

32、线性回归模型的拟合优度的判定系数越大，说明回归方程预测能力较好。（ ）

参考答案：◊

33、近邻分类中，随着近邻个数的增加，模型的复杂度增加。（ ）

参考答案：◊

34、DBSCAN目的是找到密度可达对象的最大集合。（ ）

参考答案：◊

35、要解决居民出行方式：公交、私家车、自行车、地铁，与哪些变量相关的问题可以采用多重选择模型。（ ）

（ ）

参考答案：√

36、数据抽样后数据波动性可能减小。（ ）

参考答案：√

37、cp值越小的决策树越复杂。（ ）

参考答案：√

38、K-MEANS（K-均值）聚类的中心一般用每个类数据的均值来表示。（ ）

参考答案：√

39、有的非平稳时间序列需要用高阶差分才能将其平稳化。（ ）

参考答案：√

40、DBSCAN算法能发现任意形状的空间聚类。（ ）

参考答案：√

### 三、多选题

1、关于抽样后的数据的大小说法正确的是（ ）。

A 抽样比例越低，抽样数据越大 B 抽样样本量越小，抽样数据越大

C 抽样比例越高，抽样数据越大 D 抽样样本量越大，抽样数据越大

参考答案：CD

2、缺失数据的处理方法包括（ ）。

A 填补 B 删除 C 插补 D 提取

参考答案：ABC

3、下列哪些变量不是属性数据（ ）。

A 性别 B 季节 C 工作类型 D 车流量 E 降雨量 F 图片灰度值

参考答案：DEF

4、关于主成分分析中的特征值理解正确的是（ ）。

A 可以对协方差矩阵进行特征值分解得到

B 可以对相关矩阵进行特征值分解得到

C 特征值越大的主成分越代表的原始数据的信息越多

D 特征值越小的主成分越代表的原始数据的信息越多

参考答案：ABC

5、期望预测误差一般包括下面的（ ）三部分。

A 模型假设失误带来的误差 B 偏差

C 方差 D 训练误差

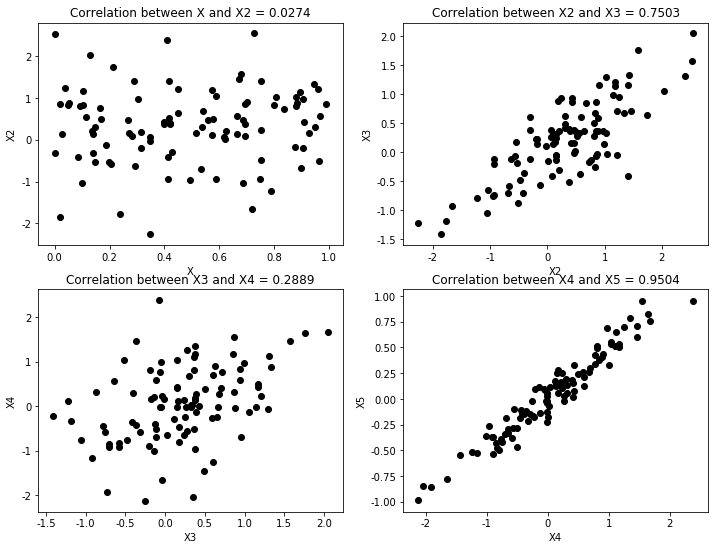
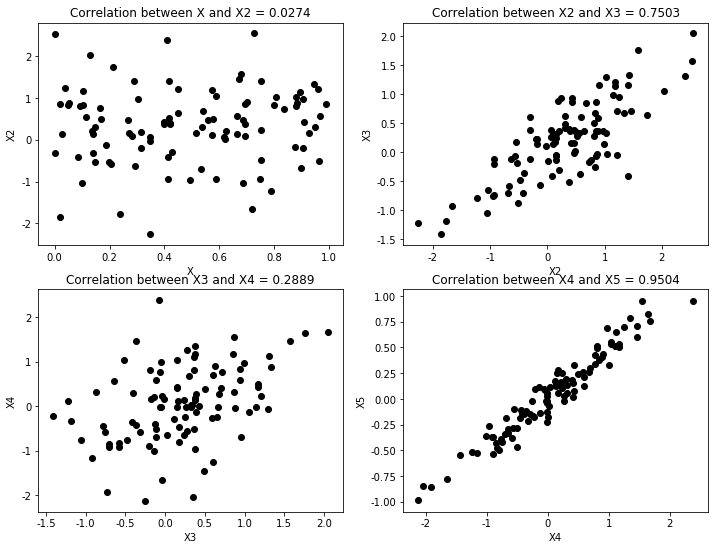
参考答案：ABC

6、分类任务的评价通常采用（ ）。

A 准确度 B ROC曲线 C 召回率 D F1 score E AUC

参考答案：ABCDE

7、从下面两个变量的相关系数图可以看出（ ）



A 左图相关系数大于0 B 右图相关系数小于0

C 左图相关系数大于右图 D 右图相关系数等于0

参考答案：AC

8、相关系数的显著性检验的原假设为（ ）。

A 两个变量之间没有线性关系

B 两个变量之间有线性关系

C 两个变量之间相关系数为0

D 两个变量之间相关系数不为0

参考答案：AC

9、从下面的相关系数矩阵中可以看出（ ）。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sepal.Length** | **Sepal.Width** | **Petal.Length** | **Petal.Width** |
| **Sepal.Length** | 1.000000 | -0.117570 | 0.871754 | 0.817941 |
| **Sepal.Width** | -0.117570 | 1.000000 | -0.428440 | -0.366126 |
| **Petal.Length** | 0.871754 | -0.428440 | 1.000000 | 0.962865 |
| **Petal.Width** | 0.817941 | -0.366126 | 0.962865 | 1.000000 |

A Sepal.Length与Sepal.Width之间呈现负线性相关，但是这种相关性不太高

B Petal.Length与Petal.Width之间有非常强的正线性相关

C Sepal.Width与Sepal.Length之间有较弱的负线性相关

D Sepal.Width与Petal.Width之间的负线性相关比Sepal.Length与Sepal.Width之间的负线性相关要低

参考答案：ABC

10、所谓拖尾，是指在自相关系数图或偏自相关系数图中（ ）

A. 自相关系数或偏自相关系数有指数型波动

B. 自相关系数或偏自相关系数有震荡型衰减的波动

C. 自相关系数或偏自相关系数不会都落入置信区间内

D. 自相关系数或偏自相关系数在某期之后的系数逐渐趋于零

参考答案：ABC

11、关于分位数回归，下面说法正确的是（ ）

A.当数据具有尖峰厚尾的分布特征或有离群点（即异常值）时，采用分位数回归更稳健。

B.分位数回归不需要进行残差检验

C.分位数回归会给出因变量不同分位数水平下的参数估计

D.分位数回归不需要做模型检验

参考答案：AC

12、关于决策树算法的复杂度一般采用哪些统计量来度量（ ）

A. 结点个数

B. 层数

C. cp值

D. 不纯度

参考答案：ABC

13、下面白噪声序列说法正确的是（ ）

A. 具有零均值

B. 具有同方差性

C. 协方差为零

D. 不是平稳时间序列

参考答案：BC

14、决策树中包含下面的这几类结点：（ ）

A. 根节点

B. 内部节点

C. 叶节点

D. 终节点

参考答案：ABCD

15、决策树中结点的最优划分可能是依据节点划分前后的（ ）来确定的。

A. 信息增益比最大

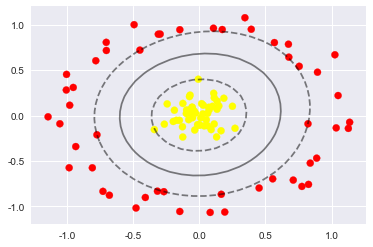
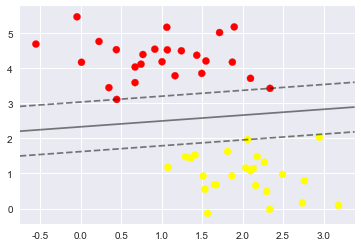
B. 信息增益最大

C. 信息增益最小

D. GINI增益最大

参考答案：ABD

16、由下图中两类数据（红色一类，黄色一类）的划分问题引出的关于线性可分与不可分的概念正确的是（ ）。



A. 左图两类数据的划分是一个线性可分问题

B. 右图的两类数据的划分是一个线性不可分问题

C. 左图两类数据的划分是一个线性不可分问题

D. 右图的两类数据的划分是一个线性可分问题

参考答案：AB

17、系统聚类中的类间距离是：（ ）

A.合并后新出现的类与其他类之间的距离

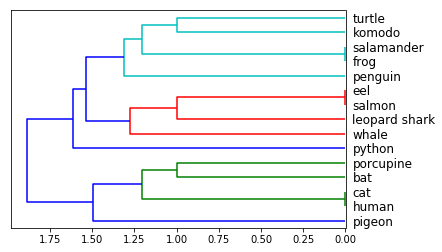
B.不能通过前一步的类间距离递推得到

C.可以通过前一步的类间距离递推得到

D.有多种可以选择的方式

参考答案：ACD

18、动物学家有的时候需要通过对新发现的物种进行归类。这里收集到了15种动物生理特征的数据。根据Warm-blooded、Gives Birth、Aquatic Creature、Aerial Creature、Has Legs、Hibernates等六个指标综合对样本进行聚类分析，考虑采用系统聚类法进行聚类。聚类后的谱系图如下所示。



请问如果最终聚为两类，关于聚类结果说法正确的是：

A porcupine、bat、cat、human、pigeon一类

B turtle、komodo、salamander、frog、penguin、eel、salmon、leopard shark、whale、python一类

C porcupine、bat、cat、human、pigeon python一类

D turtle、komodo、salamander、frog、penguin、eel、salmon、leopard shark、whale一类

参考答案：AB

19、设*dij*表示第*i*个样本与第*j*个样本之间的距离。那么距离必须满足的原则有：（ ）

A.

B.

C.

D.

参考答案：ABCD

20、聚类任务中，衡量个体之间的相似性通常使用下面哪些度量（ ）

A. 距离

B. 相关系数

C. 相似系数

D. 匹配系数

参考答案：ABCD

21、下列哪些变量不是属性数据（ ）。

A 性别 B 季节 C 工作类型 D 车流量 E 降雨量 F 图片灰度值

参考答案：DEF

22、数据经过标准化处理后（ ）。

A 数据的波动性可能增加 B 数据波动性可能减小 C 数据值都变小了 D 数据值将被放大 E 数据值主要集中在区间（-3,3） F 数据值主要集中在区间（,）

参考答案：ABE

23、下面关于主成分分析说法正确的是（ ）。

A 主成分是相互正交（垂直）的 B 主成分捕获了数据差异最大的方向

C 是线性降维方法 D 是非线性降维方法

参考答案：ABC

24、主成分分析用于（ ）。

A 减少分析变量 B 将数据从原始的高维空间投影到低维空间

C 将高维数据进行可视化 D 将数据进行标准化处理

参考答案：ABC

25、期望预测误差中的方差部分是由（ ）带来的。

A 最优解所在空间假设失误 B 数据的随机性

C 估计值自身的波动 D 估计值的平均可能偏离了真实值

参考答案：BC

26、分类任务的评价指标通常采用（ ）。

A 准确度 B ROC曲线 C 召回率 D F1 score E AUC

参考答案：ABCDE

27、典型相关分析适用的前提是（ ）

A 各组变量之间的关系为线性相关关系

B 两个变量之间的关系为线性相关关系

C 各组变量所组成的线性组合与构成该组合的具体变量之间的关系为线性相关

D 两个变量之间的关系为非线性相关

参考答案：AC

28、相关系数的显著性检验的原假设为（ ）。

A 两个变量之间没有线性关系

B 两个变量之间有线性关系

C 两个变量之间相关系数为0

D 两个变量之间相关系数不为0

参考答案：AC

29、从下面的相关系数矩阵中可以看出（ ）。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sepal.Length** | **Sepal.Width** | **Petal.Length** | **Petal.Width** |
| **Sepal.Length** | 1.000000 | -0.117570 | 0.871754 | 0.817941 |
| **Sepal.Width** | -0.117570 | 1.000000 | -0.428440 | -0.366126 |
| **Petal.Length** | 0.871754 | -0.428440 | 1.000000 | 0.962865 |
| **Petal.Width** | 0.817941 | -0.366126 | 0.962865 | 1.000000 |

A Sepal.Length与Sepal.Width之间呈现负线性相关，但是这种相关性不太高

B Petal.Length与Petal.Width之间有非常强的正线性相关

C Sepal.Width与Sepal.Length之间有较弱的负线性相关

D Sepal.Width与Petal.Width之间的负线性相关比Sepal.Length与Sepal.Width之间的负线性相关要低

参考答案：ABC

30、定性数据之间的相关通常采用（ ）来度量。

A Pearson相关系数

B Spearman相关系数

C Kendall tau-b相关系数

D Hoeffding’s D相关系数

参考答案：BCD

31、关于分位数回归，下面说法正确的是（ ）

A.当数据具有尖峰厚尾的分布特征或有离群点（即异常值）时，采用分位数回归更稳健。

B.分位数回归不需要进行残差检验

C.分位数回归会给出因变量不同分位数水平下的参数估计

D.分位数回归不需要做模型检验

参考答案：AC

32、关于决策树算法的复杂度一般采用哪些统计量来度量（ ）

A. 结点个数

B. 层数

C. cp值

D. 不纯度

参考答案：ABC

33、决策树算法中节点的不纯度通常采用哪些度量：（ ）

A.熵

B.cp值

C.GINI指数

D.误分率

参考答案：ACD

34、决策树中包含下面的这几类结点：（ ）

A. 根节点

B. 内部节点

C. 叶节点

D. 终节点

参考答案：ABCD

35、下面关于CART的理解正确的是（ ）。

A. CART能够探测和揭示数据中的交互作用。

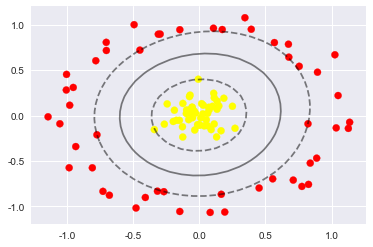
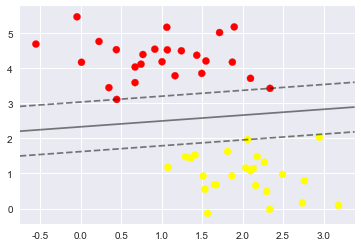
B. CART不受离群点、共线性、异方差性或者（如在经典回归中的）误差结构的影响。

C. CART对自变量和因变量不做任何形式的分布假定。

D. CART不能处理缺失数据

参考答案：ABC

36、由下图中两类数据（红色一类，黄色一类）的划分问题引出的关于线性可分与不可分的概念正确的是（ ）。



A. 左图两类数据的划分是一个线性可分问题

B. 右图的两类数据的划分是一个线性不可分问题

C. 左图两类数据的划分是一个线性不可分问题

D. 右图的两类数据的划分是一个线性可分问题

参考答案：AB

37、下面关于系统聚类说法正确的是：（ ）

A. 系统聚类也叫层次聚类

B. 系统聚类初始时，将每个个体看作一类

C. 系统聚类的过程就是将距离近的个体合并成一类

D. 系统聚类只用于Q型分类

参考答案：ABC

38、下面哪些距离是明氏距离：（ ）

A. 绝对距离

B. 欧氏距离

C. 车比雪夫距离

D. 马氏距离

参考答案：ABC

39、时间序列的各个元素之间的相关系数：（ ）

A.取值在-1到1之间

B.其绝对值越接近于1，说明时间序列的自相关程度越高

C.被称为自相关系数

D.时间间隔为的两部分数据之间的相关性被称为阶自相关系数

参考答案：ABCD

40、非平稳的时间序列：（ ）

A. 在整体上有明显的上升的趋势

B. 在局部上有明显的上升的趋势

C. 在整体上有明显的下降的趋势

D. 在局部上有明显的下降的趋势

参考答案：ABCD

41、一个时间序列可以由下面( )部分构成.(多选)

A. 长期趋势

B. 季节变动

C. 循环波动

D. 不规则变动

参考答案：ABCD

42、构成时序的各个组成元素的相关系数：（ ）(多选)

A.取值在-1到1之间

B.其绝对值越接近于1，说明时间序列的自相关程度越高

C.被称为自相关系数

D.时间间隔为的两部分数据之间的相关性被称为阶自相关系数

参考答案：ABCD

43、下列哪些变量不是属性数据（ ）。

A 性别 B 季节 C 工作类型 D 车流量 E 降雨量 F 图片灰度值

参考答案：DEF

44、关于主成分分析中的特征值理解正确的是（ ）。

A 可以对协方差矩阵进行特征值分解得到

B 可以对相关矩阵进行特征值分解得到

C 特征值越大的主成分越代表的原始数据的信息越多

D 特征值越小的主成分越代表的原始数据的信息越多

参考答案：ABC

45、期望预测误差一般包括下面的（ ）三部分。

A 模型假设失误带来的误差 B 偏差

C 方差 D 训练误差

参考答案：ABC

46、分类任务的评价指标通常采用（ ）。

A 准确度 B ROC曲线 C 特效度 D 灵敏度 E 错分成本

参考答案：ABCDE

47、典型相关分析适用的前提是（ ）

A 各组变量之间的关系为线性相关关系

B 两个变量之间的关系为线性相关关系

C 各组变量所组成的线性组合与构成该组合的具体变量之间的关系为线性相关

D 两个变量之间的关系为非线性相关

参考答案：AC

48、相关系数的显著性检验的原假设为（ ）。

A 两个变量之间没有线性关系

B 两个变量之间有线性关系

C 两个变量之间相关系数为0

D 两个变量之间相关系数不为0

参考答案：AC

49、从下面的相关系数矩阵中可以看出（ ）。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sepal.Length** | **Sepal.Width** | **Petal.Length** | **Petal.Width** |
| **Sepal.Length** | 1.000000 | -0.117570 | 0.871754 | 0.817941 |
| **Sepal.Width** | -0.117570 | 1.000000 | -0.428440 | -0.366126 |
| **Petal.Length** | 0.871754 | -0.428440 | 1.000000 | 0.962865 |
| **Petal.Width** | 0.817941 | -0.366126 | 0.962865 | 1.000000 |

A Sepal.Length与Sepal.Width之间呈现负线性相关，但是这种相关性不太高

B Petal.Length与Petal.Width之间有非常强的正线性相关

C Sepal.Width与Sepal.Length之间有较弱的负线性相关

D Sepal.Width与Petal.Width之间的负线性相关比Sepal.Length与Sepal.Width之间的负线性相关要低

参考答案：ABC

50、定性数据之间的相关通常采用（ ）来度量。

A Pearson相关系数

B Spearman相关系数

C Kendall tau-b相关系数

D Hoeffding’s D相关系数

参考答案：BCD

51、关于分位数回归，下面说法正确的是（ ）

A.当数据具有尖峰厚尾的分布特征或有离群点（即异常值）时，采用分位数回归更稳健。

B.分位数回归不需要进行残差检验

C.分位数回归会给出因变量不同分位数水平下的参数估计

D.分位数回归不需要做模型检验

参考答案：AC

52、关于决策树算法的复杂度一般采用哪些统计量来度量（ ）

A. 结点个数

B. 层数

C. cp值

D. 不纯度

参考答案：ABC

53、决策树算法中节点的不纯度通常采用哪些度量：（ ）

A.熵

B.cp值

C.GINI指数

D.误分率

参考答案：ACD

54、决策树中包含下面的这几类结点：（ ）

A. 根节点

B. 内部节点

C. 叶节点

D. 终节点

参考答案：ABCD

55、决策树中结点的最优划分可能是依据节点划分前后的（ ）来确定的。

A. 信息增益比最大

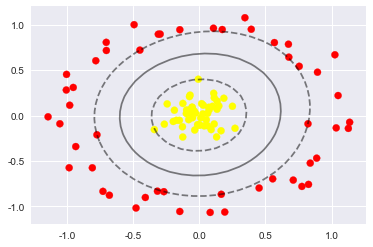
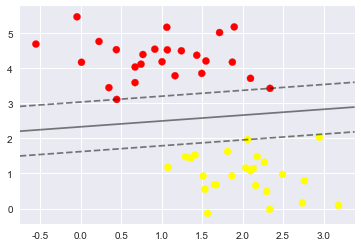
B. 信息增益最大

C. 信息增益最小

D. GINI增益最大

参考答案：ABD

56、由下图中两类数据（红色一类，黄色一类）的划分问题引出的关于线性可分与不可分的概念正确的是（ ）。



A. 左图两类数据的划分是一个线性可分问题

B. 右图的两类数据的划分是一个线性不可分问题

C. 左图两类数据的划分是一个线性不可分问题

D. 右图的两类数据的划分是一个线性可分问题

参考答案：AB

57、所谓截尾，是指在自相关系数图或偏自相关系数图中（ ）

A.自相关系数或偏自相关系数在滞后的前几期可能处于置信区间之外，

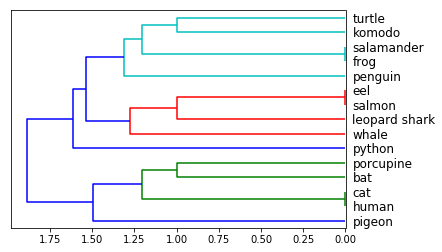
B.自相关系数或偏自相关系数在某期之后的系数基本上都落入置信区间内

C. 自相关系数或偏自相关系数在某期之后的系数逐渐趋于零

D. 自相关系数或偏自相关系数有指数型波动

参考答案：ABC

58、动物学家有的时候需要通过对新发现的物种进行归类。这里收集到了15种动物生理特征的数据。根据Warm-blooded、Gives Birth、Aquatic Creature、Aerial Creature、Has Legs、Hibernates等六个指标综合对样本进行聚类分析，考虑采用系统聚类法进行聚类。聚类后的谱系图如下所示。



请问如果最终聚为两类，关于聚类结果说法正确的是：

A porcupine、bat、cat、human、pigeon一类

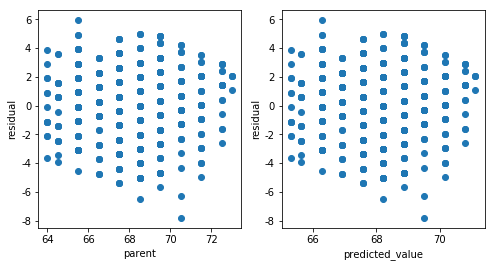
B turtle、komodo、salamander、frog、penguin、eel、salmon、leopard shark、whale、python一类

C porcupine、bat、cat、human、pigeon python一类

D turtle、komodo、salamander、frog、penguin、eel、salmon、leopard shark、whale一类

参考答案：AB

59、关于下面的残差图叙述正确的是（ ）。



A. 自变量与残差之间的关系不明显，基本上无关

B. 符合随机误差对于因变量而言具有同方差性的假定

C. 残差大体均匀的分布在[-8，6]之间，

D. 残差均值与0非常接近，故符合ε零均值的假定。

参考答案：ABCD

60、关于二元选择模型说法正确的是（ ）。

A. 二元选择模型一般选用极大似然法进行参数估计

B. 二元选择模型的参数估计值表示自变量变动对因变量的边际影响

C. 二元选择模型的参数估计值表示自变量的变动对因变量取“1”的概率的影响

D. 二元选择模型通常假设观测数据与一个隐藏的标准正态分布或者Logistic分布有关。

参考答案：ACD

61、主成分分析用于（ ）。

A 减少分析变量 B 将数据从原始的高维空间投影到低维空间

C 将高维数据进行可视化 D 将数据进行标准化处理

参考答案：ABC

62、回归预测任务的评价指标通常采用（ ）。

A 平均绝对误差 B 均方根误差 C 特效度 D 灵敏度 E 决定系数

参考答案：ABE

63、下列哪些变量不是属性数据（ ）。

A 性别 B 季节 C 工作类型 D 车流量 E 降雨量 F 图片灰度值

参考答案：DEF

64、关于主成分分析中的主成分说法正确的是（ ）。

A 可以得到的主成分的个数与原始变量的个数一样

B 通常只保留前几个特征值大的主成分

C 通常只保留前几个特征值小的主成分

D 主成分保留个数是随意指定的

参考答案：AB

65、期望预测误差中的偏差部分是由（ ）带来的。

A 估计值的平均可能偏离了真实值 B 数据的随机性

C 最优解所在空间假设失误 D 估计值自身的波动

参考答案：AB

66、聚类分析的内部评价指标通常有（ ）。

A DBI B 兰德指数 C Jaccard系数 D FM指数 E DI

参考答案：AE

67、典型相关分析适用的前提是（ ）

A 各组变量之间的关系为线性相关关系

B 两个变量之间的关系为线性相关关系

C 各组变量所组成的线性组合与构成该组合的具体变量之间的关系为线性相关

D 两个变量之间的关系为非线性相关

参考答案：AC

68、下面关于时间序列的平稳性判断的说法正确的是（ ）

A. 自相关系数检验可以用于时间序列的平稳性检验

B. 白噪声检验可以用于时间序列的平稳性检验

C. 若时间序列的自相关系数基本上（通常为>3时）都落入置信区间（即上图中的两条水平线之间），且逐渐趋于零，则该时间序列具有平稳性

D. 若时间序列的自相关系数更多地落在置信区间外面，则该时间序列具有平稳性。

参考答案：ABC

69、关于下面的购物篮事务数据叙述正确的是（ ）

|  |  |  |
| --- | --- | --- |
|  | **TID** | **项集** |
| **1** | 1 | {面包，牛奶} |
| **2** | 2 | {面包，尿布，啤酒，鸡蛋} |
| **3** | 3 | {牛奶，尿布，啤酒，可乐} |
| **4** | 4 | {面包，牛奶，尿布，啤酒} |
| **5** | 5 | {面包，牛奶，尿布，可乐} |

A 事务1中的{面包，牛奶}是一个2-项集

B 项集{啤酒，尿布，牛奶}的支持度计数为3

C 项集{啤酒}的支持度计数为3

D {面包，牛奶，尿布}是一个2-项集

参考答案：AC

70、定性数据之间的相关通常采用（ ）来度量。

A Pearson相关系数

B Spearman相关系数

C Kendall tau-b相关系数

D Hoeffding’s D相关系数

参考答案：BCD

71、公司针对不同性别、年龄的潜在客户研究与手机购买意向（是否购买）相关的变量得到如下的结果。

手机屏幕截图

描述已自动生成

关于这个结果说法正确的是（ ）。

A. 性别变量是显著的

B. 年龄变量也是显著的

C. 年龄越大，购买手机的可能性越小

D. 采用Logit连接函数

参考答案：ABD

72、关于下面的线性判别结果的说法正确的是（ ）

precision recall f1-score support

setosa 0.33 0.29 0.31 7

versicolor 0.25 0.50 0.33 4

virginica 0.00 0.00 0.00 4

micro avg 0.27 0.27 0.27 15

macro avg 0.19 0.26 0.21 15

weighted avg 0.22 0.27 0.23 15

A. 这是在15个观测上的预测结果

B. 这15个观测中setosa类别的观测有7个

C. 此次判别的预测准确率为22%

D. virginica类别的预测误差为0

参考答案：ABC

73、决策树算法中节点的不纯度通常采用哪些度量：（ ）

A.熵

B.cp值

C.GINI指数

D.误分率

参考答案：ACD

74、决策树中包含下面的这几类结点：（ ）

A. 根节点

B. 内部节点

C. 叶节点

D. 终节点

参考答案：ABCD

75、关于支持向量机算法的说法正确的是：（ ）

A. 超平面到所有两类数据点的距离之和称为分类器的边缘

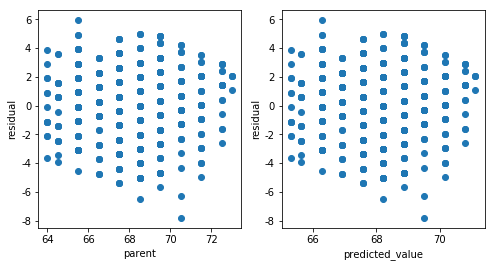
B. 具有最大边缘的超平面即是我们需要的最大边缘超平面。

C. 以超平面为以及相应的分类决策函数来分类的支持向量机称为线性可分支持向量机。

D. 非线性支持向量机通过引进核函数解决线性不可分问题。

参考答案：BCD

76、关于下面的残差图叙述正确的是（ ）。



A. 自变量与残差之间的关系不明显，基本上无关

B. 符合随机误差对于因变量而言具有同方差性的假定

C. 残差大体均匀的分布在[-8，6]之间，

D. 残差均值与0非常接近，故符合ε零均值的假定。

参考答案：ABCD

77、系统聚类中的类间距离是：（ ）

A.合并后新出现的类与其他类之间的距离

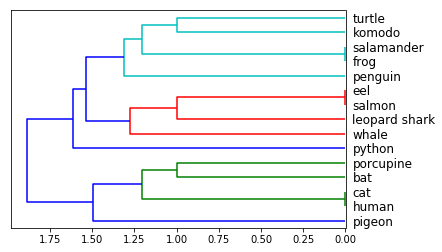
B.不能通过前一步的类间距离递推得到

C.可以通过前一步的类间距离递推得到

D.有多种可以选择的方式

参考答案：ACD

78、动物学家有的时候需要通过对新发现的物种进行归类。这里收集到了15种动物生理特征的数据。根据Warm-blooded、Gives Birth、Aquatic Creature、Aerial Creature、Has Legs、Hibernates等六个指标综合对样本进行聚类分析，考虑采用系统聚类法进行聚类。聚类后的谱系图如下所示。



请问如果最终聚为两类，关于聚类结果说法正确的是：

A porcupine、bat、cat、human、pigeon一类

B turtle、komodo、salamander、frog、penguin、eel、salmon、leopard shark、whale、python一类

C porcupine、bat、cat、human、pigeon python一类

D turtle、komodo、salamander、frog、penguin、eel、salmon、leopard shark、whale一类

参考答案：AB

79、非平稳的时间序列：（ ）

A. 在整体上有明显的上升的趋势

B. 在局部上有明显的上升的趋势

C. 在整体上有明显的下降的趋势

D. 在局部上有明显的下降的趋势

参考答案：ABCD

80、所谓拖尾，是指在自相关系数图或偏自相关系数图中（ ）

A. 自相关系数或偏自相关系数有指数型波动

B. 自相关系数或偏自相关系数有震荡型衰减的波动

C. 自相关系数或偏自相关系数不会都落入置信区间内

D. 自相关系数或偏自相关系数在某期之后的系数逐渐趋于零

参考答案：ABC