简历

姓　　名	于晗丹	**出生年月**	1999.01
民　　族	汉	**籍　　贯**	浙江绍兴
电　　话	17857312377	**政治面貌**	中共党员
学　　历	硕士	**邮箱**	2300936182@qq.com
在读学校	墨尔本大学	**专　　业**	Information Technology (AI)
个人主页	https://handanyu.github.io/		

教育背景

2021.07-2023.07　　墨尔本大学　　Master of Information Technology(Artificial Intelligence)

主修专业课：机器学习（score: H1）、NLP（score: H1）、AI Planning 等

Projects:

1. 基于 twitter 的情感分析，并参与 Kaggle Competition (71/343)

2. Rumour Detection on Twitter，并参与 Kaggle Competition (17/163)

2017.09-2021.07　　杭州师范大学　　信息与计算科学（本科）

主修专业课：数据结构，数据挖掘，数据分析，大数据技术与应用，数据库应用、机器学习等 (final GPA: 4.28)

主要荣誉奖项：全国数学建模（浙江省）二等奖，浙江省高数竞赛（数学组）二等奖，美国数学建模 H 奖，浙江省政府奖学金，浙江省优秀毕业生，校三好学生等。

毕业论文课题：《EM 算法在聚类中的研究及应用》，主要通过利用半监督学习以及最大熵原理改进 EM 算法，减小了 EM 算法对初始值敏感性同时提高了聚类效果。

实习经历

2022.02-2022-06　　杭州德睿智药有限公司　　算法工程师

- 使用 PyTorch 对最新分子生成相关论文进行复现，对比各模型在训练集和测试集上的指标(e.g., ROCAUC)表现；
- 复现在 GNN 中基于梯度的对抗性扰动的节点特征增强方法，将其运用于药物分子分类，准确率提高了 2%左右；
- 使用 PyTorch 框架复现以 Tensorflow 为框架的开放项目。

2021.04-2021.07　　杭州观远数据有限公司　　算法工程师

- 利用 pandas 进行数据分析挖掘，精度误差分析，并撰写误差分析报告；
- 利用 K-Means 算法对大小客户进行聚类分析；
- 编写 MySQL 脚本实现看板设计；
- 按照 pipline 模式重构规整代码；
- 使用 git 对项目版本进行管理。

2020.12-2021.04　　杭州海康机器人技术有限公司　　数据标定员

- 协助深度学习小组，收集训练集，训练测试目标检测算法模型，并进行数据校验；
- 使用 Halcon 以及 Python 内部库 pyzbar 编写脚本代码自动对物流面单二维码以及条形码进行标注；
- 利用 Python 的自动化工具 pywinauto，pyautoGUI 编写自动化脚本，使 VM 软件实现自动化训练模型，缩短了约 10%的训练时间。

2021.09-2021.12　　杭州博世电动工具中国有限公司　　数据库开发

- 利用 pandas 进行数据处理；
- 基于 python 利用 pySimpleGUI 制作项目管理绩效软件并结合 Streamlit 进行实时数据可视化：主要负责 UI 设计以及功能模块实现（功能主要包括：项目、人员的增删改查）。通过该软件准确及时收录项目信息，提高部门项目管理效率。

校内项目经历

2022.05　　Kaggle Competition　　组长

项目名称：Rumour Detection on Twitter

- 编写 shell 脚本，运用 Python 中 tweepy 库根据 tweet ids 抓取 tweets；
- 运用 NLTK 库以及正则表达式对 tweets 中的文本数据进行预处理，具体包括移除停用词、词干提取、去除标点等；
- 对 tweets 进行文本特征提取以及数值特征提取，其中文本特征包括句子长度、词频统计、情感分析等，数值特征包括点赞数、转发数、回帖数、用户发帖率等。同时利用 LightGBM 进行特征筛选；
- 通过在不同特征数据集上使用预训练模型 BERTweet 获取上下文信息，并构建不同二分类器包括传统机器学习分类算法模型(i.e., LR, Naive Bayes)以及以 ReLU 作为中间层的激活函数，Sigmoid 作为输出层的激活函数的全连接神经网络对 BERTweet 进行微调；并使用 development dataset 上的 F1 score 以及 AUC 对各模型进行评估；
- 最终对于测试集的预测，运用基于投票思想的集成方法，将之前评估得到的 top3 分类模型得到的预测结果进行投票表决，每个测试样本取投票最多的那个作为该样本最终的预测值，并最终得到在测试集上的 F1 score 为 90%，Kaggle 的排名为 17/163。

2021.11　　Kaggle Competition　　个人

项目名称：基于 twitter 的情感分析

- 分别在 TFIDF 和基于最大熵的 TFIDF 上构建多分类模型，具体包括 Naive Bayes, Softmax Regression 和 KNN。其中运用网格搜索法进行超参数选择。并通过在训练集上的 F1 score 来对模型进行评估。最终基于最大熵的 TFIDF 的 Softmax Regression 分类模型在训练集上的 F1 score 为 0.758，在测试集上的 F1 score 为 0.756。且在 Kaggle 中的排名为 71/343；
- 并通过分析 gender bias 发现基于 Naive Bayes 的多分类模型最容易产生 gender bias。

2020.01　　美国数学建模　　组长

项目名称：基于 LSTM 的苏格兰鱼群迁移问题研究

- 通过影响鱼群迁移的因素进行机理分析，主要任务是对苏格兰海域浅表面海水温度进行预测；
- 建立 LSTM 模型，其中特征是苏格兰海域浅表面海水温度，并选择 Adam 作为优化器。
- 协同小组成员完成论文一篇并发表于《应用数学进展》。

2019.08-2019.11　　统计调查项目　　重要成员

项目名称：“理性公益，科技向善”——“互联网+公益”模式对公益事业的影响及对市民满意度调查（以杭州市为例）

- 问卷数据处理（包括数据清洗，数据归一化等）；
- 采用多值 Logistic 回归分析、列联表分析、关联规则挖掘等方法进行数据分析；
- 采用因子分析以及层次分析法构建满意度评价体系；
- 协同小组成员完成调查报告一篇，并最终整理成论文发表于《统计学与应用》。

自我评价

- 熟练使用 Python 进行文本文档处理，以及使用 numpy, pandas, scipy 等包进行数据分析和挖掘，利用 matplotlib 库进行可视化；并掌握 Python 数据结构。
- 熟练使用 shell 编写脚本语言，掌握使用 git 管理项目；
- 熟练使用 MySQL 对数据进行增删查改；
- 掌握机器学习知识（主要掌握分类、回归和聚类模型）；
- 掌握模型评估指标（如 Recall, ROC curve, mAP 等);
- 掌握损失函数（如交叉熵，KL 散度，平方损失等）；
- 熟悉 LSTM，GAN，AutoEncoder 等深度学习模型；
- 了解 PyTorch 框架，能自行使用 PyTorch 框架搭建基本深度学习模型；
- 掌握常用的数据分析、挖掘方法以及相关算法；
- 了解 Hadoop, Spark 的运行机制。
- 自我约束能力强，自学能力强，对新知识有很强对好奇心和渴求欲。