《数据分析与挖掘》考试大纲

一、 适用的专业

管理科学与工程。

二、 考试方法和考试时间

考试为闭卷考试，考试时间为 3 小时。

三、 考试的主要内容

1、 数据挖掘理论基础

数据挖掘的定义；可以进行数据挖掘的模式类型；数据挖掘的技术；数据挖掘的面向类型；数据挖掘的主要问题；数据对象与属性；数据基本统计描述；数据可视化；度量数据的相似相异性；数据清理；数据集成；数据集成；数据归约；数据变化与离散化；数据仓库的概念；数据仓库建模；数据仓库的设计与使用；数据仓库的实现；数据泛化。

2、 数据挖掘模式

数据挖掘频繁项集、闭项集、关联规则的基本概念；频繁项集挖掘方法；模式评估方法；模式挖掘：一个路线图；多层、多维空间中的模式挖掘；基于约束的频发模式挖掘；挖掘高维数据和巨型模式；挖掘压缩或近似模式；模式探索与应用。

3、 分类

分类的基本概念；决策树归纳；贝叶斯分类方法；基于规则的分类；模型评估与选择；提高分类准确度的方法；贝叶斯信念网络；向后传播分类的方法；支持向量机；用频繁模式分类；惰性学习法或从

近邻学习；其他分类方法如遗传算法、粗糙集方法、模糊集方法；有关分类的相关问题：多类分类、半监督分类、主动学习、迁移学习。

4、 聚类分析

聚类分析的定义；划分的方法；层次方法；基于密度的方法；基于网格的方法；聚类评估；基于概率模型的聚类；聚类高维数据；聚类图和网络数据；具有约束的聚类。

5、 离群点检测

离群点和离群点分析；离群点检测方法；统计学方法；基于临近性的方法；基于聚类的方法；基于分类的方法；挖掘情境离群点和集体离群点；高维数据中离群点检测。

6、 数据挖掘的前沿和趋势

挖掘复杂的数据类型；数据挖掘的其他方法；数据挖掘的应用；数据挖掘与社会

7、 机器学习的基本理论与知识

线性模型：基本形式，线性回归，对数几率回归，线性判别分析，多分类学习；决策树：基本流程，划分选择，剪枝处理，连续与缺失值，多变量决策树；神经网络：神经元模型，感知机与多层网络，误差逆传播算法，全局最小与局部最小；支持向量机：间隔与支持向量，对偶问题，核函数，软间隔与正则化，支持向量机回归，核方法；贝叶斯分类器：贝叶斯决策论，极大似然估计，朴素贝叶斯分类器，EM 算法；半监督学习：未标记样本，生成式方法，半监督 SVM，图半监督方法，基于分歧的方法，

半监督聚类。

四、 试卷结构

试卷满分 100 分，基础知识题目（简答题）占 20%，解答题占 60%，综合性论述题占 20%。

五、 主要参考书

韩家炜 编著，数据挖掘概念与技术，北京：机械工业出版社，2012。

周志华 著，机器学习，北京：清华大学出版社，2016