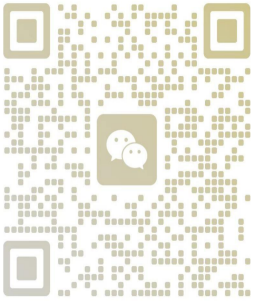
单项选择题



1.下列各项不属于数据的是（　）。
A.文本　　B.图像　　C.视频　　D.印象
2.下列各项不属于大数据特征的是（　）。
A.体量大　　B.种类多　　C.真实性　　D.数据生成慢
3.数据异常值的处理方法不包括（　）。
A.极小值替换　　B.删除　　C.忽略　　D.视为缺失值进行填补
4.下列各项不能用于描述数据集中趋势的是（　）。
A.方差　　B.平均数　　C.中位数　　D.峰值
5.下列各项不属于 Hadoop 的特点是（　）。
A.存储迅速　　B.成本高　　C.计算能力强　　D.灵活性强
6.在工业网络实时监控系统中，需要连续不断地采集和处理数据。以下（　　）不属于这种计算模式。
A.在线处理　　B. 实时处理　　C. 流式计算　　D. 批量计算
7. 下面不是研究数据方法的是（）。
A.统计学　　B.机器学习　　C.心理分析　　D.数据挖掘
8. 下面不属于大数据的处理过程的是（）。
A.数据获取　　B.数据清洗　　C.数据分析　　D.数据安全
9. 下面不属于大数据计算模式的类型的是（）。
A.批量计算　　B.手动计算　　C.流式计算　　D.交互式计算
10. 下列各项属于合规数据的是（　）。
A.非法收集隐私信息数据　　B.取得使用者同意的个人资料数据
C.泄露的隐私信息数据　　D.垄断数据
11. 在 Hadoop 生态系统中，主要负责节点集群的任务调度和资源分配，将存储和计算资源分配给不同应用程序的组件是（　）。
A.HDFS　　B.MapReduce　　C.YARN　　D.Storm
12. 下列属于图数据的主要特性的是（）。
A.数据驱动计算　　B.不规则问题　　C.高数据访问率　　D.以上均是
13. 可以用来查看数值型变量的分布的可视化方法是（）。
A.箱线图　　B.直方图　　C.小提琴图　　D.以上方法均可以
14. 如果只是研究两个数值变量之间的关系，最常见的可视化方法是（）。
A.直方图　　B.散点图　　C.饼图　　D.折线图
15. 下列各项不属于批处理系统的特点的是（　）。
A. 可以实现实时的分析报告或自动响应
B. 可以实现无缝扩展以处理峰值数据量或数据请求
C. 支持数据在不同系统之间进行交换
D. 支持作业执行状态的监控
16. 下列各项属于非结构化数据的是（　）。
A.图像　　B.二维数据表　　C.HTML 文档　　D.以上均是
17. 在大数据的处理流程中，（　）步骤是将数据转化为图形，以更直观的方式展示和表达。
A.存储与管理　　B.可视化　　C.采集与预处理　　D.分析与挖掘
18 . 下列关于异常值的描述中，错误的是（）。
A.　可以使用箱线图检测异常值
B.　当异常值的数量不是很多时，可以直接将含有异常值的观测记录删除
C.　可以将异常值视为缺失值，按处理缺失值的方法处理异常值
D.　异常值的存在不属于数据质量问题，不会影响模型的预测能力
19. 下列各项关于分类的描述中，错误的是（）。
A. 可以借助分类方法根据电子邮件的标题判断其是否为垃圾邮件

B. 在进行建模之前就要有明确的分组预测目标
C. k 近邻算法是一种简单但强大的分类算法
D. 用来建立分类模型的输入数据称为测试集
20. 假设散点图中的观测点分布较为分散，没有任何规律，说明两个变量之间的关系为（）。
A. 完全线性相关 B. 线性相关 C. 非线性相关 D. 不相关
21. 在 Hadoop 生态系统中，主要负责跨节点存储结构化或非结构化数据，并以日志文件的形式管理数据的组件是（ ）。
A.HDFS B.MapReduce C.YARN D.Storm
22. 下列各项不属于批处理系统的特点的是（ ）。
A. 为开发者提供了一个简单、快捷的开发框架
B. 支持各种数据格式的处理
C. 支持数据在不同系统之间进行交换
D. 可以实现实时的分析报告或自动响应
23. 为表示一组数据的分布特征，反映数据分布是否对称时，常用的可视化方法是（）。
A.箱线图 B.气泡图 C.折线图 D.散点图
24. 如果要反映某学生在 6 个学期中每学期平均成绩的变化情况，采用（ ）可视化方法较为合适。
A.饼图 B.折线图 C.散点图 D.直方图
25. 下列各项属于结构化数据的是（ ）。
A.图像 B.二维数据表 C.声音 D.文本
26. 在大数据的处理流程中，下列各项中最先进行的是（）。
A.存储与管理 B.可视化 C.采集与预处理 D.分析与挖掘
27 . 下列关于缺失值的描述中，错误的是（）。
A. 缺失值是指数据集中有些变量的一个或多个取值无法获得
B. 数据缺失的现象大量存在
C. 回归插补的方法不会改变数据分布
D. 当缺失数据的记录所占比例在数据集中少于 10%时，可以将缺失值直接删除
28. 下列各项关于聚类的描述中，错误的是（）。
A. 可以借助聚类方法进行异常检测
B. 在进行建模之前就要有明确的分组预测目标
C. 可以利用聚类分析发现具有相似功能的基因组
D. 根据数据本身的自然结构对数据进行分组
29. 假设散点图中的观测点恰好落在一条直线上，说明两个变量之间的关系为（）。
A. 完全线性相关 B. 线性相关 C. 非线性相关 D. 不相关
30. 下列各项属于数据仓库的特点的是（）。
A. 数据以主题为导向，提供决策支持
B. 数据源单一
C. 数据质量低
D. 不支持历史数据分析

判断题

1.根据数据在收集过程中是否控制有关因素，可以将数据分为观测数据和实验数据。（ ）
2. 时间序列分析中采用对数变换来消除异方差。（ ）
3. 关系型数据库不是用来存储和访问具有彼此相关性数据的数据库。（ ）
4. 气泡图中气泡的面积大小没有实际意义。（ ）
5. 数据科学是通过科学方法探索数据，以获得有价值的发现。()
6. 数据科学的发展不仅可以推动学科的发展，而且能够助推相关产业的发展与进步。()
7. 网页数据是一种半结构化数据。()
8. 在分类方法中，决策树法的结果复杂难懂、可解释性较差。（ ）

9. MapReduce 编程模型的首要步骤是对存储系统中的文件按列处理，并产生键值对。()
10. MapReduce 基于分而治之的算法范式，利用多台计算机完成数据处理（）
11. 银行业通过大数据技术可以有效分析经营过程中可能存在的风险因素。()
12. 数据脱敏技术可以有效降低敏感数据泄露的风险。()
13. 时间序列数据是按时间顺序排列的观测值序列，用于所描述现象随时间变化的情况。（　）
14. 数据预处理的主要目的是为了提高数据质量，将原始数据变成更加方便计算或处理的格式，使数据形态更加符合建模要求，进而提升数据挖掘的质量和效率。（　）
15.数据可视化对于提升数据的理解、分析与推断没有帮助。（　）

简答题

1.变量的定义是什么?
用于刻画观测数据集的特征的量叫做变量。
2.请列举三种常用的电子商务推荐算法。
协同过滤推荐算法；基于内容的推荐算法；基于关联规则的推荐算法；基于人口统计的推荐算法。
3.请列举五种常见的数据缺失值插补方法。
均值插补；回归插补；随机回归插补；多重插补；K 近邻算法插补
4.数据可视化的基本原则包括哪些方面?
数据可视化的效果要能准确的表达数据中的信息而不产生偏差或歧义；能够清晰地表达数据中的信息；其设计的可视化图表能够令人赏心悦目。
5.数据的定义是什么?
数据是对现象或事物进行测量和记录的结果，可用来制表、计算和分析等，也可以统指一切保存在电脑中的信息，能够进行电子化的记录，包括文本、图像、音频、视频等。
6.大数据的成因是什么?
数据的存储和管理能力的增强；数据采集能力增强；大数据的挖掘和分析等技术的同步发展。
7.数据整理的内容主要包括哪四个方面?
数据的提取；数据的连接；数据的聚合；去除冗余和重复。
8.通过相关系数矩阵处理共线性问题的算法步骤是什么?
计算相关系数矩阵；确定最大的成对相关系数对应的预测变量 A 和预测变量 B；计算变量 A 与其他所有变量之间的平均绝对值相关系数，对变量 B 也做同样的计算；比较 A 与 B，谁的平均绝对值相关系数最大，删除谁；重复以上步骤，直到两两之间绝对值相关系数低于某一特定阈值。
9.定量变量的定义是什么?
当一个变量的取值可以在一个范围内连续取值时，该变量就是定量变量。
10.大数据的处理流程主要包括哪 5 个步骤?
数据的采集与预处理；数据的存储与管理；数据的可视化；数据的分析与挖掘；大数据的处理。
11.数据离散化的定义是什么?
数据的离散化是指将数据由数值型变量变成分类型变量，即将变量的取值由原来的一个区间内连续取值映射到为若干个有限个值。
12.数据可视化的作用是什么?
快速获取信息；数据的探索性分析；挖掘数据隐藏的规律。