2024年大数据挖掘分析专业考试题库（含答案）

一、单选题

1.个人信息的收集、处理和利用应当遵循()的原则，不得违反法律、法规的规定和双方的约定收集、处理和利用个人信息。()

A、正规、合法、必要

B、合法、正当、必要

C、合法、合规、正当

D、合法、合理、合规

参考答案： B

2.Tableau 能够对数据进行处理包括()。

A、将多个数据源数据拼接为一个宽表；

B、修改、删除、新增数据行；

C、对数据进行行列转换、重命名、格式修改；

D、对数据进行计算、合并生成新的数据列

参考答案： A

3.()是数据仓库体系架构的重要组成部分，具备数据仓库的部分特征和 OLTP 系统的部分特征。

A、E.SB;

B、D.M

C、ODS

D、E.TL

参考答案：C

4.RFM 方法中的F 说明客户的()

A、兴趣度

B、粘性

C、当前价值

D、未来价值

参考答案：A

5. 关于主成分数目的选取，正确的是()

A、保留多少个主成分取决于累计方差在方差总和中所占百分比

B、一般选择50%以上

C、选择前两个就可以

D、选择的数目和变量的个数一致

参考答案：A

6. 下列关于数据重组的说法中，错误的是()

A、数据重组是数据的重新生产和重新采集

B、数据重组能够使数据焕发新的光芒

C、数据重组实现的关键在于多源数据融合和数据集成

D、数据重组有利于实现新颖的数据模式创新

参考答案： A

7. 在 SQL中，创建数据库用的命令是()

A、CREATESCHEMA

B、CREATETABLE

C、CREATEVIEW

D、CREATEDATABASE

参考答案： D

8. 当时间序列数据点的一阶差分近似为一常数，可配合以下哪种预测模型()

A、直线

B、二次抛物线

C、三次抛物线

D、指数曲线

参考答案： A

9. 对于企业来说，数据使用的关键是()

A、数据收集

B、数据存储

C、数据分析

D、数据再利用

参考答案： D

10. 线性回归算法寻找()与预测目标之间的线性关系。

A、属性

B、根因

C、表象

参考答案：A

11. 下面不属于明细数据质量评价指标的是()。

A、接入率;

B、自动采集率

C、及时率

D、完整率

参考答案： B

12. 以下哪些分类方法可以较好地避免样本的不平衡问题?()

A、KNN

B、SVM

C、Bayes

D、神经网络

参考答案： A

13. ()算法是最广泛使用的聚类算法，算法简单，易于理解和操作。

A、gglomerative

B、C.URE

C、K-means

D、k-中心点算法

参考答案： C

14贝叶斯决策是根据()进行决策的一种方法。

A、极大似然概率

B、先验概率

C、边际概率

D、后验概率

参考答案： D

15. 回归分析的第一步是()

A、确定解释量和被解释变量

B、确定回归模型

C、 建立回归方程

D、进行检验

参考答案： A

16. 当所有观测值都落在回归直线上，则这两个变量之间的相关系数为()

A、1

B、-1

C、+1 或-1

D、0

参考答案： C

17. ()是进行项目投资效益评价的最终依据。

A、现金流量

B、盈亏平衡点

C、净现金流量

D、现金流入量

参考答案： C

18. 被广泛用于购物篮分析的是()。

A、关联分析；

B、分类和预测

C、聚类分析

D、演变分析

参考答案： A

19. 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

参考答案： D

20. 下列不属于关联分析的关键要素的是()

A、支持度

B、置信度

C、满意度

D、提升度

参考答案： C

21. NoSQL 含义是指()

A、NO!SQL;

B、NomberSQL;

C、NotOnlySQL

D、NOLLSQL

参考答案： C

22. 资金的时间价值是()

A、同一资金在同一时点上价值量的差额

B、同一资金在不同时点上价值量的差额

C、不同资金在同一时点上价值量的差额

D、不同资金在不同时点上价值量的差额

参考答案： B

23. 以下哪种方法不属于于监督学习模型()

A、决策树

B、线性回归

C、关联分析

D、判别分析

参考答案： C

24. 在多元回归模型的检验中，目的是检验每一个自变量与因变量在指定显著性水平上是否存在线性相关关系的检验是()

A、r检验

B、t检验

C、f检验

D、DW 检 验

参考答案： B

25. 关于混合模型聚类算法的优缺点，下面说法正确的是()

A、当簇只包含少量数据点，或者数据点近似协线性时，混合模型也能

很好地处理。

B、混合模型比K 均值或模糊C 均值更一般，因为它可以使用各种类型的分布。

C、混合模型很难发现不同大小和椭球形状的簇。

D、混合模型在有噪声和离群点时不会存在问题。

参考答案： B

26. 大数据背景下，数据支撑业务的目的是()

A、建立数据科学

B、完成数据应用

C、配备数据硬件

D、吸纳数据人才

参考答案： B

27. 下面关于因子分析的说法正确的是()

A、因子分析就是主成分分析

B、因子之间可相关也可不相关

C、因子受量纲的影响

D、可以对因子进行旋转，使其意义更明显

参考答案： D

28. 快速实现简单的 MapReduce 统计，不必开发专门的 MapReduce 应用，十分适合数据仓库的统计分析的是()。

A、Map;

B、Reduce

C、H.ive

D、SQL 语 句

参考答案： D

29. 企业要建立预测模型，需准备建模数据集，以下四条描述建模数据集正确的是()

A、数据越多越好

B、尽可能多的适合的数据

C、数据越少越好

D、以上三条都不正确

参考答案： B

30. 以下哪个类型的变量在作预测客户流失的模型中最有解释力度?

A、人口基本数据，比如年龄和性别

B、基本社会状态数据，比如收入和职业

C、业务数据，比如消费频次

D、业务数据的衍生变量，比如最近3个月消费频次的变化情况

参考答案： D

31. 将复杂的地址简化成北、中、南、东四区，是在进行?

A、数据正规化

B、数据一般化

C、数据离散化

D、数据整合

参考答案： B

32.Hadoop 是一个开发和运行处理大规模数据的软件平台，是 Appach 的一个用()语言实现开源软件框架。

A、java

B、

C.++

C、R 语言

参考答案： A

33. 大数据特征错误的是()。

A、容量大;

B、类型多

C、价值高

D、系统多

参考答案： D

34.Apriori 算法是最基本的一种关联规则算法，它采用布尔关联规则的挖掘频繁项集的算法，利用()搜索的方法挖掘频繁项集。

A、逐层

B、逐级

C、自底向上

D、自上而下

参考答案： A

35. 分类算法以()定理为基础，采用概率方法对数据进行建模

A、决策树

B、K-最邻近

C、SVM

D、 贝叶斯

参考答案： D

36. 自然界中某种事物发生时其他事物也会发生，则这种联系称之为()。

A、连接

B、联络

C、关联

D、联系

参考答案： C

37. 源业务系统接入数据中心的方式主要有：JDBCESB 和()。

A、D.XP;

B、E.SP

C、0GG

D、E.TL

参考答案： C

38. 下列哪个不属于个人信息影响评估原则()

A、个人信息的处理目的、处理方式等是否合法、正当、必要

B、对个人的影响及风险程度

C、谁主管谁负责

D、所采取的安全保护措施是否合法、有效并与风险程度相适应。

参考答案： C

39. 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

参考答案： C

40.Hadoop 框架中两大核心是：()和 MapReducE

A、H.CFS;

B、H.DFS

C、H.EFS

D、H.FFS

参考答案： B

41. 将数据转换为可视化的形式，便于直观快速发现数据规律。的数据探索方法是()。

A、汇总统计法

B、概率统计法

C、可视化法

参考答案： C

42. 矩估计的基本原理是()

A、用样本矩估计总体矩

B、使得似然函数达到最大

C、使得似然函数达到最小

D、小概率事件在一次试验中是不可能发生的

参考答案： A

43.数据预处理目前常用的异常数据识别方法包括业务判别法、()、箱线图判别法、统计判别法

A、 聚类判别法；

B、 回归判别法

C、抽样判别法

参考答案： A

44.算法的核心思想是()逐层构造一个树。

A、 自上而下

B、自下而上

C、 自左向右

D、 自右向左

参考答案： A

45.下列关于大数据的分析理念的说法中，错误的是()

A、在数据基础上倾向于全体数据而不是抽样数据

B、在分析方法上更注重相关分析而不是因果分析

C、在分析效果上更追究效率而不是绝对精确

D、在数据规模上强调相对数据而不是绝对数据

参考答案： D

46.什么是 KDD?()

A、数据挖掘与知识发现

B、领域知识发现

C、文档知识发现

D、动态知识发现

参考答案： A

47.某家长为了使孩子在第3-6 年上大学的4 年中，每年年初得到10000元助学基金，他应在2年前在银行存入多少钱?(年利率按5%计算)()

A、33771

B、30291

C、32163

D、45256

参考答案： A

48.关于K 均值和DBSCAN 的比较，以下说法不正确的是()。

A、K均值丢弃被它识别为噪声的对象，而DBSCAN一般聚类所有对象

B、K 均值使用簇的基于原型的概念，而DBSCAN 使用基于密度的概念

C、K 均值很难处理非球形的簇和不同大小的簇，DBSCAN 可以处理不同大小和不同形状的簇

D、K 均值可以发现不是明显分离的簇，即便簇有重叠也可以发现，但是DBSCAN 会合并有重叠的簇

参考答案： A

49.SQL 查询语句中HAVING 子句的作用是()

A、指出分组查询的范围

B、指出分组查询的值

C、指出分组查询的条件

D、指出分组查询的内容

参考答案： C

50.一组数据中出现次数最多的数据称为()。

A、分位数

B、中位数

C、众数

参考答案： C

51.JSON中的中括号一般来表示()。

A、数组;

B、标点符号

C、对象

D、注释

参考答案： C

52.模型构建指基于()数据构建数据挖掘模型。

A、线上

B、线下

C、实时

D、历史

参考答案： D

53.Tableau 在处理离线地图时，需要将标记设置为()。

A、路径;

B、区域

C、边形

D、 已填充地图

参考答案： A

54.以下关于人工神经网络(ANN)的描述错误的有()

A、神经网络对训练数据中的噪声非常鲁棒

B、可以处理冗余特征

C、训练ANN 是一个很耗时的过程

D、至少含有一个隐藏层的多层神经网络

参考答案： A

55.美国海军军官莫里通过对前人航海日志的分析，绘制了新的航海路线图，标明了大风与洋流可能发生的地点。这体现了大数据分析理念中的()

A、在数据基础上倾向于全体数据而不是抽样数据

B、在分析方法上更注重相关分析而不是因果分析

C、在分析效果上更追究效率而不是绝对精确

D、在数据规模上强调相对数据而不是绝对数据

参考答案： B

56.当时间序列的环比增长速度大体相同时，适宜拟合()

A、指数曲线

B、抛物线

C、直线

D、对数曲线

参考答案： A

57. 将多个指标转化为少数几个指标的一种统计分析方法是()。

A、数据预处理;

B、数据降维

C、主成分分析

D、假设检验

参考答案： C

58. 设 X={1,2,3} 是频繁项集，则可由X可产生()个关联规则。

A、3

B、4

C、5

D、6

参考答案： D

59. 当一个连续变量的缺失值占比在85%左右时，以下哪种方式最合理()

A、直接使用该变量

B、根据是否缺失，生成指示变量，仅使用指示变量作为解释变量

C、使用多重插补的方法进行缺失值填补

D、直接删除该变量

参考答案： B

60. 大数据分析挖掘流程正确的是()。

A、业务理解→数据理解→数据准备→建立模型→模型评估；

B、业务理解→数据准备→数据理解→建立模型→模型评估；

C、业务理解→数据准备→数据理解→模型评估→建立模型；

D、业务理解→数据准备→模型评估→数据理解→建立模型

参考答案： A

61. ()是统计学的基础，是统计学里面最重要的概率分布

A、正态分布；

B、静态分布

C、动态分布

D、稳态分布

参考答案： A

62. 因子分析的主要作用有()

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

参考答案： A

63. 数据中心侧的数据流转方式未为()

A、D.XP;

B、E.SP

C、0GG

D、E.TL

参考答案： D

64.给定历史时间数据，通过拟合时序模型，分析研究时序数据的发展变化规律，得出观测数据的历史统计特征，再据此进行外推预测目标的分析方法是()。

A、 聚类;

B、回归

C、时间序列

D、汇总统计

参考答案： C

65.智能健康手环的应用开发，体现了()的数据采集技术的应。

A、统计报表

B、网络爬虫

C、API 接口

D、传感器

参考答案： D

66.假设检验中显著性水平是()

A、推断时犯取伪错误的概率

B、 推断时取伪弃真的概率

C、正确推断的概率

D、是推断的可信度

参考答案： B

67. 以下哪些算法是分类算法()

A、DBSCAN

B、C4.5

C、K-Mean

D、EM

参考答案： B

68. 以下关于大数据应用说法错误的是()。

A、大数据起源互联网，目前处于成熟期；

B、目前金融、电信、零售、公共服务等领域在积极的探索和应用大数据；

C、互联网是大数据的发源地；

D、互联网上形成了多种相对成熟的应用模式。

参考答案： A

69. 下列关于计算机存储容量单位的说法中，错误的是()

A、1KB<1MB<1GB

B、基本单位是字节(Byte)

C、一个汉字需要一个字节的存储空间

D、一个字节能够容纳一个英文字符

参考答案： C

70. 当置信水平一定时，置信区间的宽度()

A、随着样本量的增大而减小

B、随着样本量的增大而增大

C、与样本量的大小无关

D、先随着样本量的增大而减小，到一定程度后会随着样本量的增大而增大。

参考答案： A

71. 倒传递神经网络(BP 神经网络)的训练顺序为何?(A:调整权重；B:计算误差值；C:利用随机的权重产生输出的结果)

A、BCA

B、CAB

C、BAC

D、CBA

参考答案： D

72. 个人信息保护影响评估报告和处理情况记录应当至少保存()年。

A、一

B、十

C、五

D、三

参考答案： D

73. 资金的时间价值是()

A、同一资金在同一时点上价值量的差额

B、同一资金在不同时点上价值量的差额

C、不同资金在同一时点上价值量的差额

D、不同资金在不同时点上价值量的差额

参考答案： B

74. 有一条关联规则为A→B,此规则的信心水平(confidence) 为60%,则代表()

A、买 B 商品的顾客中，有60%的顾客会同时购买A

B、同时购买A,B 两商品的顾客，占所有顾客的60%

C、买 A 商品的顾客中，有60%的顾客会同时购买B

D、两商品 A,B在交易数据库中同时被购买的机率为60%

参考答案： C

75. 有一组数据其均值是20, 对其中的每一个数据都加上10, 那么得到的这组新数据的均值是()。

A、20

B、10

C、15

D、30

参考答案： D

76. 与大数据密切相关的技术是()。

A、蓝牙;

B、云计算

C、Wi-Fi

D、博弈论

参考答案： B

77. 在数据分析和处理方面具有分析方法丰富、分析模型扩展强、数据挖掘能力强等特点的分析工具是()。

A、Weka

B、SPSS

C、SAS

D、R

参考答案： D

78. 用于分类与回归应用的主要算法有：()

A、Apriori 算法、HotSpot 算法

B、RBF 神经网络、K均值法、决策树

C、K 均值法、SOM 神经网络

D、决 策 树 、BP 神经网络、贝叶斯

参考答案： D

79. ()提供的支撑技术，有效解决了大数据分析、研发的问题，比如虚拟化技术、并行计算、海量存储和海量管理等。

A、点计算

B、线计算

C、 云计算

D、面计算

参考答案： C

80. 描述一组对称(或正态)分布数据的离散程度时，最适宜选择的指标是()

A、极差

B、标准差

C、均值

D、变异系数

参考答案： B

81. 考虑下面的频繁 3- 项集的集合：{1, 2, 3}, {1, 2, 4}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5} 假定数据集中只有5个项，若采用合并策略，则由候选产生过程得到4-项集不包含()

A、1,2,3,4

B、1,2,3,5

C、1,2,4,5

D、1,3,4,5

参考答案： C

82. 相关分析与回归分析的一个重要区别是()

A、前者研究变量之间关系的密切程度，后者研究变量间的变动关系，并用方程式表示

B、前者研究变量之间的变动关系，后者研究变量间关系的密切程度

C、两者都研究变量间的变动关系

D、两者都不研究变量间的变动关系

参考答案： A

83.SQL 语句中删除表的命令是()

A、DROPTABLE

B、DELETETABLE

C、ERASETABLE

D、DELETEDBF

参考答案： A

84. 已知事件A的概率P(A)=0.6,U 为必然事件，则P(A+U)=1,P(AU)=

A、0.4

B、0.6

C、0

D、1

参考答案： B

85. 迈克尔·波特的“五力模型”中，五种竞争力量中不包括()

A、新进入者

B、供应商议价能力

C、其他利益相关者力量

D、行业中现有竞争者

参考答案： C

86. 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 使用等宽划分(宽度为50)方法将它们划分成四个箱，求15在哪个箱子?()

A、第 1 个

B、第 2 个

C、第 3 个

D、第 4 个

参考答案： A

87. ()是研究一种或者多种因素的变化对试验结果的观测值是否有显著影响的统计方法。

A、因子分析；

B、数据降维

C、方差分析

D、假设检验

参考答案： C

88. 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

89. 以下哪个指标不能用于线性回归中的模型比较()

A、R 方

B、调整R方

C、AIC

D、BIC

参考答案： A

90. 下面关于聚类分析说法错误的是()

A、一定存在一个最优的分类

B、聚类分析是无监督学习

C、聚类分析可以用于判断异常值

D、聚类分析即：物以类聚，人以群分

参考答案： A

91. 分类变量使用以下哪个统计量进行缺失值填补较合适

A、均值

B、最大值

C、众数

D、中位数

参考答案： C

92. 若数据量较大，下面哪种方式比较适合()

A、系统聚类

B、快速聚类(k-means)

C、A 和 B 都可以

D、A 和 B都不可以

参考答案： B

93. 在对历史数据集进行分区之前进行数据清洗(缺失值填补等)的缺点是什么

A、违反了建模的假设条件

B、加大了处理的难度

C、无法针对分区后各个数据集的特征分别做数据清洗

D、无法对不同数据清理的方法进行比较，以选择最优方法

参考答案： D

94. 数据仓库是随着时间变化的，下面的描述不正确的是()

A、数据仓库随时间的变化不断增加新的数据内容

B、捕捉到的新数据会覆盖原来的快照

C、数据仓库随时间变化不断删去旧的数据内容

D、数据仓库中包含大量的综合数据，这些综合数据会随着时间的变化不断地进行重新综合

参考答案： C

95.SQL 语言中，删除一个表中所有数据，但保留表结构的命令是()

A、DELETE

B、DROP

C、CLEAR

D、REMORE

参考答案： A

96. 依照《中华人民共和国数据安全法》和有关法律、行政法规的规定，()负责统筹协调网络数据安全和相关监管工作。

A、工业和信息化部

B、 国家安全部门

C、国家网信部门

D、通信主管部门

参考答案：C

97.假设属性 ine 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的值映射到0至1的范围内。对属性 ine 的73600元将被转化为()

A、0.751

B、0.163

C、0.457

D、0.716

参考答案： D

98.建立一个模型，通过这个模型根据已知的变量值来预测其他某个变量值属于数据挖掘的哪一类任务?()

A、根据内容检索

B、建模描述

C、预测建模

D、寻找模式和规则

参考答案： C

99.自动化高级分析实验室，实现与统一数据资源库互联，实现数据的自助组表、自助分析功能，满足不同层级、不同水平的用户需求的是()

A、初级分析；

B、综合分析

C、典型分析

D、高级分析

参考答案： D

100. 下列选项中属于现金流入的项目是()

A、所得税

B、建设投资

C、经营成本

D、营业收入

参考答案： D

101. 一 组N 个观测值按数值大小排列，分成100份，处于 X%位置的值称第X 个百分位数称为()。

A、分位数

B、中位数

C、众数

参考答案： A

102. 开始将N 个样品各自作为一类，将规定样品之间的距离和类与类之间的距离，然后将距离最近的两类合并成一个新类，计算新类与其他类的距离，重复进行两个最近类的合并，每次减少一类，直至所有的样品合并为一类，此种聚类方法是()

A、K-means

B、SOM 聚 类

C、系统聚类

D、有序聚类

参考答案：C

103. 指数平滑法可以用以下哪种指标来反映对时间序列资料的修正程度()

A、平滑常数

B、指数平滑数初始值

C、跨越期

D、季节指数

参考答案： A

104. 变量的量纲比如以厘米或者米为单位对下面哪种方法会有影响()

A、方差分析

B、 回归分析

C、聚类分析

D、主成分分析

参考答案： C

105. 关于Tableau 的特点，以下说法错误的是()

A、学习成本低，简单易用；

B、图表精美；

C、开发快速，分享便捷；

D、需要IT 大量人员参与

参考答案： D

106. 在 ID3 算法中信息增益是指()

A、信息的溢出程度

B、信息的增加效益

C、熵增加的程度最大

D、熵减少的程度最大

参考答案： D

107. 一组数据中最大值与最小值的差值称为(),也称全距。

A、极差

B、极距

C、距离

D、方差

参考答案： A

108. ()是一种开源软件编程语言，主要用于统计分析，绘图和数据挖掘，内置多种统计及分析功能。

A、H.IVE;

B、H.ADOOP

C、R;

D、H.FDS

参考答案： C

109. 决策树中的 InformationGain 的计算是用来?

A、剪枝

B、使树成长

C、处理缺失值和异常值

D、砍树

参考答案：B

110.泊松回归是一种广泛应用的()回归模型。

A、线性

B、 非线性

C、预测

D、估算

参考答案： B

111.数据库系统是由()组成的

A、数据库、数据库管理系统和用户

B、数据文件、命令文件和报表

C、数据库文件结构和数据

D、常量、变量和函数

参考答案：A

112.下列四项中，不属于数据库特点的是()

A、数据共享

B、数据完整性

C、数据冗余很高

D、数据独立性高

参考答案： C

113.某小区60%居民订晚报，45%订青年报，30%两报均订，随机抽一户。则至少订一种报的概率为()

A、0.82

B、0.85

C、0.80

D、0.75

参考答案： D

114.关于统计学和大数据之间的关系， 一下说法错误的是()。

A、面临大数据，统计学的研究对象有所改变；

B、在大数据环境中，需要首先将未知的问题转化为可用的统计方法；

C、在大数据分析过程中，传统的统计分析过程“定量一定位一再定性”转变为“定量一定性”；

D、在大数据环境中，需要将统计研究的对象范围扩展到一切数据。

参考答案： A

115.在数据分析和处理方面具有分析方法丰富、分析模型扩展差、数据挖掘能力强等特点的分析工具是()。

A、Weka

B、SPSS

C、SAS

D、R

参考答案： B

116.下列说明错误的是()

A、性别=“男”=>职业=“司机”,是布尔型关联规则

B、性别=“女”=>avg(收入)=2300,是一个数值型关联规则

C、肝炎=>ALT(丙氨酸转氨酶)升高，是一个单层关联规则

D、性别=“女”=>职业=“秘书”,是多维关联规则

参考答案： C

117.Hive 是基于Hadoop的一个数据()工具

A、分析；

B、仓库

C、制图

D、可视化

参考答案： B

118.数据记录内容完整比例，包括指标单位维度、业务维度组合记录条数完整，指标字段值完整称为()

A、指标数据自动采集率；

B、指标数据接入率

C、指标数据接入及时率；

D、指标数据完整率

参考答案： D

119.EXCEL中，求标准差的函数是()

A、AVERAGE

B、MEDIAN

C、MODE

D、STDEV

参考答案： D

120.对于下列实验数据：1,108,11,8,5,6,8,8,7,11,描述其集中趋势用()最为适宜，其值是()。

A、平均数，14.4

B、中位数，8.5

C、众数，8

D、 以上都可以

参考答案： C

121. 《个人信息保护法》发布执行时间()。

A、2021 年9月1日；

B、2021年10月1日；

C、2021 年11月1日

D、2021 年12月1日

参考答案： C

122.检测一元正态分布中的离群点，属于异常检测中的基于 ()的离群点检测

A、统计方法

B、邻近度

C、密度

D、聚类技术

参考答案： A

123.下列关于聚类挖掘技术的说法中，错误的是()

A、不预先设定数据归类类目，完全根据数据本身性质将数据聚合成不

同类别

B、要求同类数据的内容相似度尽可能

C、要求不同类数据的内容相似度尽可能

D、与分类挖掘技术相似的是，都是要对数据进行分类处理

参考答案： B

124.按数据的结构程度来划分，分为()

A、结构化数据、半结构化数据、非结构化数据

B、强结构化数据、弱结构化数据

C、截面数据、面板数据

D、一级数据、二级数据、三级数据

参考答案： A

125.当前国内社会中，最为突出的大数据环境是()

A、互联网

B、物联网

C、综合国力

D、自然资源

参考答案： A

126.()是一种松散耦合的服务和应用之间标准的集成方式。

A、E.SB;

B、D.M

C、ODS

D、E.TL

参考答案： A

127. 在建立线性回归(LinearRegression) 之前我们可以利用何种方法挑选重要属性，以降低模型的复杂度?

A、皮尔森相关系数

B、卡方检定

C、T-检定

D、Z-Score

参考答案： A

128. 给定一个置信概率和置信区域，若误差超过置信区域，则认为误差不是随机误差引起，视为异常值的判别方法()。

A、聚类判别法;

B、回归判别法

C、抽样判别法

D、统计判别法

参考答案： D

129. ROC 曲线凸向哪个角，代表模型越理想?

A、左上角

B、右上角

C、左下角

D、右下角

参考答案： A

130. . 给出下列结论：

1)在回归分析中，可用指数系数 R 方的值判断模型的拟合效果， R 方越大，模型的拟合效果越好；

2)在回归分析中，可用残差平方和判断模型的拟合效果，残差平方和越大，模型的拟合效果越好；

(3)在回归分析中，可用相关系数 r 的值判断模型的拟合效果，r 越小，模型的拟合效果越好；

(4)在回归分析中，可用残差图判断模型的拟合效果，残差点比较均匀地落在水平的带状区域中，说明这样的模型比较

合适. 带状区域的宽度越宽，说明模型的拟合精度越高. 以上结论中，正确的有()个.

A、1

B、2

C、3

D、4

参考答案： A

131. 回归是一种预测建模技术，研究()和()的依存关系。

A、预测值、实际值

B、自变量、因变量

C、绝对误差、平方误差

D、测试样本、集合

参考答案： B

132. ODS是指()

A、企业数据中心；

B、数据仓库

C、操作型存储

D、总线

参考答案： B

133. 用简单随机重复抽样方法抽取样本单位，如果要使抽样平均误差降低50%, 则样本容量需要扩大到原来的()

A、2倍

B、3倍

C、4倍

D、5倍

参考答案： C

134. 假设{BCE}为一频繁项目集(FrequentItemset), 则根据AprioriPrinciple 以下何者不是子频繁项目？

A、BC

B、CE

C、C

D、CD

参考答案： D

135.假如学生考试成绩以“优”.“良”.“及格”和“不及格”来记录，为了说明全班同学考试成绩的水平高低，其集中趋势的测度()

A、可以采用算术平均数

B、可以采用众数或中位数

C、只能采用众数

D、只能采用四分位数

参考答案： C

136.将原始数据进行集成、变换、维度规约、数值规约是在以下哪个步骤的任务?()

A、数据获取

B、分类和预测

C、数据预处理

D、数据可视化

参考答案： C

137.在有指导的数据挖掘中，有关测试集的说法错误的是()

A、测试集和训练集是相互联系的

B、测试集是用以测试模型的数据集

C、通常测试集大约占总样本的三分之一

D、K-次交叉验证中，测试集只有1个，训练集有K-1 个

参考答案：A

138.以下那一项不是大数据提供的用户交互方式是()。

A、企业报表;

B、查询

C、大数据分析挖掘

D、可视化

参考答案： A

139.表示职称为副教授、性别为男的表达式为()

A、职称=’副教授’OR 性别=’男’

B、职称=’副教授’AND 性别=’男’

C、BETWEEN’副教授’AND’男’

D、IN(’副教授’,’男’)

参考答案： B

140.Logistic回归是在商业领域上使用最广泛的预测模型，常用于()分类变量预测和概率预测。

A、四值

B、三值

C、二值

D、一值

参考答案： C

141.检查异常值常用的统计图形是()

A、柱状图

B、箱线图

C、 帕累托图

D、气泡图

参考答案： B

142.以下哪条属于个人信息影响的评估场景：()

A、处理敏感个人信息，利用个人信息进行自动化决策

B、委托处理个人信息、向第三方提供个人信息、公开个人信息

C、向境外提供个人信息

D、 以上皆是

参考答案： D

143. 《个人信息保护法》对于企业的影响不包括()

A、需要强化个人信息处理这主体责任

B、明确了可以量化的中国版执行罚则

C、需要设立负责处理个人信息保护相关事务的专门机构和指定代表

D、收集和处理数据时可以不遵循“最小化”原则

参考答案： D

144.实际由源业务系统自动接入的指标数据占指标体系中应接指标总数的比例称为()

A、指标数据自动采集率；

B、指标数据接入率

C、指标数据接入及时率；

D、指标数据完整率

参考答案： A

145. 假设检验中，拒绝域的边界称为()

A、临界值

B、临界点

C、置信水平

D、边际值

参考答案： A

146. 以下关于大数据的概念和理解不正确的是()

A、大数据是指无法再容许的时间内用常规的软件工具对其内容进行抓取、管理和处理的数据集合，大数据规模的标准是持续变化的，当前泛指单一的数据集的大小在几十TB和 PB之间；

B、大数据是一项技术，能够对复杂海量数据进行实时获取、传输、存储、加工和利用的高薪技术；

C、大数据是一种挑战，现有的数据采集、传输、存储、处理和分析技术已无法适用于现有的需要；

D、大数据是一个时代，拥有大数据是时代的特征、解读大数据是时代的任务、应用大数据是时代的机遇。

参考答案： A

147. 数据收集的标准为()而非动用企业全部数据。

A、一致性、可靠性、时效性

B、相同性、可靠性、时效性

C、C相关性、可靠性、时效性

D、一致性、可靠性、实际性

参考答案： C

148.MySQL 是()

A、操作系统；

B、数据库

C、聊天软件

D、浏览器

参考答案： B

149. 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、主成分分析

参考答案： B

150. 某超市研究销售记录发现，购买奶的很概率会购买包，这种属于数据挖掘的哪类问题?()

A、聚类分析

B、关联规则

C、分类分析

D、自然语言处理

参考答案： B

151. 苹果公司对 IPHONE的降价行为属于()