链滴

我们公司的统计与数据挖掘考试，考试时间是 1 个小时，满分 100 分

作者：relyn
原文链接：https://ld246.com/article/1473999212325
来源网站：链滴

<p>姓名：　　分数：
一、选择题（48 分）
1、以下两种描述分别对应哪两种对分类算法的评价标准？（　）
(a)警察抓小偷，描述警察抓的人中有多少个是小偷的标准。 (b)描述有多少比例的小偷给警察抓了的准。
A. Precision, Recall　B. Recall, Precision　C. Precision, ROC　D. Recall, ROC
2.当不知道数据所带标签时，可以使用哪种技术促使带同类标签的数据与带其他标签的数据相分离？（）
A. 分类 B. 聚类 C. 关联分析 D. 隐马尔可夫链
3.使用交互式的和可视化的技术，对数据进行探索属于数据挖掘的哪一类任务？（　）
A. 探索性数据分析 B. 建模描述 C. 预测建模 D. 寻找模式和规则
4.下面哪种不属于数据预处理的方法？（　）
A 变量代换　B 离散化　C 聚集　D 估计遗漏值
5.假设 12 个销售价格记录组已经排序如下：5, 10, 11, 13, 15,35, 50, 55, 72, 92, 204, 215，将它们分成四个箱，等频（等深）划分时，15 在第几个箱子内？（　）
A 第一个 B 第二个 C 第三个 D 第四个
6.以下哪种方法不属于特征选择的标准方法：（　）
A 嵌入 B 过滤 C 包装 D 抽样
7.下面不属于创建新属性的相关方法的是：（　）
A 特征提取 B 特征修改 C 映射数据到新的空间 D 特征构造
8.考虑值集{1、2、3、4、5、90}，其截断均值（p=20%）是（　）
A 2　B 3　C 3.5　D 5
9.假设属性 income 的最大最小值分别是 12000 元和 98000 元。利用最大最小规范化的方法将属性值映射到 0 至 1 的范围内。对属性 income 的 73600 元将被转化为：（　）
A 0.821　B 1.224　C 1.458　D 0.716
10.以下哪些算法是基于规则的分类器（　）
A. C4.5　B. KNN　C. Naive Bayes　D. ANN
11.决策树中不包含以下哪种结点？（　）
A,根结点（root node）B,内部结点（internal node）　C,外部结点（external node）　D,叶结点（leaf node）
12.以下哪项关于决策树的说法是错误的（　）
A.冗余属性不会对决策树的准确率造成不利的影响　B.子树可能在决策树中重复多次
C.决策树算法对于噪声的干扰非常敏感　　D.寻找最佳决策树是 NP 完全问题
13.因子分析的主要作用：（　）
A、对变量进行降维　B、对变量进行判别　C、对变量进行聚类　D、以上都不对
14.关于 K-means 聚类过程正确的是：（　）
A、使用的是迭代的方法 B、均适用于对变量和个案的聚类 C、对变量进行聚类 D、以上都不对
15.东北人养了一只鸡和一头猪。一天鸡问猪：＂主人呢？＂猪说：＂出去买蘑菇了。＂鸡听了撒丫就跑。猪说：＂你跑什么？＂鸡叫道：“有本事主人买粉条的时候你小子别跑！＂以上对话体现了数分析方法中的（　）
A. 关联　B. 聚类　C. 分类　D. 自然语言处理
16.已知甲班学生“统计学”的平均成绩为 86 分，标准差是 12.8 分，乙班学生“统计学”的平均成是 90 分，标准差是 10.3 分，下列表述正确的是（　）
A. 乙班平均成绩的代表性高于甲班 B. 甲班平均成绩的代表性高于乙班
C. 甲、乙两班平均成绩的代表性相同 D. 甲、乙两班平均成绩的代表性无法比较
17.当你用跑步时间（RunTime）、年龄（Age）、跑步时脉搏（Run_Pulse）以及最高脉搏（Maximm_Pulse）作为预测变量来对耗氧量（Oxygen_Consumption ）进行回归时，年龄（Age）的参数计是-2.78. 这意味着什么？（　）
A、年龄每增加一岁，耗氧量就增大 2.78　. B、年龄每增加一岁，耗氧量就降低 2.78.
C、年龄每增加 2.78 岁，耗氧量就翻倍。　　D、年龄每减少 2.78 岁，耗氧量就翻倍。
18.下面那一项可用于比较身高和体重的变异度（　）
A．方差　B．标准差　C．变异系数　D．全距
19.正态曲线下，横轴上从均数到 +∞ 的面积为（　）
A．97.5%　B．95%　　C．50%　D．5%

20.统计图中的散点图主要用来（ ）。
A.观察变量之间的相关关系　　B.主要用来表示总体各部分所占的比例
C.主要用来表示次数分布　　D.主要用来反映分类数据的频数分布
21、客户画像可以使用哪种分析方法？（)
A.聚类 B.因子分析 C.两者都可以 D.两者都不可以
22、个体之间的相似性主要用哪种数据挖掘方法？（)
A.聚类 B.因子分析 C.关联规则 D.社交网络分析
23 变量之间的相关性主要用哪种数据挖掘方法？（)
A.聚类 B.因子分析 C.关联规则 D.社交网络分析
24 客户之间的联系主要用哪种数据挖掘方法？（)
A.聚类 B.因子分析 C.关联规则 D.社交网络分析
二、填空题（22 分）
1、数据预处理包括_____、<strng>、</e>、和数据离散化。<r>
2、列举出处理空缺值的三种方法___、<stong>和。
3、四种计量尺度分别是定类、_____、<strng>和</strng>。
4. 标号 12345 的 5 个球，一次取两个，和为 3 或者 6 的概率是_____。
三、简答题 （30 分）
1、某银行信用卡模型建设过程中，申请评分卡模型训练过程出现过拟合现象，请阐述什么是过拟合象？如何解决过拟合现象？如果一个模型在训练过程出现欠拟合现象，那原因又是如何？如何避免？6 分）
2、什么是聚类分析? 请详细描述 k-means 算法的计算原理、步骤以及优缺点。（8 分）
3、已知每 10 万人中有 1 人得艾滋病。现在有一种检查，如果被测者患病则一定能查出来。如被测者没病，有 1% 的测试出错也显示阳性。现在一个人检查结果是阳性。问真正得病的概率?? （4 分）
4、何谓数据规范化？规范化的方法有哪些？写出对应的变换公式。（6 分）
5、何谓聚类？它与分类有什么异同？（6 分）
打赏区有答案