数据挖掘

理论、算法与应用

MarkZZZ WeChat: MarkZZZ20XX

课程简介

本课程系统讲授数据挖掘的核心理论与算法，以严格的数学框架为基础，涵盖关联规则、频繁模式、聚类分析、谱聚类、降维、异常检测、推荐系统、图挖掘、流数据挖掘、隐私保护数据挖掘以及因果数据挖掘等主题。课程强调从第一性原理出发的数学推导，包括 Apriori 算法的格理论正确性证明、谱聚类的 Cheeger 不等式、差分隐私的组合定理、PageRank 的 Perron-Frobenius 收敛性分析等。通过严格的定理证明与算法分析，帮助学生建立深厚的数据挖掘理论素养，具备独立开展数据挖掘研究和解决前沿问题的能力。

适合人群

- 计算机科学、应用数学、统计学、运筹学、数据科学等方向的博士生及高年级硕士生
- 具备扎实数学基础，希望深入理解数据挖掘理论根基的研究人员
- 对大规模数据分析、机器学习理论、隐私计算、图数据挖掘感兴趣的交叉学科研究者
- 已修读本科级数据挖掘或机器学习课程，寻求理论深化的工程师与自学者

前置知识

- **线性代数**：矩阵运算、特征值分解、奇异值分解（SVD）、谱定理、矩阵范数、半正定矩阵
- **概率论与数理统计**：随机变量、期望与方差、大数定律、中心极限定理、假设检验、极大似然估计、集中不等式（Markov/Chebyshev/Chernoff）
- **图论基础**：图的表示（邻接矩阵、Laplacian）、连通性、随机游走
- **基础算法与复杂度**：排序、哈希、树/图的基本算法、NP 完全性基础
- **机器学习基础**：监督学习、无监督学习、优化基础（梯度下降）
- **编程基础**：Python/NumPy/Pandas（课程以理论为主，算法分析为辅）

1 课程大纲

讲次	主题	内容概要
1	数据表示与预处理	数据矩阵与特征空间、缺失值处理（MCAR/MAR/MNAR 机制、均值/中位数/多重插补/MICE）、异常值检测与处理、归一化（Min-Max/Z-score/Robust Scaler）、特征工程（多项式特征/交叉特征/Bin 化）、数据质量度量（完整性/一致性/唯一性/时效性）、高维数据的维度诅咒
2	关联规则挖掘	支持度-置信度-提升度框架、Apriori 算法（反单调性引理与完整性/正确性证明）、候选生成与剪枝机制、FP-growth 算法（FP-tree 构造与条件模式基）、关联规则的兴趣度度量（Lift/Conviction/Leverage）、闭频繁项集与极大频繁项集的关系
3	频繁模式的理论	频繁项集格结构（半格/闭包系统）、闭频繁项集（Closed Frequent Itemsets）的完备性证明、极大频繁项集（Maximal Frequent Itemsets）的冗余分析、CHARM 算法的理论基础、基于采样的频繁模式挖掘（Toivonen 算法）、误差界推导（Hoeffding 不等式的应用）
4	聚类基础	K-means（Lloyd 算法收敛性证明、k-means++ 初始化的 $O(\log k)$ 竞争比证明）、层次聚类（单链/全链/Ward 方法、树状图 Dendrogram）、DBSCAN 的正确性（核点/边界点/噪声点的定义与性质）、聚类评估指标（轮廓系数、Calinski-Harabasz、Davies-Bouldin、调整兰德指数）
5	谱聚类	图 Laplacian 理论（非归一化/归一化 Laplacian 的谱性质）、Fiedler 向量与图连通性、谱嵌入（Shi-Malik 的 NCut 松弛）、Cheeger 不等式（完整证明：$h(G)^2/2 \le \lambda_2 \le 2h(G)$）、有效电阻与谱稀疏化、谱聚类的算法流程与复杂度分析
6	降维	PCA（最大方差推导/最小重构误差/谱分解等价性完整证明）、核 PCA（RKHS 框架）、LDA（Fisher 线性判别准则、类间散度矩阵/类内散度矩阵）、t-SNE（Student-t 分布的动机、KL 散度梯度推导）、UMAP（拓扑数据分析基础、黎曼度量近似）

讲次	主题	内容概要
7	异常检测	统计检验方法（Grubbs 检验、Bonferroni 校正、Shapiro-Wilk）、基于密度的方法（LOF 局部离群因子定义与性质）、基于隔离的方法（Isolation Forest 的期望路径长度分析）、基于深度学习的方法（Autoencoder、DAGMM）、时间序列异常检测
8	推荐系统	协同过滤（用户/物品相似度、稀疏性问题）、矩阵分解模型（ALS 收敛性分析、SGD 的随机性与收敛性）、SVD++（隐式反馈的建模）、偏置项处理、基于内容的推荐、冷启动问题、推荐系统的评估指标（RMSE/MAE/NDCG/MAP）
9	图挖掘	PageRank 算法（随机游走解释、Perron-Frobenius 定理、幂迭代收敛性证明）、社区发现（模块度优化：Louvain 算法、谱分析）、影响最大化（贪心算法的 $(1-1/e)$ 近似比证明、CELF 优化）、图嵌入（DeepWalk/Node2Vec/GNN 简介）
10	流数据挖掘	数据流模型（时间/空间约束）、Count-Min Sketch（误差界 ϵ-δ 分析完整证明）、Bloom Filter（误报率分析）、蓄水池采样（Vitter 算法正确性证明）、概念漂移检测（ADWIN 算法、Page-Hinkley 检验）、滑动窗口上的频繁项集
11	隐私保护数据挖掘	差分隐私（ε-DP 定义、全局敏感度、Laplace 机制、Gaussian 机制）、组合定理（顺序组合/并行组合完整证明）、差分隐私下的 ERM（输出扰动/目标函数扰动/梯度扰动）、局部差分隐私（Randomized Response、RAPPOR）、私密数据发布
12	因果数据挖掘与前沿	因果图模型（DAG、d-分离、Markov 条件）、因果发现算法（PC 算法、FCI、LiNGAM）、反事实推断（潜在结果框架、ATE/ATT 估计）、可解释数据挖掘（SHAP、LIME 在数据挖掘中的应用）、大规模数据挖掘系统（Spark MLlib、分布式协同过滤）、课程总结与研究前沿

2 参考书目

1. Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011. (数据挖掘经典教材)
2. Charu C. Aggarwal, *Data Mining: The Textbook*, Springer, 2015. (理论与算法综合参考)
3. Cynthia Dwork and Aaron Roth, *The Algorithmic Foundations of Differential Privacy*, Founda-

tions and Trends in Theoretical Computer Science, 2014. (差分隐私权威参考)
4. Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar, *Introduction to Data Mining*, 2nd ed., Pearson, 2018. (入门教材)
5. Charu C. Aggarwal, *Outlier Analysis*, 2nd ed., Springer, 2017. (异常检测专著)
6. Charu C. Aggarwal, *Recommender Systems: The Textbook*, Springer, 2016. (推荐系统专著)
7. Charu C. Aggarwal, *Graph Data Management and Mining*, Springer, 2010. (图数据挖掘)
8. Charu C. Aggarwal and Philip S. Yu, *Privacy-Preserving Data Mining*, Springer, 2008. (隐私保护挖掘)
9. Judea Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, 2009. (因果推断权威著作)
10. Bernhard Schölkopf and Alexander J. Smola, *Learning with Kernels*, MIT Press, 2002. (核方法经典专著)
11. David Easley and Jon Kleinberg, *Networks, Crowds, and Markets*, Cambridge University Press, 2010. (图与网络分析)
12. Mohammed J. Zaki and Wagner Meira Jr., *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, 2nd ed., Cambridge University Press, 2020. (综合参考)