西南财经大学招收硕士生考试说明及考试大纲

适用专业：大数据管理

考试科目：《数据挖掘综合》

第一部分：考试内容及要求

一. 数据挖掘概述

考试内容

数据挖掘的概念 知识发现过程 数据挖掘数据类型 数据挖掘功能和模式 数据挖掘可利用的技术 数据挖掘应用 数据挖掘的主要问题

考试要求

1.了解数据库系统技术的演变过程；理解数据挖掘的概念；掌握知识发现过程的 7 个步骤。

2.掌握数据挖掘的数据类型；掌握数据挖掘功能和模式；理解数据挖掘与统计学、机器学习的联系和区别；了解数据挖掘的应用领域；了解数据挖掘的主要问题。

二. 数据预处理

考试内容

数据属性 数据基本描述统计 数据预处理概述 数据质量 数据预处理的主要步骤 数据清理 数据集成 数据变换 数据规约 数据离散化

考试要求

1.了解数据对象与属性类型。

2.理解数据的基本统计描述，掌握均值、中位数、众数、极差、四分位数、方差、标准差和四分位数极差的概念和计算方法；了解数据基本统计描述的图形显示；了解度量数据的相似性和相异性。

3.了解进行数据预处理的原因及其重要性；了解数据质量涉及的因素；掌握数据预处理的主要步骤。

4.了解数据清理的概念；了解处理数据缺失值的方法；了解处理噪音数据的方法。

5.理解数据集成的概念；掌握冗余和相关性分析的方法（χ^2 检验，Pearson 积矩系数）。

6.了解数据变换的策略；掌握数据规范化的计算方法（最小-最大规范化、z 分数规范化、按小数定标规范化）。

7.理解数据归约的概念；了解数据归约的策略；了解线性回归、对数线性模型、直方图、聚类、抽样等数据归约方法。

8.理解数据离散化和概念分层的概念；了解数据离散化的方法（分箱、直方图分析、聚类分析、相关分析）。

三.数据仓库和联机分析处理

考试内容

数据仓库基本概念　OLTP 和 OLAP 数据立方体 数据仓库的数据模型 概念分层 典型的 OLAP 操作 数据仓库的设计 数据仓库的实现 数据仓库和数据挖掘

考试要求

1.理解数据仓库的概念和关键特征；了解 OLTP 和 OLAP 的概念和主要区别。

2.了解数据仓库模型的种类；了解元数据库的概念以及与其他数据的区别。

3. 理解数据立方体的概念；了解数据仓库的数据模型（星型模式、雪花模式、事实星座模式）；了解典型的 OLAP 操作方法。

4.了解数据仓库设计的四种视图,了解数据仓库的设计过程和步骤；了解 OLAP 查询处理的步骤。

5.了解三类数据仓库应用；了解多维数据挖掘的重要性。

四. 挖掘频繁模式、关联和相关性

考试内容

频繁项集概念　频繁项集挖掘方法　Apriori 算法　FP-growth 算法

考试要求

1.理解项集、闭项集、频繁项集和关联规则的概念；了解规则兴趣度的两种度量（支持度和置信度）。

2.了解关联规则挖掘的步骤。

3.了解 Apriori 算法的步骤；了解 FP-growth 算法的步骤和优缺点；掌握相关性度量提升度（lift）的计算方法。

五. 分类和预测

考试内容

数据分类和预测的概念 判定树归类算法 信息增益 树剪枝 回归分析 分类法的准确性 组合分类器 类不平衡问题

考试要求

1.理解数据分类的概念；了解分类的两个过程；理解监督学习和非监督学习的区别；了解分类和预测的数据预处理方法；掌握评估分类和预测方法的标准。

2.了解决策树的概念和优缺点；了解决策树归纳分类的主要步骤；了解常用的属性选择度量，掌握信息增益度量的求法；理解两种常用的树剪枝方法。

3.了解评估分类器性能的度量；了解评估分类和预测准确率的方法（混淆矩阵、灵敏度和特小型、F 度量）。

4.了解 K-折交叉验证和自助法的基本思想；了解 ROC 曲线的概念和特点。

5.了解组合分类器的概念和常用的组合分类方法；了解装袋和提升的基本思想以及两者的区

别；了解随机森林的基本思想。

6. 了解类不平衡问题的概念；了解提高类不平衡数据分类准确率的一般方法。

六. 聚类分析

考试内容

聚类分析的概念 聚类方法的分类 算法方法的距离度量 划分方法 层次方法 基于密度的方法 基于网格的方法 聚类评估

考试要求

1. 理解聚类分析的概念；了解聚类分析的应用领域；了解比较聚类方法的标准；了解数据挖掘对聚类的典型要求；了解比较聚类方法的各个方面。

2. 理解划分方法的概念和一般特点，以及典型算法；理解层次方法的概念和一般特点，以及典型算法；理解基于密度的聚类方法的概念和一般特点，以及典型算法；理解基于网格的聚类方法的概念和一般特点，以及典型算法。

3. 理解 K-均值算法的步骤和优缺点。

4. 了解聚类方法的距离度量。

5. 了解聚类评估概念和主要任务；了解测定聚类质量的方法。

第二部分：考试方法和考试时间

数据挖掘综合考试采用闭卷、笔试形式，考试时间为 180 分钟。

第三部分：试卷结构及参考书目

（一）题分：试卷满分为 150 分

（二）题型比例：

名词解释　约 20％

简答题　约 50%

计算编程题　约 30%

（三）参考书目：

1. 韩家炜，Micheline Kamber 裴健著, 范明, 孟小峰译,《数据挖掘概念与技术》，机械工业出版社，2012，ISBN：9787111391401.

2. 周志华著，《机器学习》，清华大学出版社，2016, ISBN: 9787302423287.

3. 李航著，《机器学习方法》，清华大学出版社，2022, ISBN: 9787302597308.

4. .Magnus Lie Hetland 著，袁国忠译，《Python 基础教程（第 3 版・修订版）》，人民邮电出版社，2023, ISBN: 9787115613691.