数据挖掘

定义:数据挖掘、机器学习、AI

数据 类型、统计、预处理(清洗、转换、整合)

数据类型:分类的(标称、序数)、数值的(区间、比率)

数据集类型:维度、稀疏性、分辨率;记录数据(事务数据、数据矩阵、稀疏数据矩阵)、基于图形的数据(对象联系、图形对象)、有序数据(时序数据、序列数据、时间序列数据、空间数据)

频率、众数、百分位数、均数、中位数、极差、方差、协方差矩阵、相关矩阵

分类

有监督学习、步骤

算法:决策树-ID3,C4.5,CART、支持向量机(了解)、朴素贝叶斯、ANN(了解, 深度学习)

模型评估和选择:混淆矩阵(计算指标)和准则、交叉验证

组合方法:Bagging、Boosting、随机森林

频繁模式

定义:关联规则(支持度、置信度)

算法:Apriori、FP-growth(熟练掌握)

聚类

定义:非监督学习

算法:partition-based—k-means、Hierarchical-based—two ways、Density-based—DBSCAN、AP(基础概念)、Local density-based(基础概念)、复杂网络(基本概念)CPM、MCL

四个大题(10个选择20分+3个计算30分(kmeans、频繁、决策树)+4个简答20分+2个论述30分)