《智能制造大数据技术》重点

第一章

1.了解智能制造的基本概念。

2.掌握工业大数据的 5V（规模性（Volume）、多样性（Variety）、高速性（Velocity）、价值密度低（Value）、真实性（Veracity））特性。

3.掌握工业大数据分析流程

第二章

1.数据属性的类型：标称、二元、序数、数值（区间标度、比率标度）。

数据是否连续：连续、离散。

数据是否随时间变化：时序。

2.数据的基本统计描述方法：集中趋势度量和离散趋势度量。

集中趋势度量：度量数据分布的中部或者中心位置。均值、中位数、众数和中列数等。

离散趋势度量：度量数据的离散状态。极差、平均差、方差、标准差、协方差、四分位数以及数据的离散系数等。

3.统计度量的图形展示：折线图、直方图、条形图、箱线图、散点图以及平行坐标图。

4.数据的相似性度量：

样本间的相似程度通过邻近性度量表示。

标称属性的邻近性度量；二元属性的邻近性度量；数值属性的距离度量（欧几里得距离、曼哈顿距离、切比雪夫距离、马氏距离）方法。

变量间的相似程度可以通过相关性分析确定。时间序列自相关、皮尔逊相关系数、秩相关系数

5.因果关系（不考）。

第三章

1.数据预处理的主要任务：数据清洗、数据集成、数据变换和数据归约。

数据清洗：格式内容清洗（时间、日期、数值、全半角等显示格式不一致；内容中有不该存在的字符；内容与该字段应有内容不符），填补缺失的值，光滑噪声（分箱、小波变换、经验模态分解）、异常值清洗（简单统计分析、3σ原则、箱线图分析、基于模型检测、基于距离、基于密度、基于聚类），逻辑错误清洗（去重、去除不合理值）。

数据集成：将来自多个数据源的数据整合成一致的数据存储。

数据变换：数据变换通过数据规范化或者离散化将数据变换成适于挖掘的形式。如：数据规范化、数据离散化。

数据归约：得到数据的归约表示，而使得信息内容的损失最小化。数据归约方法包括维归约（主成分分析、小波变换）、数量归约。使用参数或非参数模型，得到原数据的较小表示。参数模型包括回归和对数线性模型。非参数方法包括直方

图、聚类、抽样等。

第四章
1.频繁模式的基本概念：项、事务、数据库以及模式支持度。
2.三种频繁项集挖掘算法：Apriori 算法、FP-Growth 算法以及垂直数据结构算法的原理、算法流程和实例分析。
3.关联规则挖掘：关联规则的概念、产生（置信度阈值）和评估（提升度、卡方距离）。

第五章
1.聚类的基本概念。
2.三种聚类算法：划分聚类算法（K-means、K-medoids 和 K-means++算法的聚类思想和聚类过程）、层次聚类算法（凝聚的与分裂的层次聚类算法的聚类思想和聚类过程）、基于密度聚类算法（DBSCAN 算法的聚类思想、所涉及参数的基本概念和聚类过程）的原理、算法流程和实例分析。
3.聚类分析性能评估：内部准则法（轮廓系数和 CH 指标）、外部准则法、相对准则法（不考）。

第六章
1.分类的基本概念。
2.四种分类方法：决策树（ID3，C4.5，CART（不考））、朴素贝叶斯、支持向量机、人工神经网络。
3.决策树：基本的决策树理论、属性选择度量及剪枝操作。
4.贝叶斯分类：朴素贝叶斯分类思想。
5.支持向量机基本理论，引出对应的优化问题及相应解法。
6.人工神经网络：BP 的原理及基本过程，常用的激活函数及损失函数。
7.分类模型评价与选择：划分测试集的三种方法（保持法、交叉验证法、自助法），分类器的评价指标（准确率（Accuracy）、错误率（Error Rate）、灵敏性（Sensitive）、特效性（Specificity）、精度（Precision）、召回率（Recall）、F 度量），对分类模型效果的评估（统计显著性检验、ROC 曲线）。
8.组合分类技术：Bagging（并）、Adaboost（串）算法流程、区别。

第七章
1.回归分析的概念、基本步骤。
2.线性回归模型：最小二乘估计法、加权最小二乘方法。
3.高维系数存在的共线性问题及其影响，对应的解决方法（岭回归、Lasso 回归、主成分回归、偏最小二乘回归各自特点）。
4.非线性回归：非线性最小二乘估计法、支持向量回归。
5.模型验证：模型拟合度量（残差图（不考）、杠杆率图（不考）、拟合效果度量（Y 和$\hat{Y}$的相关系数 $\mathrm{Cor}(Y,\hat{Y})$ （不考）、总离差平方和 SST、回归平方和 SSR、残差平方和 SSE））。