数据预处理技术 复习资料

一、判断题

1. Numpy 中的 T 属性可以用于数组的转置操作。（✓）
2. 使用 stack()方法时，必须指定 axis 参数的值。（✕）
3. 在 Numpy 中，itemsize 属性表示数组中每个元素的存储字节数。（✓）
4. swapaxes()方法用于交换 Numpy 数组的两个轴。（✓）
5. 使用 Numpy 的 zeros()函数可以创建一个所有元素都是 0 的 ndarray，且元素的数据类型默认为 float64。（✓）
6. pandas 库是专门为数据分析而设计的，不依赖于其他任何库。（✕）
7. Series 对象的数据类型可以是整数、字符串或浮点数等，但索引必须是整数类型。（✕）
8. 在创建 Series 对象时，如果不指定 index 参数，pandas 会自动生成从 0 开始的整数索引。（✓）
9. DataFrame 对象只能有一组索引，即行索引。（✕）
10. 使用 pd.DataFrame()函数创建 DataFrame 对象时，data 参数可以是 ndarray、dict、list 或可迭代对象。（✓）
11. groupby()方法中的 sort 参数默认为 True，表示对分组索引进行排序。（✓）
12. 堆叠合并数据中的'ignore_index=True'参数会忽略合并结果的索引，并生成新的索引。（✓）
13. 数据预处理阶段的任务包括数据清理、数据集成、数据变换，但不包括数据挖掘。（✓）
14. groupby()方法中的 axis 参数默认为 0，表示沿列方向进行分组操作。（✓）
15. pivot_table()方法是 pandas 中用于实现数据透视表功能的方法。（✓）
16. 堆叠合并数据只能沿着行方向进行，不能沿着列方向。（✕）
17. pandas 中的 merge()方法仅支持主键合并数据，不支持其他类型的合并。（✕）
18. 数据清理主要解决的数据问题包括数据缺失、数据重复和数据异常，但不包括数据冗余。（✓）
19. 在 pandas 中，drop()方法可以用于删除数据，包括根据指定的行标签索引或列标签索引删除异常值。（✓）
20. pandas 中 plot()函数用于绘制箱形图，且默认会显示网格线。（✕）

二、单选题

31. 在数据清理中，重复值主要有哪两种处理方式？

A: 删除和填充

B: 插补和替换

C: 删除和保留

D: 替换和忽略

32. 填充缺失值时，'pad'或'ffill'方法的含义是什么？

A: 使用缺失值后面的有效值填充

B: 使用缺失值前面的有效值填充

C: 使用众数填充

D: 使用随机数填充

33. 数据清理的主要目的是什么？

A: 增加数据量

B: 提高数据质量

C: 降低数据分析成本
D: 减少数据存储空间
34. 在使用 pd.read_sql()函数从数据库中读取数据时，sql 参数指的是什么？
A: 数据库连接字符串
B: 要执行的 SQL 查询语句
C: 数据库的名称
D: 表的名称
35. 使用 read_json()函数读取 JSON 文件时，orient 参数用于指定什么？
A: 文件的编码格式
B: 文件的路径
C: JSON 字符串的格式
D: 列索引
36. head()方法默认显示数据的前几行？
A: 1 行
B: 3 行
C: 5 行
D: 10 行
37. sort_values()方法中的 by 参数用于指定根据哪个索引名进行排序？
A: 行索引名
B: 列索引名
C: 自动生成的整数索引
D: 时间戳索引
38. 下列哪个选项不是创建 Series 对象时可以使用的数据类型？
A: ndarray
B: List
C: Dict
D: set
39. DataFrame 对象具有几组索引？
A: 1 组
B: 2 组
C: 3 组
D: 4 组
40. Numpy 中，unique()函数的作用是什么？
A: 对数组进行排序
B: 查找数组中的唯一元素
C: 生成随机数
D: 计算数组的平均值
41. 使用 Numpy 的哪个函数可以快速创建一个全零的 ndarray？
A: Zeros
B: Ones
C: Empty
D: identity
42. 在 Numpy 中，使用 stack 方法组合多个数组时，哪个参数用于指定新轴的位置？
A: Arrays

B: Axis
C: Shape
D: dtype
43. 在 Numpy 中，使用哪个函数可以实现数组的排序？
A: sort()
B: order()
C: arrange()
D: rank()
44. 下列哪个属性表示 ndarray 的维数？
A: Shape
B: Ndim
C: Size
D: dtype
45. 在堆叠合并数据中，如果设置 ignore_index=True，合并结果会发生什么变化？
A: 保留原索引
B: 生成新的索引
C: 删除所有索引
D: 不影响索引

三、填空题
51. unique 函数在 Numpy 中用于查找数组中的__唯一__元素。
52. describe 方法用于一次性描述数据的多个统计指标，如平均值、最大值、最小值等，而不用逐个调用__统计__函数。
53. head 方法默认显示数据的前__5__行。
54. DataFrame 对象有两组索引，分别是行索引和__列索引__。
55. pandas 中用于合并数据的函数或方法主要包括__主键__合并数据、重叠合并数据和堆叠合并数据等。
56. 数据集成是将多个数据源的数据合并到一个数据源，形成__一致__的数据存储的过程。
57. 箱型图是一种用于显示一组数据分散情况的统计图，它通常由上边缘、上四分位数、中位数、下四分位数、下边缘和__异常值__组成。
58. pandas 中提供了删除缺失值的方法__dropna__，它用于删除缺失值所在的一行或一列数据。
59. 处理异常值之前，需要先辨别哪些值是“真异常”和“伪异常”，再根据实际情况正确地处理__异常__值。
60. 重复值主要有两种处理方式：删除和保留，其中删除重复值是比较常见的方式，目的在于保留__唯一__的数据记录。