第1章 作业

1. 【单选题】1. 下列属于数据挖掘任务的是（ ）

A. 根据性别划分公司的顾客

B. 计算公司的总销售额

C. 预测一对骰子的结果

D. 利用历史记录预测公司的未来股价

[答案] D

2. 【单选题】2. 可以在不同维度合并数据，从而形成数据立方体的是（ ）

A. 数据库

B. 数据源

C. 数据仓库

D. 数据库系统

[答案] C

3. 【单选题】3. 目的是缩小数据的取值范围，使其更适合于数据挖掘算法的需要，并且能够得到和原始数据相同的分析结果的是（ ）

A. 数据清洗

B. 数据集成

C. 数据变换

D. 数据归约

[答案] D

4. 【单选题】4. 下述四种方法哪一种不是常见的分类方法（ ）

A. 决策树

B. 支持向量

C. K-Means

D. 朴素贝叶斯分类

[答案] C

5. 【单选题】5. 下列任务中，属于数据挖掘技术在商务智能方面应用的是（ ）

A. 欺诈检测

B. 垃圾邮件识别

C. 根据因特网的搜索引擎查找特定的Web页面

D. 定向营销

[答案] D

6. 【单选题】6. 异常检测的应用包括（ ）

A. 网络攻击

B. 预测某股票的未来价格

C. 计算公司的总销售额

D. 根据性别划分公司顾客

[答案] A

7. 【单选题】7. 将原始数据进行集成、变换、维度规约、数值规约是哪个步骤的任务（ ）

A. 频繁模式挖掘

B. 分类和预测

C. 数据预处理

D. 数据流挖掘

[答案] C

8. 【单选题】8. KDD是（ ）

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

[答案] A

9. 【单选题】9. 下列有关离群点的分析错误的是（ ）

A. 一般情况下离群点会被当作噪声而丢弃

B. 离群点即是噪声数据

C. 在某些特殊应用中离群点有特殊的意义

D. 信用卡在不常消费地区突然消费大量金额的现象属于离群点分析范畴

[答案] B

10. 【单选题】10. 下列关于模式识别的相关说法中错误的是（ ）

A. 模式识别的本质是抽象出不同事物中的模式并由此对事物进行分类

B. 医疗诊断属于模式识别的研究内容之一

C. 手机的指纹解锁技术不属于模式识别的应用

D. 自然语言理解也包含模式识别问题

[答案] C

11. 【单选题】11. （ ）不属于数据挖掘的应用领域。

A. 商务智能

B. 信息识别

C. 搜索引擎

D. 医疗诊断

[答案] D

12. 【单选题】12. 目前数据分析和数据挖掘面临的挑战性问题不包括（ ）

A. 数据类型的多样化

B. 高维度数据

C. 离群点数据

D. 分析与挖掘结果可视化

[答案] C

13. 【填空题】1.常见的机器学习方法有\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_。

[答案] 监督学习; 半监督学习; 无监督学习;

14. 【填空题】2.\_\_\_\_\_是从大规模的数据中抽取或挖掘出感兴趣的知识或模式的过程或方法。

[答案] 数据挖掘;

15. 【填空题】3.\_\_\_\_\_是指在数据集中频繁出现的模式。

[答案] 频繁模式;

16. 【填空题】4.\_\_\_\_\_是指全局或局部范围内偏离一般水平的观测对象。

[答案] 离群点;

17. 【填空题】5.\_\_\_\_\_是数据仓库系统的主要应用。

[答案] 联机分析处理;

18. 【填空题】6.\_\_\_\_\_是通过建立模型预测离散的标签，而\_\_\_\_\_则是通过建立连续值模型推断新的数据的某个数值型属性。

[答案] 分类; 回归;

19. 【填空题】7.数据库是面向\_\_\_\_\_的设计，数据仓库是面向\_\_\_\_\_设计的。

[答案] 事务; 主题;

20. 【填空题】8.数据挖掘主要侧重解决四类问题：\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_和预测（定量、定性）。

[答案] 分类; 关联; 聚类;

21. 【填空题】9.数据分析是指采用适当的\_\_\_\_\_方法对收集到的数据进行分析、概括和总结。

[答案] 统计分析;

22. 【填空题】10.\_\_\_\_\_是一个目标类数据的一般特性或特性的汇总。

[答案] 特征化,数据特征化;

23. 【判断题】1.无监督学习可以在没有标记的数据集上进行学习。（ ）

[答案] 对

24. 【判断题】2.聚类就是把一些对象划分为多个组或者聚簇，从而使同组内对象间比较相似而不同组对象间差异较大。（ ）

[答案] 对

25. 【判断题】3.事务数据库的每个记录代表一个事务。（ ）

[答案] 对

26. 【判断题】4.数据仓库和数据库其实是相同的，都是数据或信息的存储系统。（ ）

[答案] 错

27. 【判断题】5.区分是将目标类数据对象的一般特性与一个或多个对比类对象的一般特性进行比较。（ ）

[答案] 对

28. 【判断题】6.离群点因偏离一般水平而不需要考虑和研究。（ ）

[答案] 错

29. 【判断题】7.聚类过程的输入对象有与之关联的目标信息。（ ）

[答案] 错

30. 【判断题】8.数据挖掘的主要任务是从数据中发现潜在的规则，从而能更好的完成描述数据、预测数据等任务。（ ）

[答案] 对

31. 【判断题】9.数据挖掘的目标不在于数据采集策略，而在于对于已经存在的数据进行模式的发掘。（ ）

[答案] 对

32. 【判断题】10.数据仓库一般存储在线交易数据，数据库存储的一般是历史数据。（ ）

[答案] 错

33. 【判断题】11.数据分析是指采用适当的统计分析方法对收集到的数据进行分析、概括和总结，对数据进行恰当的描述，并提取出有用的信息的过程。（ ）

[答案] 对

34. 【判断题】12.数据分析的定义：数据分析就是对数据进行分析。专业的说法，数据分析是指根据分析目的，用适当的统计分析方法及工具，对收集来的数据进行处理与分析，提取有价值的信息，发挥数据的作用。（ ）

[答案] 对

35. 【判断题】13.数据库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策。（ ）

[答案] 错

第2章 作业

1. 【单选题】1. 下面哪个不属于数据的属性类型（ ）。

A. 标称

B. 序数

C. 区间

D. 相异

[答案] D

2. 【单选题】2. 属于定量的属性类型是（ ）。

A. 标称

B. 序数

C. 区间

D. 相异

[答案] C

3. 【单选题】3. 一所大学内的各年纪人数分别为：一年级200人，二年级160人，三年级130人，四年级110人。则年级属性的众数是（ ） 。

A. 一年级

B. 二年级

C. 三年级

D. 四年级

[答案] A

4. 【单选题】4. 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的值映射到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

[答案] D

5. 【单选题】5. 考虑数据集{12 24 33 24 55 68 26}，其四分位数极差是（ ）。

A. 31

B. 24

C. 55

D. 3

[答案] A

6. 【单选题】6. 光年所属的属性类型为（ ）。

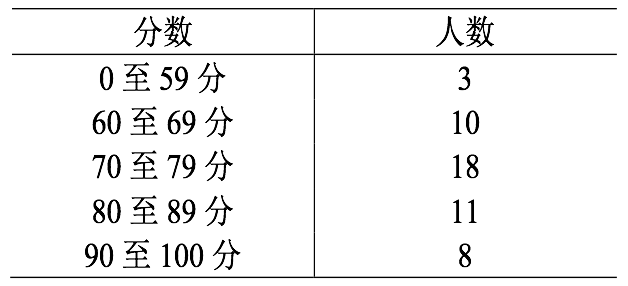
A. 标称属性

B. 序数属性

C. 区间标度属性

D. 比率标度属性

[答案] D

7. 【单选题】8. 某班数学期末考成绩分组数据如下，则数据的中位数区间是（ ）。

A. 60至69分

B. 70至79分

C. 80至89分

D. 90至100分

[答案] B

8. 【单选题】9. 军衔所属的属性类型为（ ）。

A. 标称属性

B. 序数属性

C. 二元属性

D. 数值属性

[答案] B

9. 【单选题】11. 计算p1(2，1，4，10)和p2(3，0，3，8)两个对象之间的曼哈顿距离（ ）。

A. 4

B. 2

C. 5

D. 3

[答案] C

10. 【单选题】12. 用AM 和PM 表示的时间的属性类型是（ ）。

A. 标称

B. 序数

C. 区间

D. 二元

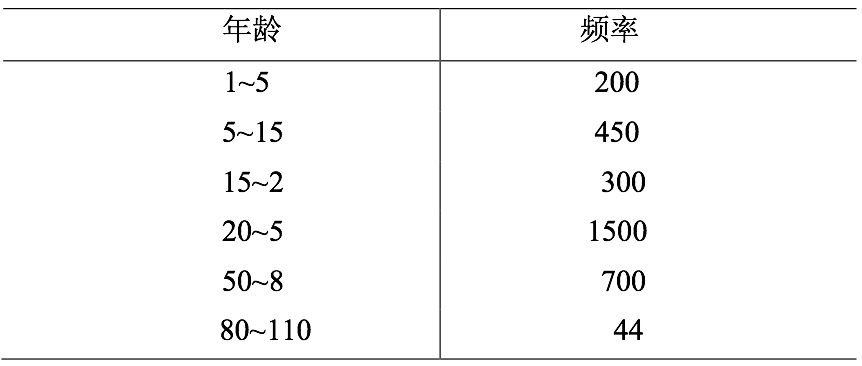
[答案] D

11. 【填空题】1. 属性可以分为\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_和\_\_\_\_\_四类。 <table cellpadding="0" cellspacing="0" width="180"><colgroup><col width="240" style="width: 180pt;"/></colgroup><tbody><tr class="firstRow"></tr></tbody></table>

[答案] 标称属性,标称; 序数属性,序数; 二元属性,二元; 数值属性,数值;

12. 【填空题】2. 中心趋势度量包括\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_和\_\_\_\_\_。

[答案] 均值; 中位数; 众数; 中列数;

13. 【填空题】4. 假设给定的数据集的值已经分组为区间。区间和对应的频率如下。
则数据的近似分组中位数是\_\_\_\_\_。

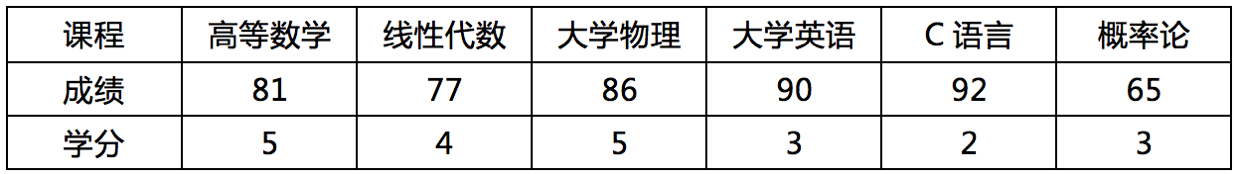
[答案] 32.94;

14. 【填空题】5. 某部门的月薪情况如下（单位：千元），30,33,48,50,53,53,57,60,64,68,70,70,90,则该部门员工的月薪的均值为\_\_\_\_\_。

[答案] 57.38;

15. 【填空题】6. 数据集{5,10,11,13,15,15,35,50,55,72,92,204,215}的中位数为\_\_\_\_\_，众数为\_\_\_\_\_。

[答案] 35; 15;

16. 【填空题】7. 假设小明某一学期的考试成绩及每门课的学分如下表所示，则小明本学期的加权平均成绩为\_\_\_\_\_。

[答案] 81.45;

17. 【填空题】8. 给定两个对象的元组x=(4,3,5,1)与y=(1,6,7,3)，它们之间的余弦相似度为\_\_\_\_\_。

[答案] 0.86;

18. 【填空题】9. 小明参加数学竞赛选拔赛，他十次测试成绩为：76,84,90,86,81,87,86,82,85,83。则小明同学十次测试成绩的方差是\_\_\_\_\_。

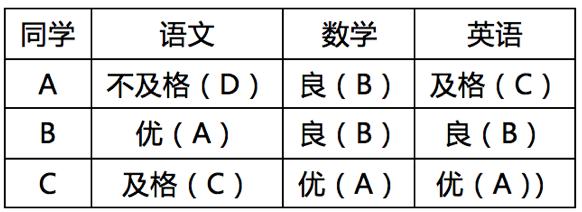
[答案] 13.2,13.20;

19. 【填空题】10. 已知点A的空间坐标为(6,30,9)，点B的空间坐标为(10,35,2)，则A与B之间的切比雪夫距离为\_\_\_\_\_。

[答案] 7;

20. 【填空题】11. 只有非零值才重要的二元属性被称作\_\_\_\_\_。

[答案] 非对称二元属性,非对称二元;

21. 【填空题】13. 某组同学的成绩评定结果如下，
则A与B，A与C，B与C之间的相异性是\_\_\_\_\_、\_\_\_\_\_和\_\_\_\_\_。

[答案] 1.06; 0.82; 0.82;

22. 【填空题】14. 给定两个词向量元祖A(6,2,5,8)和B(4,5,2,6)，则这两个对象之间的闵可夫斯基距离（其中h=4）是\_\_\_\_\_。

[答案] 3.73;

23. 【填空题】15. 假定用于分析的数据包含属性age，数据元组中age 的值如下(按递增序)：  
13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,45,52,70。则该组数据的中列数是\_\_\_\_\_。

[答案] 41.5,41.50;

24. 【判断题】1. 人的血型属于标称属性。

[答案] 对

25. 【判断题】2. 酒店的星级属于数值属性。

[答案] 错

26. 【判断题】3. 某品种玉米苗中抽取10株，测得它们的株高为：21,42,39,14,19,22,37,41,40,25（单位：cm），则玉米苗株高的标准差为10.21。

[答案] 对

27. 【判断题】4. 离散属性总是具有有限个值。

[答案] 错

28. 【判断题】5. 标称属性的值提供了足够的信息用于区分对象。

[答案] 对

29. 【判断题】6. 数值属性的值提供足够的信息确定对象的顺序。

[答案] 错

30. 【判断题】7. 欧几里得距离、曼哈顿距离、闵可夫斯基距离和切比雪夫距离均满足非负性、同一性和三角不等式。

[答案] 对

31. 【判断题】8. 余弦相似性利用向量空间中两个向量夹角的余弦值来衡量两个个体间的差异。余弦值越接近0，夹角越大，向量之间匹配越大。余弦值越接近1，夹角越小，向量之间匹配越小。

[答案] 错

32. 【判断题】9. 二元属性的相异性有两种，一种是对称的二元相异性，另一种是非对称的二元相异性。

[答案] 对

33. 【判断题】10. 已知点X的空间坐标为(3,4,2)，点Y的空间坐标为(1,6,1)，则X与Y的欧几里得距离为3。

[答案] 对

34. 【判断题】4. 计算由不对称的二元变量描述的对象间的相异度可以使用 Jaccard 系数；计算用分类变量描述的对象间的相异度可以采用属性值匹配的方法(属性值匹配，相似度为1，否则为0)；

[答案] 对

第3章 作业

1. 【单选题】1. 下面属于维归约常用的线性代数技术的有（ ）

A. 主成分分析

B. 特征提取；

C. 特征加权

D. 离散化

[答案] A

2. 【单选题】2. 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. 频繁模式挖掘

B. 分类和预测

C. 数据预处理

D. 数据流挖掘

[答案] C

3. 【单选题】3. 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 使用如下每种方法将它们划分成四个箱。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

[答案] B

4. 【单选题】4. 数据清理不包括以下哪些处理（ )

A. 缺失值的处理

B. 噪声的处理

C. 重复数据的处理

D. 不一致数据的处理

[答案] C

5. 【单选题】5. 数据规范化方法包括（ ）

A. 数据归约

B. 数据泛化

C. 数据集成

D. 最小最大规范化

[答案] D

6. 【单选题】6. 给定一组二维样本S,S = {S1, S2, S3, S4, S5}= {(1,2),(3,0),(4,3),(5,7),(1,6)},在距离阈值d大于等于4、非邻点样本的阈值部分p大于等于3时的噪声数据为（ ）

A. S4, S5

B. S1, S4

C. S2, S5

D. S1, S5

[答案] A

7. 【单选题】7. 两组向量x={6,4,7,10,8},y={5,6,1,4,12}的协方差为（ ）

A. 0.1

B. 0.2

C. 0.3

D. 0.4

[答案] B

8. 【单选题】8. 假定用于分析的数据包含属性age,数据元祖中age的值如下(按递增序)：13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,33,35,35,35,35,36,40,45,46,52,70。对age值35进行小数定标规范化转换，转换后的值为（ ）

A. 0.32

B. 0.38

C. 0.35

D. 0.40

[答案] C

9. 【单选题】9. 已知某工厂车间工人的年终奖，其均值为34 349元，标准差为16 928元，对于观测值为57 000元，使用z-score方法对其进行规范化为（ ）

A. 1.34

B. 1.43

C. 0.35

D. 0.57

[答案] A

10. 【单选题】10. 假定用于分析的数据包含属性age,数据元祖中age的值如下(按递增序)：13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,33,35,35,35,35,36,40,45,46,52,70。使用min-max规范化，将age值35转换d到[0.0,1.0]区间的值为（ ）

A. 0.35

B. 0.70

C. 0.175

D. 0.386

[答案] D

11. 【多选题】1. 下列数据变换类型及方法正确的是（ ）

A. 数据平滑：去噪，将连续数据离散化，增加粒度

B. 数据聚集：对数值属性进行监督或无监督离散化

C. 特征构造：构造出新的属性

D. 数据规范化：使数据按照比例缩放，落入特定区域

[答案] A C D

12. 【多选题】2. 数据变换的类型包括（ ）

A. 数据预处理

B. 数据泛化

C. 数据离散化

D. 特征构造

[答案] C D B

13. 【多选题】3. 数据归约技术包括（ ）

A. 维归约

B. 数量归约

C. 数据压缩

D. 数据清理

[答案] A B C

14. 【多选题】4. 下面哪些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

[答案] A B C

15. 【多选题】5. 在现实世界的数据中，元组在某些属性上缺少值是常有的。描述处理该问题的各方法正确的有（ ）

A. 忽略元组

B. 从数据中挑选一个数据填写

C. 使用属性的平均值填充空缺值

D. 使用与给定元组属同一类的所有样本的平均值

[答案] C D A

16. 【多选题】6. 数据清理的原则包括（ ）

A. 相异性原则

B. 连续性原则

C. 唯一性原则

D. 空值原则

[答案] D C B

17. 【填空题】2. 高质量数据的要求有\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_。

[答案] 准确性; 一致性; 完整性;

18. 【填空题】3. 数据预处理的技术手段包括\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_。

[答案] 数据清理; 数据变换; 数据集成; 数据归约;

19. 【填空题】5. 假定用于分析的数据包含属性age,数据元祖中age的值如下(按递增序)：13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,33,35,35,35,35,36,40,45,46,52,70。已知age的标准偏差为12.94年，使用z-score规范化对age值35进行转换，转换后得到的值为\_\_\_\_\_。

[答案] 0.386,0.38,0.39;

20. 【填空题】6. 在现实世界的数据中，元组在某些属性上的缺失值是常有的。处理该问题的常用方法有\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_。

[答案] 删除数据对象或属性; 忽略遗漏值; 估计遗漏值;

21. 【填空题】8. 对数据进行偏差检测的3个原则分别是\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_。

[答案] 唯一性原则; 空值原则; 连续性原则;

22. 【填空题】9. 数据归约技术包括\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_。

[答案] 维归约; 数据压缩; 数量归约;

23. 【填空题】10. 在使用分箱法实现特征离散化时，可以用每个箱中的\_\_\_\_\_\_\_\_\_\_\_\_\_或\_\_\_\_\_\_\_\_\_\_\_\_\_替换箱中所有的值。

[答案] 中位数,中值; 平均值,平均数;

24. 【填空题】11. 某工厂车间工人的年终奖金结果如下（已按照递增排序）：15 750，20 000，25 000，36 000，48 000，50 000，52 000，56 000，57 000，68 000，10 000，135 000（元）。使用最小-最大值规范化将奖金57 000转换到[0.0, 1.0]区间为 \_\_\_\_\_ ；使用小数定标规范化将奖金57 000转换到[-1,1]区间为 \_\_\_\_\_

[答案] 0.35; 0.057;

25. 【填空题】12. 数据清理一般需要对 \_\_\_\_\_ 和\_\_\_\_\_ 进行处理。

[答案] 缺失值; 噪声,噪声数据;

26. 【填空题】 13. 我们要进行数据预处理，是因为原始数据大多都是“ \_\_\_\_\_ ”。

[答案] 脏数据;

27. 【填空题】14. 四种处理缺失数据的方法是\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_。

[答案] 直接删除缺失属性的记录; 使用全局常量填充缺失值; 人工填写; 使用属性的中心趋势度量值填充缺失值;

28. 【填空题】15. 数据平滑的方法包括 \_\_\_\_\_ 、\_\_\_\_\_ 和 \_\_\_\_\_ 。

[答案] 分箱; 聚类; 回归;

29. 【判断题】1. 噪声是指被测量的变量产生的错误或误差。

[答案] 错

30. 【判断题】2. 数据规约技术可以得到数据集的规约表示，虽然小，但仍大致保持原数据的完整性。

[答案] 对

31. 【判断题】3. 数据变换是通过平滑聚集、数据概化、规范化等方式将数据转换成适用于数据挖掘的形式。

[答案] 对

32. 【判断题】4. 数据预处理的任务包括数据清理、数据集成、数据规约和数据泛化。

[答案] 错

33. 【判断题】5. 唯一性原则是指一个属性的每个值都是唯一的，不能和这个属性的其他值相同。

[答案] 对

34. 【判断题】6. 数据变换策略主要包括光滑、聚集、数据泛化、规范化、属性构造和离散化。

[答案] 对

35. 【判断题】7. 每个属性的最大值和最小值之间没有缺失值既满足连续性原则。

[答案] 错

36. 【判断题】8. 数据挖掘所处理的数据必须具有准确性、完整性、一致性、时效性、可信性和可解释性。

[答案] 错

37. 【判断题】9. 数据规约就是指对数据集进行简化表示。

[答案] 对

38. 【判断题】10. 数据集成有助于减少结果数据集的冗余和不一致，可以提高集成之后的挖掘过程的准确性和速度。

[答案] 对

39. 【判断题】11. 在确定数据中的离群点时，一般不必检查整个数据集

[答案] 对

第4章 作业

1. 【单选题】数据仓库是随着时间变化的，下面的描述不正确的是（ ）

A. 捕捉到的新数据会覆盖原来的快照

B. 数据仓库随时间的变化不断增加新的数据内容

C. 数据仓库随着事件变化不断删去旧的数据内容

D. 数据仓库中的综合数据会随着时间的变化不断地进行重新综合

[答案] C

2. 【单选题】关于基本数据的元数据是指（ ）

A. 基本元数据包括与企业相关的管理方面的数据和信息

B. 基本元数据包括日志文件和简历执行处理的时序调度信息

C. 基本元数据包括数据源，数据仓库和应用程序等结构相关的信息

D. 基本元数据包括关于装载和更新处理，分析处理以及管理方面的信息

[答案] D

3. 【单选题】 下面关于数据粒度的描述不正确的是（ ）

A. 数据越详细，粒度就越小，级别也就越高

B. 粒度是指数据仓库小数据单元的详细程度和级别

C. 数据综合度越高，粒度也就越大，级别也就越高

D. 粒度的具体划分将直接影响数据仓库中的数据量以及查询质量

[答案] C

4. 【单选题】有关数据仓库的开发特点，不正确的描述是（ ）

A. 数据仓库开发要从数据出发

B. 数据仓库使用的需求在开发时就要明确

C. 数据仓库的开发是一个不断循环的过程，是启发式的开发

D. 在数据仓库环境中，并不存在操作型环境中所固定的和较确切的处理流，数据仓库中数据分析和处理更灵活，且没有固定的模式

[答案] A

5. 【单选题】在有关数据仓库测试，下列说法不正确的是（ ）

A. 在完成数据仓库的实施过程中，需要对数据仓库进行各种测试

B. 在数据仓库进行测试之前一般不必要制定非常详细的测试计划

C. 系统测试需要对数据仓库的组件进行大量的功能测试和回归测试

D. 当数据仓库的每个单独组件完成后，就需要对他们进行单元测试

[答案] B

6. 【单选题】OLAP技术的核心是（ ）

A. 在线性

B. 快速响应

C. 互操作性

D. 多维分析

[答案] D

7. 【单选题】关于OLAP和OLTP的说法，下列不正确的是（ ）

A. OLAP管理大量历史数据，OLTP仅管理当前数据

B. OLAP主要用于事务和查询处理，而OLTP用于数据分析

C. OLAP中综合提炼的数据主要来自OLTP所依赖的底层数据库

D. OLAP数据较之OLTP数据要进行更多的数据维护或预处理的操作

[答案] B

8. 【单选题】数据仓库设计的三级数据模型不包含（ ）

A. 概念模型

B. 逻辑模型

C. 关系模型

D. 物理模型

[答案] C

9. 【单选题】改变数据立方体维次序的操作是（ ）

A. 切片

B. 切块

C. 上卷

D. 旋转

[答案] D

10. 【单选题】在给定的数据立方体的一个维度上进行的选择操作为（ ）

A. 切片

B. 切块

C. 下卷

D. 旋转

[答案] A

11. 【单选题】下列选项中关于粒度描述错误的是：

A. 粒度是指数据仓库的数据单位中保存数据细化或综合程度的级别；

B. 粒度影不响存放在数据仓库中的数据量的大小

C. 粒度影响数据仓库所能回答查询问题的细节程度；

D. 粒度组织数据的方式有：①简单堆积结构；②轮转综合结构；③简单直接结构；④连续结构。

[答案] B

12. 【单选题】简述数据仓库设计的三级模型及其基本内容,不正确的是。

A. 概念模型设计：对问题域内事务进行描述，是在较高的抽象层次上的设计，其主要内容包括：界定系统边界和确定主要的主题域；

B. 逻辑模型设计：对概念模型细化，定义实体属性及其关系，主要内容包括：分析主题域、确定粒度层次划分、确定数据分割策略、定义关系模式、定义记录系统；

C. 物理数据模型设计：在数据库中建立表及索引，主要内容包括确定数据存储结构、确定数据存放位置、确定存储分配以及确定索引策略等。

D. 三种模型设计时主要考虑的因素有I/O存取时间、空间利用率和维护代价等。

[答案] D

13. 【多选题】关于OLAP的特性，下面正确的是（ ）

A. 集成性

B. 快速性

C. 多维性

D. 可分析性

[答案] B C D

14. 【多选题】 数据模型是数据仓库建设的基础，一个完整、灵活、稳定的数据模型对数据仓库项目的成功起到的重要作用有（ ）

A. 利于数据的整合

B. 消除数据仓库的冗余数据

C. 排除数据描述的不一致性

D. 为整个系统建设提供导航图

[答案] A B C D

15. 【多选题】根据使用情况的不同，元数据可以分为（ ）

A. 静态元数据

B. 技术元数据

C. 业务元数据

D. 动态元数据

[答案] B C

16. 【多选题】关于数据仓库的逻辑模型，正确的说法有（ ）

A. 是数据仓库设计中的核心基础

B. 对概念数据模型的分解和细化

C. 对物理模型设计和实现具有指导作用

D. 为全局服务，集成全方位数据形成统一蓝图

[答案] A B C

17. 【多选题】关于OLAP和OLTP的区别描述，正确的是（ ）

A. OLAP是信息处理，OLTP是操作处理

B. .OLAP面向底层管理人员，OLTP面向高层决策人员

C. OLAP管理大量历史数据，OLTP主要关注当前数据

D. OLAP数据是细节性数据，OLTP数据则是综合性数据

[答案] A C

18. 【填空题】OLAP按照数据存储格式划分，实现方式有 \_\_\_\_\_、\_\_\_\_\_和HOLAP三种

[答案] ROLAP ; MOLAP;

19. 【填空题】根据使用情况的不同，元数据可以分为\_\_\_\_\_和业务元数据；根据数据状态的区别又可分为\_\_\_\_\_和动态元数据。

[答案] 技术元数据 ; 静态元数据;

20. 【填空题】数据仓库按照设计顺序，依次分为\_\_\_\_\_、\_\_\_\_\_ 和\_\_\_\_\_三个设计步骤

[答案] 概念模型设计 ; 物理模型设计; 逻辑模型设计;

21. 【判断题】OLAP技术侧重于把数据库中的数据进行分析、转换成辅助决策信息，是继数据库技术发展之后迅猛发展起来的一种新技术

[答案] 对

22. 【判断题】数据仓库中间层OLAP服务器只能采用关系型OLAP。

[答案] 错

23. 【判断题】数据仓库系统的组成部分包括数据仓库，仓库管理，数据抽取，分析工具等四个部分。

[答案] 错

24. 【判断题】数据仓库测试工作中主要包括单元测试和系统测试。

[答案] 对

25. 【判断题】改变数据立方体维度的操作称为下钻。

[答案] 错

26. 【判断题】数据仓库实际的三级模型中的概念模型不是对软件实际的描述。

[答案] 对

27. 【多选题】 数据仓库就是一个面向什么的数据集合？

A. 主题的

B. 集成的

C. 非易失的

D. 时变

[答案] D C B A

28. 【多选题】 数据仓库体系结构通常采用一种三层体系结构，底层、中间层、顶层分别通常为什么？

A. OLAP服务器

B. 数据仓库服务器

C. 前端工具

D. \_\_\_\_\_

[答案] C B A

第5章 作业

1. 【单选题】回归分析中使用的距离是点到直线的垂直坐标距离，最小二乘准则是指（ ）。

A. 

B. 

C. 

D. 

[答案] D

2. 【单选题】回归分析的步骤为（ ）。① 进行相关分析 ② 建立预测模型 ③ 确定变量 ④ 确定预测值 ⑤ 计算预测误差

A. ①③④⑤②

B. ③①⑤②④

C. ③②①⑤④

D. ③⑤①②④

[答案] C

3. 【单选题】下列变量之间的关系是函数关系的是（ ）。

A. 已知二次函数y＝ax2+bx+c，其中a，c是已知常数，取b为自变量，因变量是这个函数的判别式Δ＝b2－4ac。

B. 光照时间和果树亩产量

C. 降雪量和交通事故发生率

D. 每亩施用肥料量和粮食产量

[答案] A

4. 【单选题】对于回归分析，下列说法错误的是（ ）。

A. 在回归分析中，变量间的关系若是非确定性关系，那么因变量不能由自变量唯一确定

B. 线性相关系数可以是正的，也可以是负的

C. .回归分析中，如果=1，说明x与y之间完全相关

D. 样本相关系数r∈(-1,1)

[答案] D

5. 【单选题】某地区调查了2~9岁儿童的身高，由此建立的身高y（cm）与年龄x（岁）的回归模型为ŷ=8.25x+60.13，下列叙述正确的是（ ）。

A. 该地区一个儿童的身高为142.63cm

B. 该地区2~9岁的儿童每年的身高约增加8.25cm

C. 该地区9岁儿童的平均身高是134.38cm

D. 利用这个模型可以准确地预测该地区每个2~9岁儿童的身高

[答案] B

6. 【单选题】已知对一组观察值做出散点图后确定具有线性相关关系，若对于y=bx+a，求得b=0.51， x̅=61.75" ， y̅=38.14 ，则线性回归方程为（ ）。

A. y = 0.51x + 6.65

B. y = 6.65x + 0.51

C. y = 0.51x + 42.30

D. y = 42.30x + 0.51

[答案] A

7. 【单选题】下表是x和y之间的一组数据，则y关于x的回归方程必过（ ）。 

A. 点(2, 3)

B. 点(1.5, 4)

C. 点(2.5, 4)

D. 点(2.5, 5)

[答案] C

8. 【单选题】反映由模型中解释变量所解释的那部分离差大小的是（ ）。

A. 总离差平方和

B. 回归平方和

C. 残差平方和

D. 可决系数

[答案] B

9. 【单选题】总离差平方和TSS、残差平方和RSS与回归平方和ESS三者的关系是（ ）。

A. TSS>RSS+ESS

B. TSS＝RSS＋ESS

C. TSS<RSS+ESS

D. 

[答案] B

10. 【单选题】决定系数的取值范围是（ ）。

A. <=1

B. >=1

C. 0<<1

D. -1<<1

[答案] C

11. 【单选题】回归分析中定义的（ ）。

A. 解释变量和被解释变量都是随机变量

B. 解释变量为非随机变量，被解释变量为随机变量

C. 解释变量和被解释变量都为非随机变量

D. 解释变量为随机变量，被解释变量为非随机变量

[答案] B

12. 【单选题】最大或然准则是按从模型中得到既得的n组样本观测值的什么最大的准则确定样本回归方程（ ）。

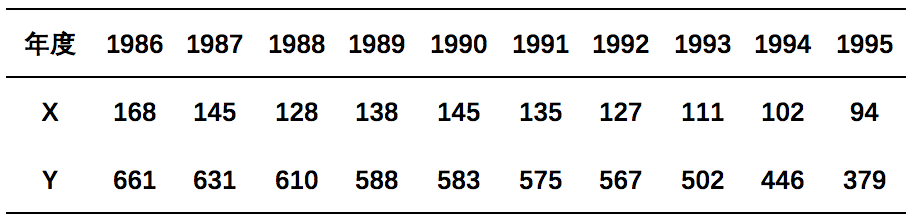
A. 离差平方和

B. 均值

C. 概率

D. 方差

[答案] C

13. 【单选题】下表为日本的汇率与汽车出口数量数据，其中X为平均汇率（日元/美元），Y为汽车出口数量（万辆）,那么，X与Y的相关系数为（ ）。 

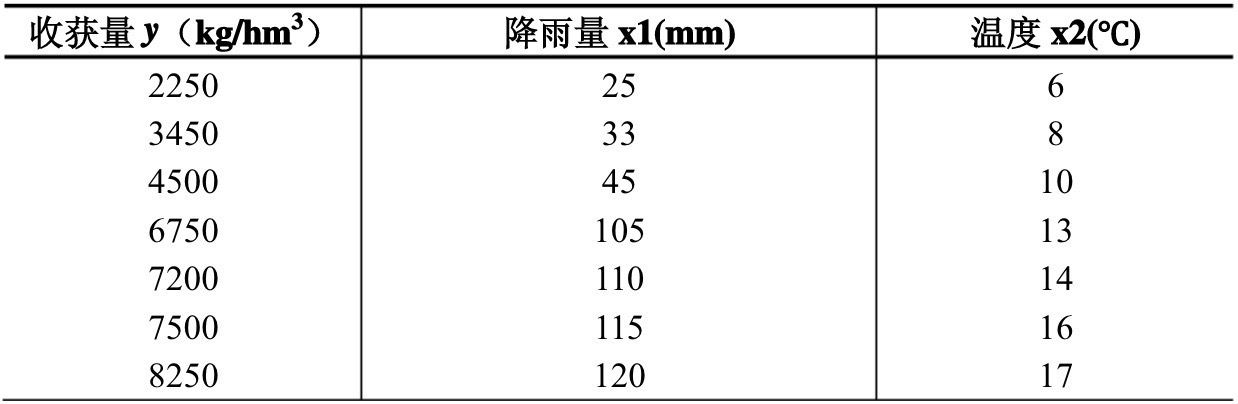
A. 0.0769

B. 0.9321

C. 0.9203

D. 0.0679

[答案] B

14. 【单选题】已知某农场通过试验取得早稻收获量与春季降雨量和春季温度的数据如下,则早稻收获量对春季降雨量和春季温度的二元线性回归方程为（ ）。 

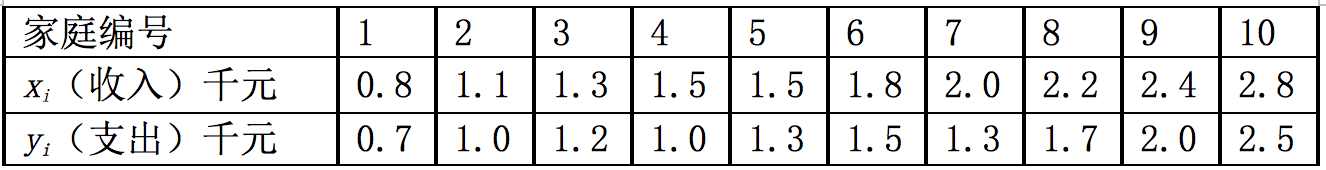
A. 

B. 

C. 

D. 

[答案] D

15. 【单选题】为研究某市家庭收入与月平均生活支出的关系，该市统计部门随机调查了10个家庭，得数据如下,则回归直线方程为()。

A. 

B. 

C. 

D. 

[答案] C

16. 【多选题】在回归分析中，检验线性相关显著性常用的三种检验方法，包含（ ）。

A. 相关系数显著性检验法

B. t检验法

C. F检验法（即方差检验法）

D. χ2检验法

[答案] A B C

17. 【多选题】回归方程 ,则下列说法不正确的有（ ）。

A. y=1.5x - 15

B. 15是回归系数a

C. 1.5是回归系数a

D. x=10时，y=0

[答案] D B A

18. 【多选题】下列关系中具有相关关系的是（ ）。

A. 人的年龄与他拥有的财富之间的关系

B. 曲线上的点与该点的坐标之间的关系

C. 苹果的产量与气候之间的关系

D. 森林中的同一种树木，其断面直径与高度之间的关系

[答案] A C D

19. 【多选题】下列变量中，属于正相关的是（ ）。

A. 收入增加，储蓄额增加

B. 产量增加，生产费用增加

C. 收入增加，支出增加

D. 价格下降，消费增加

[答案] A B C

20. 【填空题】回归分析按照涉及变量的多少，分为:\_\_\_\_\_\_\_\_\_\_

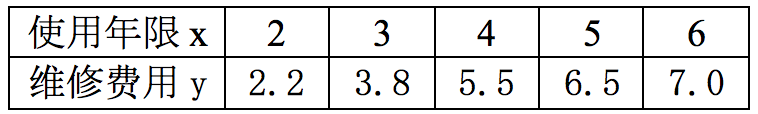
[答案] 一元回归分析; 多元回归分析;

21. 【填空题】 若线性回归方程中的回归系数b = 0,则相关系数r为:\_\_\_\_\_

[答案] 0;

22. 【填空题】某化工厂为预测产品的回收率y，需要研究它和原料有效成分含量x之间的相关关系，现取8对观测值，计算得 ，则其线性回归方程为（系数保留两位小数）\_\_\_\_\_

[答案] y=2.62x+11.47,y=11.47+2.62x;

23. 【填空题】已知关于某设备的使用年限x与所支出的维修费用y（万元），有如下统计资料。若y对x呈线性相关关系，则回归直线方程 ,表示的直线一定过定点：\_\_\_\_\_ 

[答案] （4，5）;

24. 【填空题】使用F检验对多元线性回归方程进行线性关系的显著性检验，所求F值越大，说明线性关系越：\_\_\_\_\_

[答案] 显著;

25. 【填空题】在使用决定系数R2对回归方程进行拟合优度检验时，R2越大表示回归方程的拟合程度越 \_\_\_\_\_，R2越小表示回归方程的拟合程度越\_\_\_\_\_ 。

[答案] 好; 差;

26. 【填空题】\_\_\_\_\_是确定两种或两种以上变数间相互依赖的定量关系的一种统计分析方法。

[答案] 回归分析;

27. 【填空题】F检验运用服从F分布的统计量或方差比作为统计检验，通过\_\_\_\_\_ 度量回归方程的线性关系是否显著

[答案] 显著性水平,显著性水平检验;

28. 【填空题】\_\_\_\_\_是指分类器/回归器正确地预测新的或先前未见过的数据的属性值/类标号的能力

[答案] 准确率,分类准确率;

29. 【填空题】当一些变量的值确定以后，另一些变量的值也随之完全确定的关系，这些变量间的关系完全是已知的，变量间的关系可以用函数关系表示，我们称之为\_\_\_\_\_

[答案] 确定性关系;

30. 【填空题】 变量之间有一定的依赖关系，变量之间虽然互相影响和制约，但由于受到无法估计和控制的因素的影响，使变量间的关系呈现不确定性，当一些变量的值确定以后，另一些变量值虽然随之变化，却不能完全确定，这时，变量间的关系就可以精确地用函数表示，即不能由一个或若干变量的值精确地确定另一个变量的值，我们称这样的关系为\_\_\_\_\_

[答案] 非确定性关系;

31. 【填空题】在回归分析中：被解释变量y是\_\_\_\_\_ 、解释变量x是\_\_\_\_\_ 。

[答案] 随机变量,因变量; 非随机变量,自变量;

32. 【判断题】 回归分析通常用于挖掘关联规则

[答案] 错

33. 【判断题】某块农田粮食的产量与施肥量之间的关系为不确定性关系

[答案] 对

34. 【判断题】多元线性回归是研究处理两个变量之间关系的最简单模型

[答案] 错

35. 【判断题】回归分析是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法

[答案] 对

36. 【判断题】 时序预测与回归预测一样，也是用已知的数据预测未来的值，但这些数据的区别是变量所处的时间不同

[答案] 错

37. 【判断题】回归分析按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析

[答案] 对

38. 【判断题】一元线性回归模型包含一个解释变量和两个被解释变量

[答案] 错

39. 【判断题】用总平方和表示因变量的n个观察值与其均值的误差的总和

[答案] 对

40. 【判断题】一元k次多项式回归方程为

[答案] 对

41. 【判断题】应用回归预测法时，即使变量之间不存在相关关系，也可以对这些变量应用回归预测法

[答案] 错

42. 【判断题】\_\_\_\_\_

[答案] 对

43. 【单选题】SVM通过什么实现线性不可分割问题

A. 特征降维

B. 特征筛选

C. 基尼系数

D. 核函数

[答案] D

第6章 作业

1. 【单选题】1. 下列几种数据挖掘功能中被广泛的用于购物篮分析的是（ ）

A. A.关联分析；

B. B.分类和预测；

C. C.聚类分析；

D. D.演变分析

[答案] A

2. 【单选题】2. 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. A.关联规则发现；

B. B.聚类；

C. C.分类；

D. D.自然语言处理；

[答案] A

3. 【单选题】3. 设X={1，2，3}是频繁项集，则可由X产生多少个关联规则（ ）

A. A. 4；

B. B. 5；

C. C. 6；

D. D. 7；

[答案] C

4. 【单选题】4. 置信度(confidence)是衡量哪种兴趣度度量的指标（ ）

A. A. 简洁性；

B. B. 确定性；

C. C. 实用性；

D. D. 新颖性；

[答案] B

5. 【单选题】5. 关联规则的支持度公式为（ ）

A. A.support(A=>B)=P(A∩B)；

B. B.support(A=>B)=P(B|A)；

C. C.support(A=>B)=P(A∪B)；

D. D.support(A=>B)=P(A|B)；

[答案] C

6. 【单选题】6. 下列指标中，能够度量一个规则的强度，同时衡量两个集合之间的独立性的是（ ）

A. A. 提升度；

B. B. 杠杆度；

C. C. IS度量；

D. D. 确信度；

[答案] D

7. 【单选题】7. 规则∅→A和A→∅的置信度是（ ）

A. A. 50%；

B. B. 75%；

C. C. 90%；

D. D. 100%；

[答案] D

8. 【单选题】8. 令C1 ,C2和C3分别是规则{p}→{q},{p}→{q,r},{p,r}→{q}的置信度。如果假定C1, C2和C3有不同的值，置信度最低的规则是（ ）

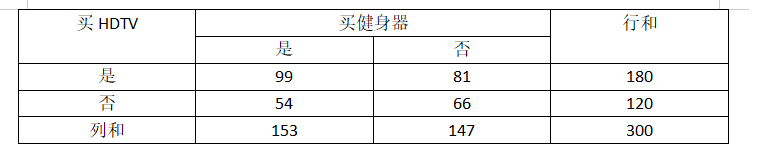
A. A. C1；

B. B. C2；

C. C. C3；

D. D. C2和C3；

[答案] B

9. 【单选题】9. 购买HDTV和购买健身器的情况如下表所示，设最小支持度阈值为0.3，最小置信度阈值为0.6，则{买HDTV }→{买健身器}的支持度为（ ） 

A. A. 0.45；

B. B. 0.55；

C. C. 0.33；

D. D. 0.27；

[答案] C

10. 【单选题】10. 上一题所给的数据中，{买HDTV }→{买健身器}的置信度为（ ）

A. A. 0.65；

B. B. 0.55；

C. C. 0.49；

D. D. 0.33；

[答案] B

11. 【单选题】11. 如果X∈Y,且Y中至少有一项不在X中，那么Y是X的（ ）

A. A. 频繁项集；

B. B. 真超项集；

C. C. 闭频繁项集；

D. D. 极大频繁项集；

[答案] B

12. 【单选题】12. 下列关于Apriori算法的分析中，错误的是（ ）

A. A. Apriori算法基于支持度的剪枝技术，用来控制候选项集的指数增长；

B. B. Apriori算法包括候选集生成和向下封闭检测两个阶段；

C. C. Apriori算法会扫描数据库2次；

D. D. Apriori算法使用逐层搜索的迭代方法；

[答案] C

13. 【单选题】13. 下表所示的购物篮事务数据集中能够提取的3-项集的最大数量是（ ） 

A. A. 3；

B. B. 6；

C. C. 10；

D. D. 20；

[答案] D

14. 【单选题】15. 下列不属于Apriori算法的缺点的是（ ）

A. A. Apriori算法分为两个阶段挖掘频繁项集；

B. B. Apriori算法产生候选项目集时没有排除无用的候选项集；

C. C. Apriori算法在每一步产生候选项目集时循环产生的组合过多；

D. D. 在扫描大型数据库时，Apriori算法会大大增加计算机系统I/O开销；

[答案] A

15. 【多选题】1. 考虑如下的频繁3-项集：{1, 2, 3}，{1, 2, 4}，{1, 2, 5}，{1, 3, 4}，{1, 3, 5}，{2, 3, 4}，{2, 3, 5}，{3, 4, 5}。选出根据Apriori 算法利用上述频繁3-项集生成的候选4-项集（ ）

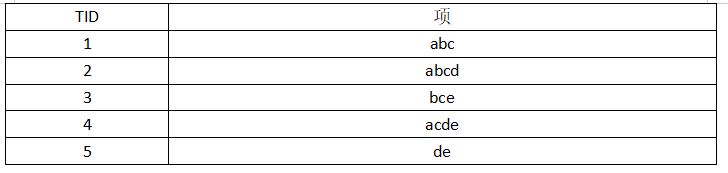
A. A. {1,2,3,4}；

B. B. {1,2,3,5}；

C. C. {1,2,4,5}；

D. D. {2,3,4,5}；

[答案] A B C D

16. 【多选题】2. 下表是一个购物篮，假定支持度阈值为40%，其中哪几个是频繁闭项集（ ）

A. A. abc；

B. B. ad；

C. C. cd；

D. D. de；

[答案] A D

17. 【多选题】3. 一个数据库有5 个事务，如下表所示。设min\_sup=60%，min\_conf = 80%。从下列选项中选出频繁2-项集（ ）

A. A. {M,K}；

B. B. {O,K}；

C. C. {K,E}；

D. D. {M,O}；

[答案] A B C

18. 【多选题】4. 以下关于非频繁模式说法,正确的是（ ）

A. A. 其支持度小于阈值；

B. B. 都是不让人感兴趣的；

C. C. 其支持度大于阈值；

D. D. 对异常数据项敏感；

[答案] A D

19. 【多选题】6. 下列关于FP-growth算法优缺点的表述中，正确的有（ ）

A. A. 相比于Apriori算法，FP-growth算法运行速度要快一个数量级；

B. B. FP-growth算法在建立FP-tree时占用空间较小；

C. C. FP-growth算法无须多次扫描数据库，节省了运行时间；

D. D. FP-growth算法处理产生的条件树时会占用很多资源；

[答案] A C D

20. 【填空题】1. Aprior算法包括\_\_\_\_\_和 \_\_\_\_\_两个基本步骤。

[答案] 连接; 剪枝;

21. 【填空题】2. 关联规则的经典算法包括\_\_\_\_\_和\_\_\_\_\_，其中\_\_\_\_\_的效率更高。

[答案] Apriori ; FP-growth; FP-growth ;

22. 【填空题】3. 如果L2={{a,b},{a,c},{a,d},{b,c},{b,d}}，则连接产生的C3=\_\_\_\_\_

[答案] {{a,b,c},{a,b,d},{a,c,d},{b,c,d}};

23. 【填空题】4. 关联规则的置信度公式为confidence(A=>B)=\_\_\_\_\_

[答案] P(B|A);

24. 【填空题】5. 同时满足\_\_\_\_\_和\_\_\_\_\_的规则称之为强关联规则。

[答案] 最小支持度阈值; 最小置信度阈值;

25. 【填空题】6. 如果一个项集的直接超集都不具有和它相同的支持度计数，则称其为:\_\_\_\_\_

[答案] 闭项集;

26. 【填空题】7. 在挖掘闭模式算法中，直接搜索闭频繁项集，并对结果进行剪枝是最常用的方法，其中剪枝的策略包括\_\_\_\_\_和\_\_\_\_\_ 。

[答案] 项合并 ; 子项集剪枝;

27. 【填空题】8. 不包含任何考察项集的事务称为:\_\_\_\_\_

[答案] 零事务;

28. 【填空题】9. 频繁出现在数据集中的模式称为:\_\_\_\_\_

[答案] 频繁模式;

29. 【填空题】10. 关联规则挖掘任务主要分为\_\_\_\_\_和\_\_\_\_\_两个子任务。

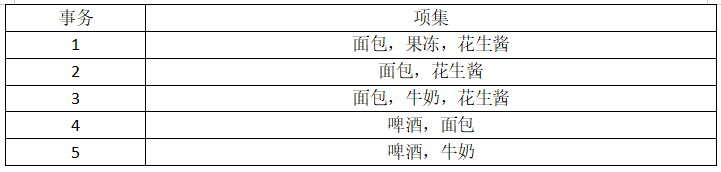
[答案] 频繁项集的产生 ; 关联规则的产生;

30. 【填空题】11.大型数据库中的关联规则挖掘包含 找出所有\_\_\_\_\_和由\_\_\_\_\_产生\_\_\_\_\_两个过程。

[答案] 频繁项集; 强关联规则; 频繁项集;

31. 【填空题】12.FP-growth算法的基本思想是用FP-growth \_\_\_\_\_ 形成频繁集。

[答案] 递归增长;

32. 【填空题】13. 某个食品连锁店每周的事务记录如下表所示，每个事务表示在一项收款机业务中卖出的商品项集，假定min\_sup=40%，min\_conf=40%，使用Apriori算法生成的强关联规则有 \_\_\_\_\_和\_\_\_\_\_两项。 

[答案] {面包}->{花生酱}; {花生酱}->{面包};

33. 【填空题】1. 事物t={牛奶，面包，啤酒}是\_\_\_\_\_项集。

[答案] 3,三;

34. 【填空题】2. FP-growth算法在一次运行中扫描\_\_\_\_\_次数据库。

[答案] 2,两,二;

35. 【填空题】3. 计算关联规则{牛奶}=>{咖啡}的支持度和置信度：\_\_\_\_\_\_\_\_\_\_(答案保留小数点后两位) 

[答案] 0.40; 0.66;

36. 【填空题】4. 从上题的数据中计算牛奶与咖啡之间的提升度和杠杆度：\_\_\_\_\_\_\_\_\_\_(答案保留小数点后一位)

[答案] 1.3; 0.1;

37. 【填空题】5. 一个数据库有5个事务，如下表所示。设min\_sup=60%，min\_conf = 80%。用Apriori算法找出所有3频繁项集(答案中不要有空格，标点符号用半角)：\_\_\_\_\_

[答案] {O,K,E},{O,E,K},{K,O,E},{K,E,O},{E,O,K},{E,K,O};

38. 【填空题】8. 计算{面包（A）=>啤酒（E）}的支持度：\_\_\_\_\_(保留小数点后一位) 

[答案] 0.2,0.20;

39. 【填空题】9. 从上题的数据中计算规则{面包（A）}=>{甜酱（B）}的置信度:\_\_\_\_\_(答案保留小数点后两位)

[答案] 0.25;

40. 【判断题】1. 关联规则挖掘过程是发现满足最小支持度的所有项集代表的规则。

[答案] 错

41. 【判断题】2. 利用先验原理可以帮助减少频繁项集产生时需要探查的候选项个数。

[答案] 对

42. 【判断题】3. 先验原理可以表述为：如果一个项集是频繁的，那包含它的所有项集也是频繁的。

[答案] 错

43. 【判断题】4. 先验原理可以表述为：如果一个项集是频繁的，那包含它的所有非空子集也是频繁的。

[答案] 对

44. 【判断题】5. 具有较高的支持度的项集具有较高的置信度。

[答案] 错

45. 【判断题】6. 如果两个项集的提升度的值小于1，则说明两个项集正相关。

[答案] 错

46. 【判断题】7. 两个项集的全置信度越大，说明两个项集的关系越紧密，反之则关系越疏远。

[答案] 对

47. 【判断题】8. 极大频繁项集的直接超集都不是频繁的。

[答案] 对

48. 【判断题】9. 可信度是对关联规则的准确度的衡量。

[答案] 对

49. 【判断题】10. Apriori算法是一种典型的关联规则挖掘算法。

[答案] 对

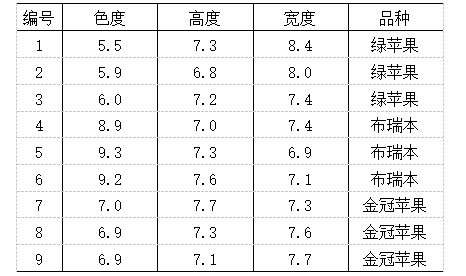
50. 【判断题】11. 关联规则是形如X=>Y的蕴含式，X和Y满足：X和Y是I的真子集，并且X和Y的交集为空集。

[答案] 对

51. 【判断题】12. 设最小支持度阈值为30%，最小置信度阈值为70%，如果一个项集的支持度为50%，则该项集是频繁项集。

[答案] 对

第7章 作业

1. 【单选题】6. 某苹果数据集如下所示， K近邻分类法（K取3）对（色度=8.8，高度=7.1，宽度=7.0）的苹果进行分类的结果为（ ） 

A. A.绿苹果；

B. B.布瑞本；

C. C.金冠苹果；

D. D.绿苹果或布瑞本；

[答案] B

2. 【单选题】7. 下列是有关于是否投保的数据集，第二列至第四列为特征，表中最后一列类别代表是否投保，按照“年薪”进行划分的信息增益率为（ ）

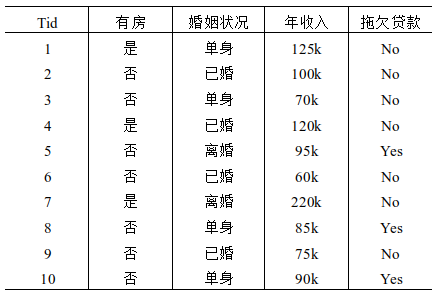
A. A.0.061；

B. B.0.327；

C. C.0.485；

D. D.0；

[答案] B

3. 【单选题】8. 考虑下表中的数据集，使用贝叶斯分类预测记录X=（有房=否，婚姻状况=已婚，年收入=120k）的类标号（ ）

A. A.Yes；

B. B. No；

C. C. Yes or No；

D. D. 不确定；

[答案] B

4. 【单选题】9. 考虑下表中的一维数据集，根据 1-最近邻、3-最近邻、5-最近邻、9-最近邻，对数据点 x=5.0分类，使用多数表决（ ）

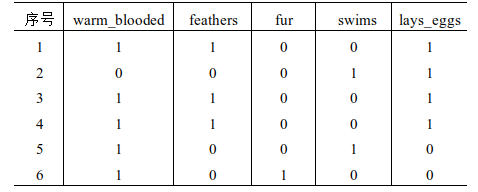
A. A. -、+、-、+；

B. B. +、+、-、-；

C. C. -、-、+、+；

D. D. +、-、+、-；

[答案] D

5. 【单选题】11. 下表给出了一个关于动物类别的训练数据。数据集包含5个属性：warm\_blooded、feathers、fur、swims、lays\_eggs。 若样本按warm\_blooded划分，对应的熵为（ ）

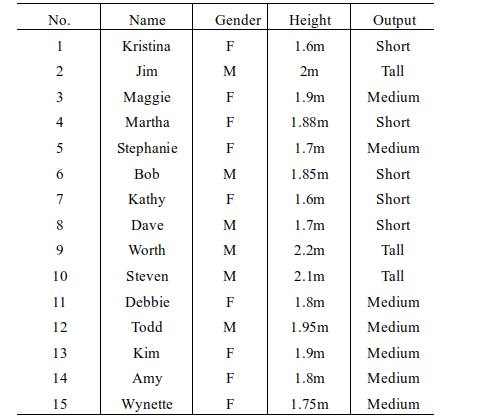
A. A.0.809；

B. B. 0.819；

C. C. 0.609；

D. D. 0.619；

[答案] A

6. 【单选题】14. 下面的例子被分为3类：{Short，Tall，Medium}，Height属性被划分为（0，1.6），（1.6，1.7），（1.7，1.8），（1.8，1.9），（1.9，2.0），（2.0，∞），根据下表，对于t=<Adam，M，1.95m>用贝叶斯分类方法进行分类，则最终结果为（ ）

A. A. Short；

B. B. Tall；

C. C. Medium；

D. D. None；

[答案] B

7. 【单选题】15. 下列是有关于是否购买电脑的数据集，其中学历，是否结婚，收入为特征，表中最后一列类别代表是否购买电脑，则数据集的信息熵为（ ）

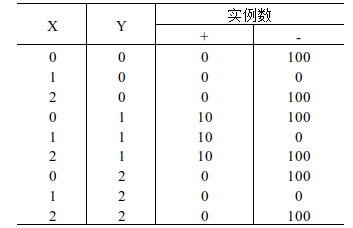
A. A. 0.254；

B. B. 0.376；

C. C. 0.971；

D. D. 0.865；

[答案] C

8. 【单选题】17. 下面的数据集包含两个属性X和Y，两个类标号"+"和"-"。每个属性取三个不同的值： 0, 1或2。"+"类的概念是Y=1， "-"类的概念是X=0或X=2。则由表构建的决策树的F1值（对"+"类定义）是（ ）。

A. A. 0.33；

B. B. 0.5；

C. C. 0.97；

D. D. 1;

[答案] B

9. 【多选题】1. 决策树分类的主要包括（ ）

A. A 对数据源进行OLAP, 得到训练集和测试集；

B. B 对训练集进行训练；

C. C 对初始决策树进行树剪枝；

D. D 由所得到的决策树提取分类规则；

E. E 使用测试数据集进行预测，评估决策树模型；

[答案] A E D C B

10. 【多选题】2. 下列哪些是分类与预测的不同之处（ ）

A. A. 分类的作用是构造一系列能描述和区分数据类型或概念的模型；

B. B. 分类被用作预测目标数据的类的标签；

C. C. 预测是建立一个模型去预测缺失的或无效的并且通常是数字的数据值；

D. D. 预测典型的应用是预测缺失的数字型数据的值；

[答案] A C

11. 【多选题】3. 下列哪些是朴素贝叶斯分类的优缺点（ ）

A. A. 朴素贝叶斯分类做了类条件独立假设，大幅降低了计算开销；

B. B. 需要大量训练数据以覆盖类条件概率空间，引入了很大开销；

C. C. 容易实现并在大多数情况下可以取得较好的结果；

D. D. 类条件独立在实际应用中缺乏准确性，因为变量之间经常存在依赖关系，这种依赖关系影响了朴素贝叶斯分类器的准确性；

[答案] C D

12. 【多选题】4. 支持向量机模型包括（ ）

A. A. 线性可分支持向量机；

B. B. 线性支持向量机；

C. C. 非线性可分支持向量机；

D. D. 非线性支持向量机；

[答案] A B D

13. 【多选题】5. 贝叶斯信念网络（BBN）有哪些特点（ ）

A. A. 构造网络费时费力；

B. B. 对模型的过分问题非常鲁棒；

C. C. 有效地避免过拟合；

D. D. 最小化计算开销；

[答案] A B

14. 【判断题】1. 当一个数据对象同时属于多个类时，很难评估分类的准确率。通常在这种情况下，我们选择的分类器一般趋向于含有这样的特征：最小化计算开销，即使给予噪声数据或不完整数据也能准确预测，在大规模数据下仍然有效工作，提供简明易懂的结果。

[答案] 对

15. 【判断题】2. KNN的主要思想是计算每个训练数据（每个训练数据都有一个唯一的类别标识）到待分类元祖的距离，取和待分类元祖距离最近的k个训练数据集，k个数据中哪个类别的训练数据占多数，则待分类元祖就属于那个类别。

[答案] 对

16. 【判断题】3. 给定决策树，选项有：（1）将决策树转换成规则，然后对结果规则剪枝；（2）对决策树剪枝，然后将剪枝后的树转换成规则。相对于选项（1），选择（2）的优点是更能泛化规则。

[答案] 错

17. 【判断题】4. 给定数据集 D，具有 m 个属性和 |D| 个训练记录，决策树生长的计算时间最多为m×D ×log(|D|)。

[答案] 对

18. 【判断题】5. 将结点划分为更小的后续结点后，结点熵可能会增加。

[答案] 错

19. 【判断题】6. 朴素贝叶斯假设属性之间是相互独立的。

[答案] 对

20. 【判断题】7. 数据分类分为两步：第一步的基本任务是建立一个模型并描述预定的数据类集；第二步的基本任务是评估模型的预测准确率，用准确率可以接受的模型对类标号未知的数据进行分类。

[答案] 对

21. 【判断题】8. 分类规则的挖掘方法通常有:决策树法、贝叶斯法、人工神经网络法、粗糙集法和遗传算法。

[答案] 对

22. 【判断题】9. 决策树是用样本的属性作为树的结构，用样本属性的取值作为树分支的结点。

[答案] 错

23. 【判断题】10. 加入正则化项控制模型的复杂度无法有效地避免过拟合。

[答案] 错

24. 【判断题】11. 随机森林算法过程中只有一个随机过程，即每棵决策树的构建所需的特征是从整体特征集中随机选取的。

[答案] 错

25. 【判断题】12. 惰性学习法的“惰性”体现在它不急于在收到测试对象之前构造分类模型。

[答案] 对

26. 【判断题】13. K近邻算法中K的取值对结果不会产生较大的影响。

[答案] 错

27. 【判断题】14. 支持向量机（SVM）对训练样本求解最小边距超平面。

[答案] 错

28. 【判断题】23. 下表为两周内天气与外出购物的数据集，利用朴素贝叶斯分类预测天气情况为（天气=晴，温度=冷，湿度=高，风力=强）时的结果为不会外出购物。

[答案] 对

第8章 作业

1. 【单选题】 下面哪种距离度量方法为欧几里得距离（ ）

A. 

B. 

C. 

D. 

[答案] B

2. 【单选题】以下哪些算法是分类算法（ ）

A. DBSCAN

B. C4.5

C. K-Mean

D. EM

[答案] B

3. 【单选题】以下哪个算法将两个簇的邻近度定义为不同簇的所有点对的平均逐对邻近度，它是一种凝聚层次聚类技术（ ）

A. MIN（单链）

B. MAX（全链）

C. 组平均

D. Ward方法

[答案] C

4. 【单选题】以下哪个聚类算法不属于基于网格的聚类算法（ ）

A. STING

B. CLIQUE

C. MAFIA

D. BIRCH

[答案] D

5. 【单选题】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

[答案] B

6. 【单选题】DIANA算法中，定义两个数据点之间的平均距离的为（ ）

A. 簇的直径；

B. 平均相异度；

C. 欧几里得距离；

D. 范式距离；

[答案] B

7. 【多选题】下列说法正确的是（ ）

A. 聚类是物理或抽象对象的集合分组成为多个类或簇（cluster）的过程；

B. 聚类中不同簇中的对象差别较小；

C. 聚类和分类要划分的类是未知的；

D. 聚类是一种无指导学习，属于观察式学习，分类则属于有指导的学习，是示例式学习；

[答案] A D

8. 【多选题】下列属于聚类分析方法的是（ ）

A. 基于层次的聚类方法；

B. 基于划分的聚类方法；

C. 基于密度的聚类方法；

D. 基于网格的聚类方法

[答案] A C B D

9. 【多选题】 聚类质量评估的主要任务包括（ ）

A. 估计聚类趋势 ；

B. 确定簇数；

C. 确定层次结构 ；

D. 确定聚类质量

[答案] A B D

10. 【多选题】 聚类分析中最常用的距离有（ ）

A. 切比雪夫距离；

B. 欧几里的距离；

C. 曼哈坦距离；

D. 明可夫斯基距离

[答案]

11. 【多选题】数据的哪些特性对聚类分析具有很强的影响（ ）

A. 高维性；

B. 规模；

C. 稀疏性；

D. 噪声和离群点

[答案] A C B D

12. 【填空题】基于 \_\_\_\_\_ 的聚类方法是基于距离判断数据对象相似度的聚类。

[答案] 划分;

13. 【填空题】基于 \_\_\_\_\_ 的聚类方法可以发现任意形状的簇。

[答案] 密度;

14. 【填空题】层次聚类方法通过将数据对象组织成若干组形成一个相应的树进行聚类，可分为 \_\_\_\_\_ 和 \_\_\_\_\_ 层次聚类

[答案] 凝聚; 分裂;

15. 【判断题】DBSCAN算法能够很好的区分原始数据的形状，但受限于用户指定的参数。

[答案] [答案] 对

16. 【判断题】在通过相似度矩阵评估簇时，如果相似度矩阵是块对角的，说明具有明显分离的簇。

[答案] [答案] 对

17. 【判断题】SSE在无监督的簇评估中能起到很好的作用。

[答案] [答案] 对

18. 【判断题】DBSCAN会把所有点划分到各自的簇中。

[答案] [答案] 错

19. 【判断题】判断簇的个数不属于簇评估。

[答案]

20. 【判断题】 K-中心点算法采用簇中对象的平均值作为参考点。

[答案] [答案] 错

21. 【判断题】DIANA算法中平均相异度指两个数据点之间的平均距离。

[答案] [答案] 对

22. 【判断题】在聚类分析当中，簇内的相似性越大，簇将的差别越大，聚类的效果就越差。

[答案] [答案] 错

23. 【判断题】K均值是一种产生划分聚类的基于密度的聚类算法，簇的个数由算法自动地确定。

[答案] [答案] 错

24. 【判断题】若一个对象不属于任何簇，那么该对象是基于聚类的离群点。

[答案] [答案] 对

25. 【判断题】分裂层次聚类和凝聚层次聚类可以对已经做出的划分决策进行回溯，聚类之间可以交换对象。

[答案] [答案] 错

26. 【判断题】同一种层次的聚类方法，选定的类间距度量不同，聚类的次序和结果也可能不同。

[答案] [答案] 对

27. 【判断题】 STING算法是一种基于网格的多分辨率聚类技术。

[答案] [答案] 对

28. 【判断题】同一层次的聚类方法，选定的类间距度量不同，聚类的结果可能相同也可能不同。

[答案] [答案] 对

29. 【判断题】一种正确选择簇数的启发式方法是使用簇内方差和关于簇数曲线的拐点。

[答案] [答案] 对

30. 【判断题】层次聚类仅局限与基于距离，不能基于密度或空间。

[答案] [答案] 错

第9章 作业

1. 【单选题】与其他观测值的差别如此之大，以至于怀疑他是由不同的机制产生的，且是一个观测值的是（ ）。

A. 边界点

B. 质心

C. 离群点

D. 核心点

[答案] C

2. 【单选题】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

[答案] A

3. 【单选题】一个对象的离群点得分是该对象周围密度的逆。这是基于（ ）的离群点定义

A. 概率

B. 邻近度

C. 密度

D. 聚类

[答案] C

4. 【单选题】下列属于分类模型进行离群点检测的优势的是（ ）。

A. 可以检测所有离群点

B. 不再需要数据和正常数据的距离

C. 建立在统计学原理上，易于理解

D. 当数据充分，检验十分有效

5. 【单选题】 离群点又被称为异常值，它的本质是（ ）。

A. 数据对象

B. 特征

C. 像素点

D. 一个数值

[答案] A

6. 【单选题】 离群点分为全局离群点，（ ）。

A. 部分离群点和条件离群点

B. 条件离群点和集体离群点

C. 部分离群点和集体离群点

D. 异常离群点和集体离群点

[答案] B

7. 【单选题】 使用基于分类的方法进行离群点的识别在实际应用中用的并不是很多，这是因为这种方式（ ）。

A. 分类准确率不高

B. 训练代价高

C. 训练数据难以获取

D. 受训练数据的影响非常大

[答案] D

8. 【单选题】基于近邻的离群点检测方法包括基于距离的离群点检测和（ ）。

A. 基于密度的离群点检测

B. 基于位置的离群点检测

C. 基于数据的离群点检测

D. 基于概率的离群点检测

[答案] A

9. 【单选题】离群点可以是合法的数据对象或者（ ）。

A. 数据关系

B. 特征

C. 像素点

D. 值

[答案] D

10. 【单选题】 如果一个对象不强属于任何簇，那么该对象是（ ）。

A. 基于距离的离群点

B. 基于统计的离群点

C. 基于聚类的离群点

D. 基于密度的离群点

[答案] C

11. 【多选题】下列属于离群点的检测方法的是（ ）。

A. 基于近邻的离群点检测

B. 基于聚类的离群点检测

C. 基于应用价值的离群点检测

D. 基于分类的离群点检测；

[答案] A B D

12. 【多选题】下列属于基于统计的离群点检测流程的是（ ）。

A. 设定数据集的分布模型

B. 不和谐检验

C. 发现离群点

D. 验证离群点

[答案] A C B

13. 【多选题】下列属于离群点检测在数据挖掘中的应用的是（ ）。

A. 市场营销管理

B. 网络监视异常

C. 工业损毁检测

D. 网络入侵检测

[答案] C B D

14. 【多选题】 下列属于离群点产生的原因的是（ ）。

A. 由于数据本身的可变性所致

B. 由于计算的误差或者操作的错误

C. 由于噪声所致

D. 由于操作人员随意添加所致

[答案]

15. 【多选题】下列关于离群点与噪声的说法，错误的是（ ）。

A. 观测值即是真实数据与噪声无关

B. 噪声指被观测数据的随机误差

C. 噪声是异常分析的研究对象

D. 噪声数据是一种离群点

[答案]