

同濟大學

實習日記本

学 院

计算机科学与技术学院

专 业

计算机科学与技术

学 号

2253214

学生姓名

李闯

接受单位

上海和今信息科技有限公司

地 址

上海市徐汇区虹桥路 333 号 1 幢 206 室

指导教师姓名

郭玉臣

大学生校外实习守则

一、高校校外实习是高等教育同社会主义建设相结合的重要环节，是培养合格人才的重要保证，每个学生必须按照教学计划的要求，认真参加各种形式的校外实习活动，并自觉地遵守本守则。

二、严格遵守实习单位的有关规章制度。如有违反，学校应会同实习单位按有关规定严肃处理。

三、自觉遵守劳动纪律、尊重实习带教人员，服从安排，严格执行安全操作规程。

四、妥善保管因实习需要而借阅的图纸、资料，严格遵守保密制度，放置泄密。

五、认真写好实习日记，客观记录实习的内容、问题及心得。

六、独立完成实习报告，系统总结、综合分析实习中的问题以及解决问题的方法和建议。

七、实习结束时，必须如数归还所借用的劳防用品、工具、仪器、资料及生活用品，如有损坏或遗失应按规定赔偿。

八、学生在实习期间造成伤残和死亡的，可按“大学生团体平安保险”有关规定给予赔偿。

实习起迄日期

自 2025 年 7 月 3 日起至 2025 年 8 月 18 日止

实 习 内 容

2025 年 7 月 3 日（星期四）

今天是我参加和鲸社区大模型夏令营的第一天。我仔细浏览了课程安排和资源库。平台设计得很清晰，分为知识教程、代码案例和实战项目三个板块。

作为夏令营的第一天，课程从 Python 基础语法开始。虽然我之前已经接触过编程，但为了夯实基础，还是决定从头开始系统学习。我上午完成了前三个关卡的练习，内容包括变量定义、数据类型、流程控制等基础语法。平台采用循序渐进的闯关模式，每个关卡都配有详细的讲解和即时练习。虽然内容相对简单，但课程设计得很用心；下午着重练习了列表、字典等数据结构的使用技巧，完成了一些综合性稍强的编程小任务。平台提供的实战题目都很贴近实际应用场景，比如用字典处理学生成绩、用集合去重等。这种理论结合实践的方式，让基础知识的掌握更加牢固。

通过第一天的学习，我有两个主要收获：首先是对 Python 的基础知识进行了系统的梳理和巩固；更重要的是完全熟悉了这个学习平台的操作逻辑和课程体系。平台的即时反馈机制和循序渐进的课程设计，让我对接下来的学习充满信心。

2025 年 7 月 6 日（星期日）

今天完成了 NumPy 模块的全部五个关卡学习。从基础的数组创建开始，逐步深入掌握了 NumPy 的核心功能。

上午完成了前两个关卡的学习。在数组创建与计算部分，重点练习了使用 `np.array()` 创建数组和各种数学运算操作。第二关的数组聚合与广播机制让我印象深刻，特别是理解不同形状数组如何进行运算的规则。

下午集中攻克了后三个关卡。数组比较与索引部分让我熟悉了布尔索引和花式索引的使用技巧。在数组排序关卡，学习了 `argsort()` 等排序方法的灵活应用。最后的综合实战将前面所学串联起来，完成了一个完整的数据处理流程。

通过今天的系统学习，我对 NumPy 的高效数组操作有了更深入的理解。这种循序渐进、理论与实践结合的课程设计，帮助我快速掌握

了 NumPy 的核心功能。虽然内容有一定难度，但通过反复练习和闯关模式，最终顺利完成了所有任务。

2025 年 7 月 8 日（星期二）

今天完成了 Pandas 模块的系统学习，收获颇丰。上午从基础的数据读取开始，通过关卡 1 掌握了 `read_csv()`、`read_excel()` 等常用数据读取方法。在处理一个格式不太规范的 CSV 文件时，我遇到了编码问题，通过调整 `encoding` 参数最终成功读取，这让我意识到真实数据往往需要额外处理。

下午重点攻克了数据索引和数据清洗两个关卡。在索引部分，`loc` 和 `iloc` 的差异让我印象深刻，特别是布尔索引的灵活运用大大提升了数据筛选效率。数据清洗关卡最具挑战性，我练习了处理缺失值的多种方法，包括 `fillna()` 和 `dropna()`，并通过一个电商数据集实战演练了异常值处理的全流程。

通过今天的学习，我不仅掌握了 Pandas 的核心操作，更重要的是培养了处理真实数据问题的思维方式。期待在后续项目中继续深化这些技能。

2025 年 7 月 12 日（星期五）

今天完成了数据可视化模块的系统学习，主要围绕 Matplotlib 和 Seaborn 两大核心库展开。课程分为六个循序渐进的关卡，每个关卡都设计了丰富的实践任务。

在 Matplotlib 部分的学习中，我从最基础的柱形图、折线图绘制开始（关卡 1），逐步深入到图表元素的精细配置（关卡 2）。特别是在设置中文显示时，通过添加 `font.sans-serif` 参数解决了中文乱码问题。关卡 3 的多图布局和双 Y 轴绘制让我印象深刻，使用 `GridSpec` 实现复杂子图排列的过程虽然有些挑战，但最终效果令人满意。

下午转入 Seaborn 的学习，这个基于 Matplotlib 的高级封装库让统计图表绘制变得异常简单。在关卡 4 中，我仅用几行代码就绘制出了专业的分布图和小提琴图。关卡 5 的回归模型可视化让我直观地理解了数据间的相关性。最后的关卡 6 中，`FacetGrid` 的多维度分析功能尤其出色，通过分层绘制清晰地展现了数据中的潜在模式。

通过今天的实践，我不仅掌握了数据可视化的基本技能，更重要的是学会了如何选择合适的图表类型来传达数据背后的故事。平台提供的泰坦尼克号等经典数据集，让每个知识点的学习都变得生动具体。这种从基础到高级的系统训练，使我对 Python 可视化工具的应用能力得到了全面提升。

2025 年 7 月 15 日（星期二）

今天完成了机器学习模块的系统学习，通过六个实战关卡深入理解了不同算法的应用场景。课程设计非常注重理论与实践的结合，每个关卡都配备了真实的数据集和具体的业务问题。

上午从基础的逻辑回归开始（关卡 1），在乳腺癌诊断案例中，我特别注意到了特征标准化的重要性。当数据未经过标准化处理时，模型的收敛速度明显变慢。通过 StandardScaler 进行预处理后，准确率提升了约 8%。接着在二手车价格预测的回归问题中，学习了如何评估回归模型的性能指标，如 MAE 和 R-squared。

下午的 KMeans 聚类（关卡 2）和决策树（关卡 3）学习让我印象深刻。在股票分类任务中，通过肘部法则确定最佳聚类数 $k=5$ ，这个结果与行业分类标准高度吻合。而在 iris 分类任务中，使用 Graphviz 可视化决策树的过程非常有趣，直观地展示了模型是如何做出分类决策的。

最后两个关卡涉及更复杂的算法。SVM（关卡 4）在处理肥胖风险数据时展现了强大的分类能力，而 XGBoost（关卡 5）在葡萄酒分类任务中的表现尤其出色。通过调整 `n_estimators` 和 `max_depth` 等参数，最终将准确率提升到了 96%。

综合实战（关卡 6）是最具挑战性的部分。在堪培拉天气预测任务中，我尝试了多种特征工程方法，包括时间特征提取和异常值处理。参考 baseline 后，通过引入滑动窗口特征，最终将预测准确率提高了 12%。这个实战项目让我深刻体会到，在实际应用中，数据预处理和特征工程往往比算法选择更重要。

通过今天的学习，我不仅掌握了常见机器学习算法的使用方法，更重要的是培养了解决实际问题的完整思路。从数据探索、特征工程到模型调优，每个环节都需要仔细思考和反复验证。这种系统化的训练为后续更复杂的 AI 学习打下了坚实基础。

2025 年 7 月 16 日（星期三）

今天完成了深度学习基础课程的系统学习，通过四个循序渐进的关卡，我对深度学习的核心概念和关键技术有了全面的理解。

上午从基础理论开始，重点复习了线性代数和概率论中的重要概念。在 PCA 降维实践中，我直观地理解了特征值分解在数据压缩中的应用。随后通过计算图模型的练习，加深了对信息流传递机制的认识。

下午的神经网络部分收获最大。在 MNIST 数据集上对比了不同激活函数和优化算法的表现，Sigmoid 和 ReLU 的差异尤为明显。通过调整学习率和正则化参数，我成功将测试准确率提升到了 98.3%。

CNN 和 RNN 的实践环节让我印象深刻。在图像卷积运算中，通过

可视化不同卷积核的效果，直观理解了特征提取的过程。LSTM 单元的实现则让我对序列建模有了更深的认识。

最后的框架对比环节非常实用。通过相同的图像分类任务在 TensorFlow 和 PyTorch 中的实现，我清晰体会到了两个框架的设计哲学和编码风格的差异。PyTorch 的动态图机制让调试过程更加直观。

特别感谢和鲸社区精心设计的课程体系，从理论到实践的完整闭环让我对深度学习有了系统性的掌握。这种扎实的基础训练，将为后续更复杂的模型学习打下坚实基础。

2025 年 7 月 18 日（星期五）

今天完成了 HuggingFace Transformers 实战课程的学习，收获远超预期。课程内容之丰富、实践性之强，让我对 Transformer 模型有了全新的认识。

通过系统化的实战训练，我不仅掌握了 HuggingFace 生态的核心使用方法，更重要的是培养了大模型应用的工程思维。每个实践环节都让我对 NLP 任务的处理流程有了更深刻的理解，从数据预处理到模型微调，再到效果评估，这一整套方法论让我受益匪浅。

课程设计的答题环节非常巧妙，通过概念性问题帮助我及时巩固关键知识点。这些问题直指模型的核心原理和应用要点，促使我不断反思实践过程中的每个技术细节。这种“动手实践+理论强化”的学习模式，效果出奇地好。

这次学习经历让我深刻体会到，在 AI 领域，理论认知和工程实践必须相辅相成。期待将所学应用到更多实际场景中，继续深化对大模型的理解和应用能力。

2025 年 8 月 2 日（星期六）

今天参与了和鲸社区举办的“大模型小语种专业领域 LoRA 微调 Workshop”，这是一次极具深度的技术实践体验。活动围绕 OpenDataLab“万卷·丝路 2.0”多语言语料库展开，让我对 LoRA 这一前沿技术有了全新的认识。

在实践环节中，我深刻体会到 LoRA 技术的高效与灵活。通过低秩矩阵分解的方式，仅需调整极少量参数就能显著提升模型在小语种任务上的表现。最令我惊讶的是，在保持原始模型架构不变的情况下，通过精心设计的参数配置，我们成功让模型掌握了阿拉伯语的专业术语和语言特征。

实验过程中，我特别关注了不同参数配置对模型性能的影响。通过对比 $r=32$ 和 $r=64$ 等不同 rank 值的表现，直观理解了低秩适应的核心思想。多维度评估体系的构建也让我大开眼界，从传统的

BLEU/ROUGE 指标到专业的术语覆盖率分析，这套评估方法为模型优化提供了全方位的指导。

特别要感谢和鲸社区提供的免费且强大的算力资源支持，让我能够顺利完成这次实践。在训练过程中，高性能的计算资源确保了模型微调的效率，使得复杂的参数调优实验成为可能。这种开放共享的教育理念，为学习者创造了极佳的技术实践环境。

这次 Workshop 不仅让我掌握了 LoRA 微调的技术细节，更重要的是培养了我解决领域适配问题的系统性思维。从数据准备、参数配置到效果评估，每个环节都需要精心设计和反复验证。这种严谨的工程实践态度，将对我未来的 AI 学习产生深远影响。

2025 年 8 月 5 日（星期一）

今天参与了和鲸社区举办的"基于大语言模型知识推理能力的蒸馏框架"Workshop，这是一次极具启发性的技术探索之旅。活动聚焦如何将大语言模型的推理能力高效迁移到轻量化模型，为解决实际应用中的算力瓶颈提供了创新思路。

在实践环节中，我深刻体会到知识蒸馏技术的精妙之处。通过构建教师-学生模型架构，我们成功将大型教师模型的复杂推理能力"提炼"到更轻量的学生模型中。最令人惊喜的是，在医学考试领域的迁移实验中，学生模型不仅保持了与教师模型相当的准确率(75.4%)，还大幅降低了推理所需的计算资源。

实验过程中，我特别关注了跨领域迁移的挑战。医学考试数据与科学考试数据在格式和表述上的差异，促使我们重构了数据预处理流程。通过设计兼容不同数据格式的解析逻辑，我们成功实现了知识蒸馏框架的灵活适配。这种解决实际问题的过程，让我对技术落地的复杂性有了更深刻的认识。

特别要感谢和鲸社区提供的完善实验环境和专业指导，让我能够顺利完成这次知识蒸馏的实践探索。平台提供的预配置环境和详实的实验文档，极大地降低了技术门槛，使我能专注于算法原理的理解和实现。

这次 Workshop 不仅让我掌握了知识蒸馏的核心技术，更重要的是培养了我解决实际工程问题的能力。从模型架构设计、数据处理到效果评估，每个环节都蕴含着丰富的工程智慧。这种理论与实践相结合的学习体验，将对我未来的技术研究和应用开发产生深远影响。

2025 年 8 月 10 日（星期四）

今天参与了和鲸社区举办的"RAG 技术在电商智能问答中的应用"Workshop，这是一次极具实用价值的技术实践。活动聚焦如何通过

检索增强生成技术优化电商场景下的问答系统，让我对这一前沿技术有了更深入的理解。

在实践环节中，我深刻体会到 RAG 技术的强大之处。通过构建专业的知识库，我们成功解决了大模型在电商领域容易产生"幻觉"回答的问题。最令人惊喜的是，在对比实验中，Qwen3-0.6B 模型在保持较高准确性的同时，响应速度比 4B 模型快了 19%，这充分证明了"合适即最优"的模型选择原则。

实验过程中，我特别关注了不同向量索引类型的性能差异。通过对比 FAISS 提供的 FlatL2、IVFFlat、IVFPQ 和 HNSW 四种索引，发现 IVFPQ 在检索速度和存储空间上展现了最佳平衡。这种工程实践中的性能调优经验，对构建实际应用系统具有重要价值。

特别要感谢和鲸社区提供的完善实验环境和详实的案例资料，让我能够深入理解 RAG 技术的核心要点。平台精心设计的电商场景案例，包括退货政策、商品咨询等典型问题，使技术学习与实际业务紧密结合。

这次 Workshop 不仅让我掌握了 RAG 技术的实现细节，更重要的是培养了我从业务需求出发设计技术方案的能力。从知识库构建、语义检索到回答生成，每个环节都需要针对具体场景进行优化。这种理论与实践相结合的学习体验，将对我未来开发智能问答系统产生深远影响。

指导教师批阅意见：

指导教师签名

月 日