2022-2023（二）《数据挖掘》期末考试范围与题型参考

➢ 主要知识点

第一章：绪论

- 数据挖掘的应用举例
- 数据挖掘基本流程
- 分类和聚类的异同

第二章：数据表示和预处理

- 属性类型、转换（连续属性转类别：等距、等频；类别转独热编码）
- 属性的常用统计描述（中心趋势、离散趋势）
- 常见距离、相似度度量
- 箱型图 Boxplot 中各部分代表的意义
- 数据规范化（min-max、零均值）

第三章：降维

- 降维的作用
- 主成分分析的原理、目标函数、算法

第四章：分类

✧ 模型评估与性能度量

- 如何评估一个分类模型（数据划分、训练、预测）
- 欠拟合和过拟合的概念、过拟合的原因、判断、解决
- 评估方法（留出法、交叉验证、自助法）
- 性能度量计算（accuracy/error、precision、recall、混淆矩阵、F1）

✧ 决策树

- 决策树属性划分度量（信息增益、gini 指数）
- 决策树基本构造过程和预测
- 了解剪枝的作用、预剪枝和后剪枝的概念

✧ 贝叶斯分类

- 贝叶斯定理、朴素贝叶斯假设（为什么要做该假设、怎么样的假设）
- 朴素贝叶斯算法步骤、不同实现方法：查表与惰性
- 了解拉普拉斯修正（原因、一般做法）

✧ k 最近邻

- K 最近邻算法步骤
- 对 k 的理解、拐点法确定 k

✧ 组合分类

- 组合分类成功的前提
- Bagging 和 boosting 基本框架、各自特点、代表性方法
- Bagging 中增加基分类器随机性的方式

第五章：聚类

- 聚类外部度量和内部度量的定义
- 内部度量的评价标准
- k 均值基本算法步骤、目标函数
- k 均值算法的两个主要问题/局限性（初始化、K 的设定，如何缓解这两个问题）
- 凝聚层次聚类基本步骤、基本连接度量（single、complete、average）

➢ 复习重点：

- 主要复习材料：上课 ppt+课后作业；可以对照每个知识点进行复习，关注具体的例子、对必要性（原因）、局限性、优缺点的讨论等；
- 要求手动实现的算法：最近邻、决策树、朴素贝叶斯、k 均值、层次聚类；
- PCA 降维需要掌握基本步骤和相关公式表示，但不要求具体计算；
- 其他一些计算包括：统计信息、距离、规范化、分类模型性能度量等；
- 代码：两个案例中对统计信息查看、基本预处理的代码实现；

➢ 题目类型

一、论述题：对基本概念的理解、解决思路和方法的整体性理解。
例子：阐述数据挖掘的主要过程、你怎么理解过拟合问题？

二、分析推导题：对数据挖掘过程中相关方法的具体实施和计算。
例子：对给定数据进行零均值规范化、主成分分析的目标是什么，给出目标函数

三、综合计算题：基于给定数据和任务，进行比较完整的数据挖掘过程以及对涉及的概念的理解，需要给出具体计算步骤。
例子：对给定数据进行分类，完成以下各个子任务。(a) 特征转换　(b) 建立朴素贝叶斯分类器　(c) 计算预测集的 F1 值。　(d) 如果数据不断增加，你觉得用查表法好还是惰性学习好？

注意：计算题要给出解题过程和必要公式，不要只给答案！