数据挖掘组2021年培训须知

总体概括

2021年暑假集训营开始时间为7月13号，持续时间为41天，划分为7周。每周上交周记，周记字数不少于1000字，每两周上交学习笔记和代码。每周进行小组会议，及时跟进新生进度。

本次训练营学习集中在掌握数据挖掘的理论知识，机器学习的算法原理，明晰差分隐私和多智能体的前沿研究，为未来进行学术研究，参加比赛打好基础。

培养目标

本次训练营的目标是学习数据挖掘的理论知识，了解和应用机器学习算法的基本原理，具备简单的后端开发能力。通过实际的项目锻炼，提高实习生的编程能力，文档能力和学习能力。

同时培养实习生具备学术研究的基本素质，具备查找和阅读论文的能力，了解多智能体和差分隐私的发展脉络和研究热点，为未来学术研究做准备。

培训安排

培训时间：

早上 8: 30 ~ 11: 30
下午 2: 30 ~ 5: 30
晚上 7: 30 ~ 10: 30

打卡时间分别为5: 00到8: 30，1: 30到2: 30，6: 30到7: 30

具体安排

培训时间表（暂定，后续可能有变动）：

时间段	学习内容	要求	备注
7.13-7.14	基础数学知识，模型的评价和选择 线性模型：线性回归，逻辑回归，多分类学习，类别不平衡问题。 学术：了解谷歌搜索，WebofScience等知名数据库，学习如何进行文献的检索。了解SCI，CCF分区，影响力因子等学术指标。	撰写学习笔记，了解基础的线性代数和概率论知识，为理解算法原理打下基础。了解模型的评价和选择 使用糖尿病数据集，代码实现线性回归，逻辑回归，掌握线性回归的多种方法，了解类别不平衡问题的解决方案。	重点掌握矩阵求导，极大似然估计等内容。
7.15-7.18	三种决策树，随机森林 朴素贝叶斯算法 ,EM算法 学术：文献检索实践	撰写学习笔记，了解决策树的基本原理。使用西瓜数据集，代码实现三种决策树，了解两种剪枝方式并应用。 在已经实现决策树的基础上，构建随机森林模型 学习笔记，了解贝叶斯算法的原理，使用垃圾邮件数据集，代码实现朴素贝叶斯算法，了解EM算法。 以多智能体一致性或差分隐私为主题，搜集近五年的重要文献。	
7.19-7.22	神经网络 支持向量机，聚类算法 学术：研读搜集的综述	了解神经网络的基本概念，了解后向传播算法，使用手写数字集，构建简单的BP神经网络进行识别。 学习笔记，了解支持向量机的原理，使用提供的数据集进行代码实现。了解常用的聚类算法，使用提供的数据集进行代码实现。 研读搜集的综述，对多智能体和差分隐私领域进行初步的了解，并总结学术研究热点	学有余力可以对深度学习进行简单入门。
7.23-7.25	Flask框架的学习 学术：精读分配的论文	学习Flask框架的基本内容，学习如何在云服务器上部署项目。可以尝试搭建自己的博客 精读分配的论文，并撰写阅读报告	
7.26-8.1	中期考核	暂定	
8.1-8.5	降维算法，集成学习 学术：论文的仿真	撰写学习笔记，了解降维算法和集成学习的原理，并进行代码实现 分配不同交叉领域的论文，认真精读并进行代码仿真，要求达到原论文水平。	要求代码具备可拓展性，考虑时空复杂度。
8.5-8.22	最终考核	暂定	

注:

1. 在暑假学习的过程中，任何作业和文档均要求使用 git 保存到代码仓库。
2. 每两天上交学习笔记和代码，原则上笔记不少于800字
3. 每周日上交周记，周结字数不得少于1000个字。
4. 每周一晚上开小组会议，每位实习生对所做工作进行汇报，并解答实习生疑问，进行技术交流。

拓展要求

学有余力的同学，可以自行安排学习内容。