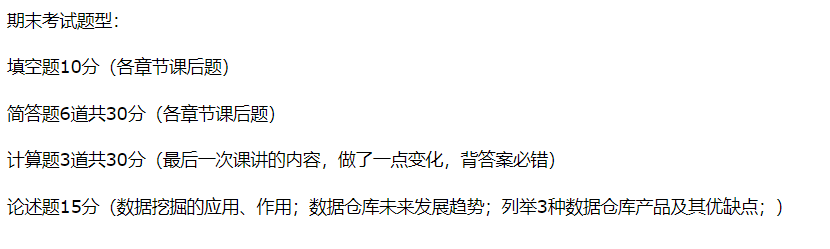
考前复习参考题:



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？

**答：ETI过程就是负责将操作型数据转换成调和数据的过程。**

**ETI过程有四个步骤组成，即抽取、清洗、转换、加载和索引，这些步骤也可进行不同的组合。**

**数据抽取：从源文件和源数据库中获取相关数据用于填充数据仓库，称为数据抽取。数据清洗：是一种使用模式识别和其他技术，在将原始数据转换和转入数据仓库之前来升级原始数据质量的技术。**

**数据转换：数据转换在整个过程中处于中心位置，它把数据从源操作型业务系统转换到企业数据仓库的数据格式。**

**数据加载和索引：即是把数据加载到数 据仓库或数据集市的过程。**

1. 贝叶斯网络的三个主要议题是什么？

**答：①贝叶斯网络预测：是指从起因推测一个结果的推理，也称为由顶向下的推理。**

**②贝叶斯网络诊断：是指从结果推测一个起因的推理，也称为由底至上的推理。**

**③贝叶斯网络学习：是指由先验的贝叶斯网络得到后验的贝叶斯网络的过程。先验的贝叶斯网络是根据用户的先验知识构造的贝叶斯网络，后验贝叶斯网络是把先验贝叶斯网络和数据相结合而得到的贝叶斯网络。**

1. 什么是聚类分析？（2分）聚类分析的应用领域有哪些？（3分）

**聚类分析是将物理的或者抽象的数据集合划分为多个类别的过程，聚类之后的每个类别中任意两个数据样本之间具有较高的相似度，而不同类别的数据样本之间具有较低的相似度。**

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？

**答：要训练条件概率 P(B|A)，可以在历史数据中统计 A 发生的次数 T(A)，然后统计在 A 发生的数据中 B 发生的次数 T(A,B)，条件概率 P(B|A) = T(B)/T(A)。要训练联合条件概率 P(C|A,B)，可以在历史数据中统计 A、B 共同发生的次数 T(A,B)，然后在 A、B 共同发生的数据中统计 C 发生的次数T(A,B,C)，联合条件概率 P(C|A,B)=T(A,B,C)/T(A,B)。以上的符号 A、B、C 可以表示某个事件，也可以表示该事件的相反事件。**

6.前馈网络和递归网络有什么本质区别?

**前馈网络和递归网络的本质区别是网络的某些输出是不是循环作为网络的输入。 前馈网络的所有输出都不能作为输入，而递归网络的某些输出可以循环作为网络的输入。**

7.请比较PCA和LDA的区别。

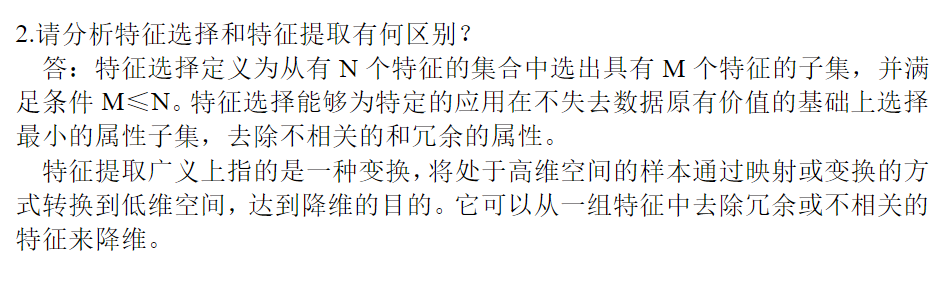
**PCA（主成成分分析）和LDA（线性判别分析）都是经典的降维算法。PCA是无监督的，也就是训练样本不需要标签；LDA是有监督的，也就是训练样本需要标签。PCA是去除掉原始数据中冗余的维度，而LDA是寻找一个维度，使得原始数据在该维度上投影后不同类别的数据尽可能分离开来。**

8.请简述数据挖掘中关联规则Apriori算法的思想。

**基本思路是层次搜索的迭代方法。通过多次扫描交易记录集，产生不同长度的频繁集，首先产生1频繁集，在此基础上经过连接修剪产生2频繁集，直到无法产生频繁集。在第n次循环后产生n频繁集中，首次产生n-候选频繁集的集合C，简称候选集。c中的每一个项集是对两个只有一个项不同的属于（n-1）频繁项集连接产生。**

**Apriori算法的基本原理是用支持度表示关联规则的强度，把具有关联规则的商品看做一个集合。从最小的集合开始，筛选出支持度大于某个值的集合，然后合并集合，再循环，直到找不集合为止。**

9.请分析特征选择和特征提取有何区别?

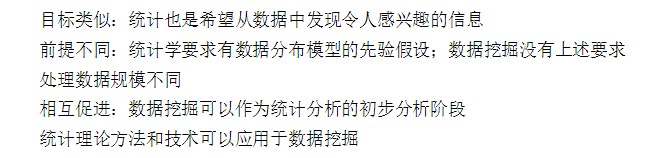


10.TF-IDF算法是什么，有什么实际意义?

TF-IDF是一种用于信息检索与数据挖掘的常用加权技术。TFIDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

常用的领域：搜索引擎、文本相似性、文本摘要

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?

**聚类与分类异同：**

**聚类和分类都是为了对数据进行划分，将相似度高的数据样本划分到一个集合中。**

**聚类要划分的类是未知的，分类则可按已知规则进行；聚类是一种无指导学习，它不依赖预先定义的类和带类标号的训练实例，属于观察式学习，分类则属于有指导的学习，是示例式学习。**

1. 什么是分类？（2分）分类的应用领域有哪些？（3分）

**分类是指把数据样本映射到一个事先定义的类中的学习过程，即给定一组输入的属性向量及其对应的类，用基于归纳的学习算法得出分类。分类问题在商业、银行业、医疗诊断、生物学、文本挖掘、因特网筛选等领域都有广泛应用。**

1. 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）

**答：信息包图定义一个主题内容和主要性能指标之间的关系，其目标就是在概念层满足用户需求。信息包图拥有三个重要对象，（度量）指标、维度和类别。**

**信息包图法也叫用户信息需求表，就是在一张平面表格上描述元素的多维性，其中的每一个维度用平面表格的一列表示。**

**总之信息包图法是一种自上而下的数据建模方法，即从用户的观点开始设计（用户的观点是通过与用户交流得到的），站在管理者的角度把焦点集中在企业的一个或几个主题上，着重分析主题所涉及数据的多维性，这种自上而下的方法几乎考虑了所有的信息以及这些信息源影响业务活动的方式。**

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）

**关联规则的分类：**

**（1）基于规则中涉及到的数据的维数，关联规则可以分为单维的和多维的。**

**（2）基于规则中数据的抽象层次，可以分为单层关联规则和多层关联规则。**

**（3）基于规则中处理的变量的类型不同，关联规则可以分为布尔型和数值型。**

**关联规则挖掘的步骤：**

1. **找出交易数据库中所有大于或等于用户指定的最小支持度的频繁项集；**

**（2）利用频繁项集生成所需要的关联规则，根据用户设定的最小可信度进行取舍，产生强关联规则。**

15.什么是关联规则？（2分）关联规则的应用有哪些？（3分）

**关联规则挖掘最初用来发现超级市场中用户购买的商品之间的隐含关联关系，并用规则的形式表示出来，称为关联规则(Association Rule)。关联规则的应用还包文本挖掘、商品广告邮寄分析、网络故障分析等。**

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 |

 答：(1)由I={a,b,c,d,e,f}的所有项目直接产生1-候选集C1,计算其支持度，去除支持度小于supmin的项集，形成1-频繁集L1。

|  |  |  |  |
| --- | --- | --- | --- |
| 项集C1 | 支持度 | 项集L1 | 支持度 |
| {a} | 3/4 | {a} | 3/4 |
| {b} | 2/4 | {b} | 2/4 |
| {c} | 2/4 | {c} | 2/4 |
| {d} | 1/4 |  |  |
| {e} | 1/4 |  |  |

(2)为发现频繁2-项集L2，首先利用L1中各项目组合连接，来产生2-候选集C2；然后扫描记录集，以获得C2中各项集的支持度。去除支持度小于supmin的项集，形成2-频繁集L2。

|  |  |
| --- | --- |
| 项集C2 | 支持度 |
| {a,b} | 1/4 |
| {a,c} | 2/4 |
| {b,c} | 0 |

结论：a->c:2/3 c->a:100%

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 |

I(买保险)=-5/13\*-8/13\*

I(年龄，买保险)=5/13\*（3/5\*-2/5\*）

+3/13（3/3\*）-0/3\*

+5/13（3/3\*）-0/3\*

I(收入，买保险)= 4/13\*（3/4\*-1/4\*）

+6/13（4/6\*）-2/6\*

+3/13（2/3\*）-0/3\*

I（信誉度，买保险）=7/13\*（2/7\*-5/7\*）

+6/13（3/6\*）-3/6\*

Gain（年龄）=I（买保险）-I（年龄，买保险）

Gain（收入）=I（买保险）-I（收入，买保险）

Gain（信誉度）=I（买保险）-I（信誉度，买保险）

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）

答：设A表示“取到的是一只次品”，Bi（i=1，2，3）表示“所取到的产品是由第i家工厂提供的”

易知B1,B2,B3样本空间S的一个划分，且有

P(B1)=0.15 P(B2)= 0.8 P(B3)=0.05

P(A| B1)=0.02 P(A| B2)=0.01 P(A| B3)=0.03

(1)由全概率公式得P(A)= P(A| B1)\* P(B1)+ P(A| B2)\* P(B2)+ P(A| B3)\* P(B3)

=0.02\*0.15+0.01\*0.8+0.03\*0.05

=0.003+0.008+0.0015=0.0125

(2)P(B1|A)= P(A| B1)\* P(B1)/P(A)=0.003/0.0125=0.24

P(B2|A)= P(A| B2)\* P(B2)/P(A)=0.008/0.0125=0.64

P(B3|A)= P(A| B3)\* P(B3)/P(A)=0.0015/0.0125=0.12

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 |

A点：48和60之间.B点：80和90之间。

计算A点的信息增益：1-2/6(-2/2log(2/2))-4/6(-3/4log(3/4)-1/4log(1/4))

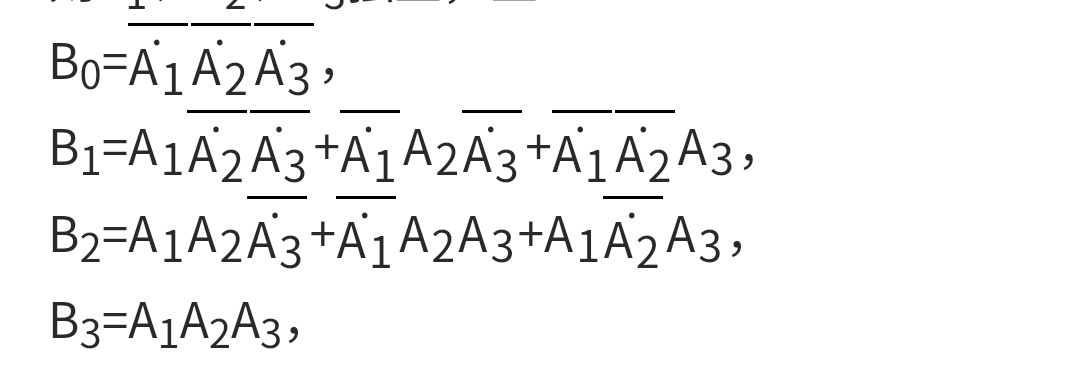
计算B点的信息增益：1-1/6(-1/1log(1/1))-5/6(-2/5log(2/5)-3/5log(3/5))

计算后，选择信息增益大的作为分割点

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。

 解：设A1、A2、A3分别是甲、乙、丙击中飞机，Bi（i=1，2，3）表示有i个人击中飞机，H表示飞机被击落。

则A1、A2、A3独立，且



于是，

P(B0)=(1-0.4)\*(1-0.5)\*(1-0.7)=0.09

P(B1)=0.4\*(1-0.5)\*(1-0.7)+(1-0.4)\*0.5\*(1-0.7)+(1-0.4)\*(1-0.5)\*0.7=0.36

P(B2)=0.4\*0.5\*(1-0.7)+(1-0.4)\*0.5\*0.7+0.4\*(1-0.5)\*0.7=0.41

P(B3)=0.4\*0.5 \*0.7=0.14

依题意有：

P(H|B0)=0, P(H|B1)=0.2 P(H|B2)=0.6 P(H|B3)=1

于是，由全概率公式有：

P(H)==0.09\*0+0.36\*0.2+0.41\*0.6+0.14\*1=0.458

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. **IBM 公司提供了一套基于可视化数据仓库的商业智能（BI）解决方案，包 括：Visual Warehouse（VW）、 Essbase/DB2 OLAP Server 5.0、IBM DB2 UDB，以及来自第 三方的前端数据展现工具（如 BO）和数据挖掘工具（如 SAS）。**
3. **Oracle 数据仓库解决方案主要包括 Oracle Express 和 Oracle Discoverer 两个部分。**
4. **Microsoft 将 OLAP 功能集成到 SQL Server 数据库中，其解决方案包括 BI 平台、BI 终端工具、BI 门户和 BI 应用四个部分。**
5. 谈一谈你对数据挖掘未来发展趋势的看法。

增长和普及： 随着数字化程度的提高，大量的数据被生成，这将促使对数据挖掘工具和技术的需求不断增长。各行各业都在寻求从数据中提取洞察以支持业务决策。

机器学习的整合： 机器学习算法在数据挖掘中发挥着重要作用。未来，随着机器学习技术的不断发展，更多的先进算法将被整合到数据挖掘工具中，以提高模型的准确性和预测能力。

自动化和自动化决策： 随着自动化技术的进步，数据挖掘工具可能变得更加自动化，减少了人工干预的需要。这将使企业更快地从数据中获得价值，并迅速做出基于数据的决策。

大数据和实时分析： 随着大数据技术的不断发展，数据挖掘将需要处理更大规模、更复杂的数据集。实时分析也将变得更加重要，以支持对数据的即时洞察和决策。

隐私和伦理问题： 随着对个人隐私和数据伦理的关注增加，数据挖掘领域可能会更加注重确保数据使用的合规性和透明性。这可能导致更严格的法规和标准。

领域专业化： 随着数据挖掘技术的成熟，不同行业和领域可能会更加专业化，开发特定领域的解决方案和工具，以满足不同行业的需求。

可解释性和可视化： 随着机器学习模型的复杂性增加，对模型结果的解释变得越来越重要。未来的趋势可能包括更加可解释的模型和更强调可视化的工具，以帮助用户理解模型的决策过程。

需要注意的是，这些观点是基于过去的趋势和技术发展的推测，未来的发展也可能受到新的技术突破、市场需求变化和法规制度的影响。

1. 请列出3种数据挖掘中学过的分类方法，并说明其优缺点。

数据挖掘中有许多分类方法，以下是其中三种常见的分类方法以及它们的优缺点：

决策树（Decision Trees）：

优点：

易于理解和解释，树状结构直观。

不需要对数据进行特别的准备，如归一化。

能够处理数值型和离散型数据。

缺点：

对噪声敏感，容易过拟合，特别是在处理复杂数据时。

不稳定，小的数据变化可能导致完全不同的树。

不太适用于处理高维稀疏数据。

K近邻算法（K-Nearest Neighbors, KNN）：

优点：

简单、直观，易于实现。

对异常值不敏感，对数据分布没有假设。

适用于多类别问题。

缺点：

需要大量内存存储训练数据，因为分类时需要计算与所有训练样本的距离。

预测时的计算复杂度高，尤其是在高维数据集上。

对数据中不相关的特征敏感，可能导致不良的分类结果。

支持向量机（Support Vector Machines, SVM）：

优点：

在高维空间中表现良好，适用于处理高维数据。

可以有效处理非线性关系，通过使用核函数进行映射。

具有较好的泛化能力，对于小样本数据也表现出色。

缺点：

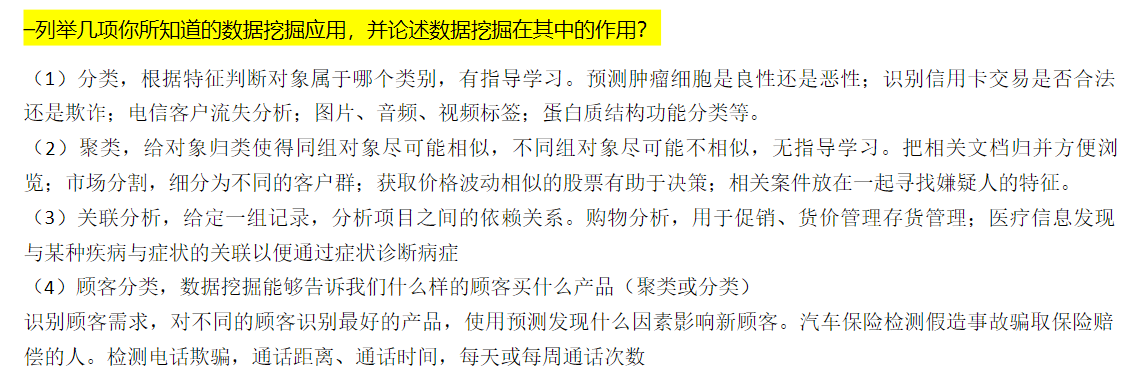
训练和预测时间较长，尤其在大规模数据集上。

对缺失数据敏感，需要对缺失值进行处理。

SVM模型的结果难以解释，不太适用于需要可解释性的场景。

这些分类方法各有优势和限制，选择使用哪种方法通常取决于数据的性质、问题的复杂性以及应用场景的需求。在实际应用中，往往需要尝试多种方法，以找到最适合特定问题的分类器。

1. 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?



1. 简述你对数据仓库未0来发展趋势的看法。

**数据仓库的发展将从数据抽取、存储管理、数据表现和方法论四个方面的发展。**

**数据抽取方面，数据库技术集中体现在系统集成化的方面，它将互连、转换、复制、调度、监控纳入标准化的统一管理，以适应数据仓库本身和数据源的变动，更有利于系统的管理与维护。**

**存储管理方面，未来的发展将会使数据库厂商生产数据仓库引擎，作为数据仓库服务器产品和数据仓库服务器并驾齐驱。**

**数据表现方面，统计分析的算法和功能将普遍集成到联机分析产品中，并与Internet/Web技术紧密结合。**

**方法论方面，数据仓库实现过程中的方法论将更加普及，将成为数据库设计的一个明确的分支，成为管理信息系统设计的必备。**