简答题（每小题5分，共30分）

1.什么是数据仓库的数据ETL过程？

2.贝叶斯网络的三个主要议题是什么？

3.什么是聚类分析？（2分）聚类分析的应用领域有哪些？（3分）

4.怎样从历史数据中训练出结点之间的条件概率或联合条件概率？

5.简单遗传算法包括哪些步骤？

6.前馈网络和递归网络有什么本质区别?

7.请比较PCA和LDA的区别。

8.请简述数据挖掘中关联规则Apriori算法的思想。

9.请分析特征选择和特征提取有何区别?

10.TF-IDF算法是什么，有什么实际意义?

11.数据挖掘与统计的区别与联系?

聚类和分类有什么区别和联系?

12.什么是分类？（2分）分类的应用领域有哪些？（3分）

13.什么是信息包图法？（2分）它为什么适用于数据仓库的概念模型的设计？（3分）

14.关联规则的分类有哪些？（2分）关联规则挖掘的步骤包括什么？（3分）

15.什么是关联规则？（2分）关联规则的应用有哪些？（3分）

计算题（每小题10分，共30分）

1.给定下表所示的一个事物数据库，写出Apriori算法生成频繁项目集，强关联规则的过程（假定最小支持度=0.5，最小置信度=0.5）。

|  |  |
| --- | --- |
| TID | 项目集 |
| 1 | a,b,c |
| 2 | a,c |
| 3 | a,d |
| 4 | b,e,f |

2.根据下表，利用ID3算法生成决策树，即选择根节点的属性。

|  |  |  |  |
| --- | --- | --- | --- |
| 年龄 | 收入 | 信誉度 | 买保险 |
| ≤40 | 高 | 良 | c2 |
| ≤40 | 高 | 优 | c2 |
| >50 | 中 | 良 | c1 |
| >50 | 低 | 良 | c1 |
| >50 | 低 | 优 | c2 |
| 41~50 | 低 | 优 | c1 |
| ≤40 | 中 | 良 | c2 |
| ≤40 | 低 | 良 | c1 |
| >50 | 中 | 良 | c1 |
| ≤40 | 中 | 优 | c1 |
| 41~50 | 中 | 优 | c1 |
| 41~50 | 高 | 良 | c1 |
| >50 | 中 | 优 | c2 |

3.某电子设备厂所用的元件是由三家元件厂提供的，根据以往的记录，这三个厂家的次品率分别为0.02、0.01、0.03，提供元件的份额分别为0.15、0.8、0.05。设这三个厂家的产品在仓库是均匀混合的，且无区别的标志。

问题1：在仓库中随机地取一个元件，求它是次品的概率。（5分）

问题2：在仓库中随机地取一个元件，若已知它是次品，为分析此次品出自何厂，需求出此元件由三个厂家生产的概率是多少？（5分）

4.根据下表，如要利用ID3算法生成决策树时，就需要将连续的数据离散化，请问是分割点应选到何处？

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Temperature | 40 | 48 | 60 | 72 | 80 | 90 |
| Play Tennis | No | No | Yes | Yes | Yes | No |

5.甲乙丙三人向同一飞机射击。设甲、乙、丙射中的概率分别为0.4、0.5和0.7。又设只有一人射中，飞机坠落的概率为0.2；若有两人射中，飞机坠落的概率为0.6；若有三人射中，飞机必坠落。求飞机坠落的概率。

论述题（每小题15分，共30分）

1.请列出3种数据仓库产品，并说明其优缺点。

2.什么是信息包图法？（7分）它为什么适用于数据仓库的概念模型的设计？（8分）

3.谈一谈你对数据挖掘未来发展趋势的看法。

4.请列出3种数据挖掘中学过的分类方法，并说明其优缺点。

5.列举几项你知道的数据挖掘应用，并论述数据挖掘在其中的作用?

6.简述你对数据仓库未来发展趋势的看法。