

小组展示参考选题

1. 小型问答专家：

背景：

随着科技的飞速发展，海量的科学文献、技术报告和研究论文不断涌现，研究人员和工程师们常常面临信息过载的问题。如何在短时间内从大量文献中提取出有价值的信息，成为了一个重要的挑战。在这样的背景下，基于 Retrieval-Augmented Generation (RAG) 的问答系统应运而生。RAG 结合了检索式和生成式模型的优势，可以在知识密集型领域中，迅速从海量文献中提取出相关信息，并以自然语言生成回答，为用户提供精准、高效的知识服务。

任务描述：

基于 ChatGLM 和 Langchain 构建一个本地的 RAG 问答系统这个系统应能在用户给定问题在本地的文献中自行搜索信息，并且利用相关信息进行专业问题的回答，具体的一些指引如下：

1. 研究领域与文献收集

在感兴趣的研究领域，研究人员首先需要自行下载不少于 50 篇相关的学术文献。

2. 本地知识库的构建

在文献收集完成后，接下来要做的是将这些文献转化为一个结构化的知识库。可以采用以下步骤：

- **文本预处理：**使用 NLP 工具对文献文本进行清理和预处理，如 PDF 识别、去除格式符号、表格、图像注释等，提取出纯文本内容。
- **文献分块：**将每篇文献分成若干小段落，通常以句子或段落或固定长度为单位。这样可以使后续的检索过程更加精准，同时减少生成模型的上下文长度压力。详情可以见 Langchain 教程。
- **知识库构建：**利用工具如 LangChain 将预处理后的文本构建为一个可检索的本地知识库。LangChain 允许创建与文本语义相关的索引，便于快速检索相关内容。

3. 基于 RAG 的小型问答专家系统

RAG 模型结合了检索模型和生成模型的优点，是构建智能问答系统的强大工具。该系统通常包括两个核心组件：

- **检索模型：**检索模型负责从本地知识库中找到与用户查询相关的文档片段。可以使用向量化模型来对文档进行向量化，并利用余弦相似度或其他相似性度量进行检索。
- **生成模型：**生成模型如 ChatGLM 或 GPT，可以根据检索到的文档片段生成具体的答案。你需要将检索得到的相关资料使用合适的 Prompt 融合到用户提问的问题中，进行最后的答案生成。

4. 系统优化

- **查询优化:** 通过调整检索模型的参数, 如查询扩展、top-k 检索数量等, 优化检索结果的质量。同时, 可以调试生成模型的温度、最大生成长度等参数, 以控制生成回答的质量和风格。将这些实验的结果进行总结汇报, 给出一套合适的配置

5. 系统展示:

对系统进行一些专业问题的问答并在 PPT 展示结果, 并且与 AI 助教直接问答进行对比, 看看搭建的专家问答系统是否更好用?

参考链接:

1. GLM 首页: <https://github.com/THUDM/GLM-4>
2. Langchain 的 RAG 搭建教程: <https://python.langchain.com/v0.2/docs/tutorials/rag/>

2.化工文献中的材料命名实体识别(NER)与信息抽取(IE)

背景介绍

在化工研究和工程应用中, 学术文献和专利中包含了大量关于材料、化学物质、反应条件、工艺流程等关键信息。这些信息是科研人员进行创新、工艺改进和技术转移的重要资源。然而, 随着文献数量的激增, 人工提取和整理这些信息变得越来越困难且耗时。命名实体识别(NER)和信息抽取(IE)技术为解决这一问题提供了有力的工具。

NER 的任务是从非结构化文本中自动识别出特定类别的实体, 例如材料名称、化学品、反应物、产品等。而信息抽取则进一步将这些实体与其相关属性或关系提取出来, 例如材料的物理化学性质、用途、反应条件等。通过这些技术, 可以自动化地将大量的文献和专利信息转化为结构化数据, 进而构建化工领域的知识库或数据库, 为科研和工业应用提供支持。

任务要求

1. 数据收集与准备

- **文献收集:** 从公开的化工期刊、数据库或专利网站中自行搜集与化工领域相关的学术文献或专利文献。建议选择包括材料科学、催化剂研究、化学工艺等领域的文献。
- **数据清洗:** 对收集到的文献文本进行清洗处理, 去除无关字符、格式符号等, 确保文本数据适合后续的自然语言处理任务。

2. 实体识别与信息抽取

- **NER/IE 模型选择：**使用一些现成的工具，如 OpenChemIE，ChemDataExtractor、甚至是大语言模型等进行文档信息的提取，体会 python 在这类工具中的使用，对化工文献进行命名实体识别，提取出与材料相关的实体。这些实体可能包括材料名称（如“钛酸钙”）、化学品（如“氯化钠”）、反应物（如“乙醇”）、产品（如“聚乙烯”）等。
- **信息抽取：**在 NER 的基础上，进一步提取材料的相关属性和关系信息。例如，提取文献中材料的物理化学性质（如“熔点”、“带隙”）、用途（如“催化剂”）和实验条件（如“反应温度”）。
- **标注与分类：**对提取出的实体和信息进行分类和标注，构建结构化的数据表或知识图谱，展示材料的特性、用途及其相互关系。

3. NER 工具比较与分析

- **工具比较：**使用多个 NER 工具，对同一批文献数据进行处理，比较不同工具的效果。
- **性能评估：**通过准确率、召回率、F1-score 等指标评估不同 NER 工具的性能，分析它们在化工领域应用中的优劣势。
- **错误分析：**分析 NER 和 IE 任务中常见的错误类型（如漏检、误检），并讨论可能的改进措施。

4. 可视化与报告撰写

- **结果可视化：**将实体识别和信息抽取的结果进行可视化展示。例如，使用知识图谱展示材料与其属性、用途、反应条件的关系，或使用表格、图表展示不同 NER 工具的性能比较结果。
- **报告撰写：**撰写一份详细的作业报告，报告应包括以下内容：
 - **引言：**介绍课题背景、研究意义和任务目标。
 - **数据收集与预处理：**描述数据的来源、获取过程和预处理步骤。
 - **方法与实现：**详细说明 NER 和 IE 任务的实现过程，介绍所使用的工具和模型。
 - **结果与讨论：**展示并分析任务结果，讨论不同工具的性能差异及其在化工领域的应用潜力。
 - **结论与展望：**总结工作成果，并提出可能的改进方向和进一步研究的建议。

3. 化工事故与安全问题的文本分析

背景介绍

化工行业因其特殊的工艺过程和使用的化学品种类繁多，常常面临着各种安全风险。化工事故一旦发生，不仅可能导致财产损失，还可能对环境和人类健康造成严重的影响。因此，化工安全问题一直是行业内外关注的焦点。通过对化

工事故相关的文本数据进行分析，能够帮助识别常见的事故原因、了解事故的影响范围、总结有效的预防措施，从而提升化工企业的安全管理水平。

在信息化时代，化工事故的信息传播速度快，公众和媒体的关注度高。分析事故报告、新闻文章和社交媒体中的相关信息，不仅可以揭示事故的主要因素，还能帮助理解公众的情绪反应，这对危机管理和舆情控制具有重要意义。

任务要求

1. 数据收集与准备

- **数据收集：**从公开渠道收集与化工事故相关的文本数据，如政府或企业发布的事故报告、新闻媒体的报道、社交媒体上的公众评论等。数据类型可以包括正式的文档、文章、评论等。
- **数据清洗：**对收集到的文本数据进行清洗和预处理，去除噪音数据（如广告、无关信息等）。

2. 文本分析

- **分词与词频统计：**使用自然语言处理工具（如 Jieba、NLTK）对文本进行分词处理，并统计每个词出现的频率。通过词频分析，可以识别出化工事故中常见的关键词（如“泄漏”、“爆炸”、“过热”等）。
- **关键词提取：**结合 TF-IDF 或其他关键词提取方法，提取事故文本中的核心关键词，并分析其在不同类型事故中的分布情况。
- **词云图可视化：**使用 wordcloud 库生成词云图，直观展示事故中高频出现的关键词，并分析不同事故类型（如火灾、爆炸、泄漏）的关键词差异。

3. 情感分析（可选）

- **情感分析：**对社交媒体或新闻报道中的文本进行情感分析，分类为积极、消极或中立情绪。可以使用现有的情感分析工具包（如 TextBlob、VADER、SnowNLP）进行初步分析。
- **舆情分析：**结合情感分析结果，探讨公众对化工事故的态度，并分析事故发生后的舆情变化，提出如何通过有效的信息传播减少恐慌和误解的策略。

4. 结果展示与报告撰写

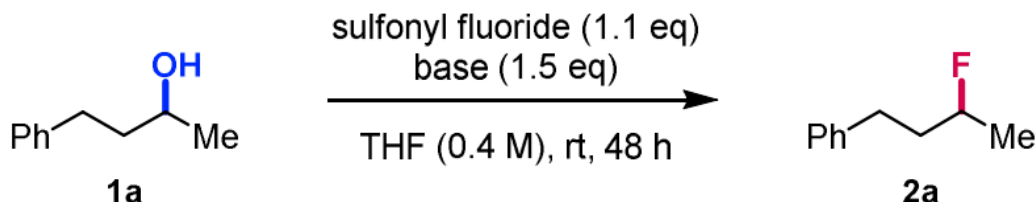
- **结果可视化：**将词频分析、关键词提取和情感分析的结果进行可视化展示。通过词云图、条形图、时间趋势图等方式，直观展示事故中常见词汇、情感趋势和关键词差异。

- **报告撰写：**撰写一份详细的分析报告，报告内容应包括：
- **引言：**介绍课题背景、研究意义和任务目标。
- **数据收集与预处理：**描述数据的来源、获取过程和预处理步骤。
- **文本分析方法：**详细说明分词、词频统计、关键词提取、情感分析等任务的实现过程，介绍所使用的工具和方法。
- **结果与讨论：**展示并分析任务结果，讨论不同事故类型的关键词特点、公众情感反应及其对安全管理的启示。
- **结论与建议：**总结工作成果，并提出如何改进化工安全管理和舆情控制的建议。

4. 有机反应数据分析

背景：

在本作业中，学生将通过 Open Reaction Database 探索化学反应数据，并使用 Python 编程和机器学习工具对化学反应进行分析和预测。此作业分为两个主要部分，学生可以选择其中一项或多项进行探索。



任务描述：

任务 1：使用 RDKit 分析化学反应数据

1. 数据获取：

- 从 Open Reaction Database 中下载化学反应数据。数据通常以 SMILES 格式表示分子结构，包含反应物、产物、试剂、溶剂等信息。具体的信息参考 <https://docs.open-reaction-database.org/en/latest/schema.html>

2. 数据处理：

- 使用 RDKit 工具将 SMILES 格式的化学物质转化为可用于数据分析的格式，如分子指纹 (Molecular Fingerprints)。
- 从分子指纹开始，探索不同反应条件 (如试剂、溶剂、反应温度、时间等) 与反应产率之间的关系。

3. 关系发掘：

- 通过数据可视化工具 (如 Matplotlib 或 Seaborn) 绘制图表，直观展示反

应条件和产率之间的关系。

- 尝试使用简单的统计方法（如相关分析）初步发掘这些关系。

4. 模型预测：

- 可以选择现有的机器学习模型（如线性回归、随机森林、支持向量机等）对反应产率进行预测。
- 可尝试调用 Scikit-learn 等 Python 库来训练和测试模型，探索如何根据输入的 SMILES 反应式和条件预测产率。

任务 2：基于 SMILES 的试剂与溶剂推荐

1. 数据处理：

- 获取 Open Reaction Database 中的反应条目，并将其转换为分子指纹或其他特征向量。
- 分离反应物、产物、试剂、溶剂的数据，为模型训练做准备。

2. 模型复现与训练：

- 复现或自行实现文献中的机器学习模型，用于试剂和溶剂的推荐。
- 参考文献：ACS Cent. Sci. 2018, 4, 11, 1465–1476
- 可以使用现有的 Python 代码或自行设计模型，基于输入的 SMILES 反应式推荐最优的试剂和溶剂。

3. 模型评估：

- 使用交叉验证等方法评估模型的表现。
- 分析模型推荐结果与实际实验条件的匹配度(Top-k 文献条件复现度等)。

参考链接：

1. Open Reaction database (<https://docs.open-reaction-database.org/en/latest/>)
2. RDKit 文档 (<https://www.rdkit.org/docs/>)

5. 钙钛矿太阳能电池材料筛选任务

背景介绍

钙钛矿太阳能电池因其优异的光电转化效率和低成本的生产工艺，近年来在光伏领域得到了广泛关注。钙钛矿材料具有 ABX₃ 的晶体结构，属于具有近立方，其中 A 和 B 是阳离子，X 通常是阴离子（如氧、氮或卤素）。不同的元素组合能够显著影响材料的光学和电子性质，特别是带隙（band gap）和形成能（formation energy）。带隙是决定材料吸收光谱范围的关键参数，而形成能则反映了材料的热力学稳定性，这两者是筛选高效且稳定的太阳能电池材料的核心指标。

Materials Project 是一个开放的材料数据库，包含了数百万种已知和预测的材料的物理化学性质。通过使用 Materials Project API，可以高效地访问和检索这些材料的晶体结构及其相关的物性数据。本次任务将引导你通过编写 Python 脚本，从 Materials Project 中获取与钙钛矿结构相关的材料数据，并利用机器学习模型预测材料的带隙和凸包能量(Energy above hull)，最终筛选出适合作为太阳能电池的钙钛矿材料。

任务描述：

1. 数据获取:

使用 Materials Project API ([API 链接](https://docs.materialsproject.org/downloading-data/using-the-api)) 和教程 (<https://docs.materialsproject.org/downloading-data/using-the-api>) 从数据库中爬取钙钛矿结构的材料数据获取材料的晶体结构、带隙、形成能等相关性质。你需要注册一个账号获取 API key, 并且选择合理的筛选条件使得数据库中只包含钙钛矿。

2. 数据处理与分析:

对爬取的数据进行预处理, 包括数据清洗、特征提取和数据分割等。利用 Py matgen、Matminer 两个工具从材料的晶体结构中提取出用于机器学习模型的特征向量。

3. 机器学习建模:

构建并训练一个机器学习模型, 利用现有的材料数据预测给定钙钛矿材料的带隙和形成能。你可以选择常见的机器学习算法, 如线性回归、随机森林、支持向量机或神经网络等。

4. 材料筛选:

通过排列组合, 扩充你的输入到一个更大的空间, 并利用机器学习模型进行快速的带隙与凸包能量预测。根据模型预测的带隙和凸包能量, 筛选出具有适合带隙和较高稳定性的钙钛矿材料。合适的带隙通常在 1.1-1.6 eV 之间, 这样的材料可以高效吸收太阳光谱中的可见光部分, 而稳定性则由形成能的大小来衡量, 通常比较小的形成能说明材料是稳定的, 你可以根据形成能计算凸包能量判断是否稳定。

5. 结果分析与讨论:

对筛选出的材料进行分析, 讨论它们在实际太阳能电池应用中的潜力。可以结合文献或已有研究, 验证模型预测结果的合理性。同时可以结合机器学习探索附录, 进行一些其他的探索。

指导建议

1. 数据处理与特征提取:

- 使用 pymatgen 从晶体结构中提取几何和电子结构特征, 作为机器学习模型的输入。例如, 可以提取局部环境描述符、键长、原子种类等。

2. 机器学习建模与优化:

- 使用 scikit-learn 或其他机器学习库对数据进行训练和测试。可以尝试不同的模型和超参数组合, 以获得最佳的预测性能。
- 建议使用交叉验证来评估模型的泛化能力。
- 其他参考资料见机器学习探索附录。

3. 材料筛选与可视化:

- 筛选出带隙在 1.1-1.6 eV 之间且具有低形成能的钙钛矿材料。
- 利用数据可视化工具展示筛选结果, 并结合文献对结果进行讨论。

6. 基于预训练大模型的 HER 催化剂筛选设计

背景介绍

氢气作为清洁能源的重要载体, 在未来的能源体系中具有巨大的应用潜力。

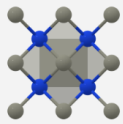
氢气的制备过程主要依赖于电解水反应，而其中的关键步骤是析氢反应（HER, Hydrogen Evolution Reaction）。在 HER 过程中，催化剂的性能决定了反应的效率。H*（吸附氢原子）的吸附能被广泛认为是评价 HER 催化剂性能的一个重要指标。一般来说，H 吸附能的绝对值越接近零，催化剂的 HER 活性越高。因此，设计和筛选具有合适 H 吸附能的催化剂对提高 HER 效率至关重要。

近年来，随着人工智能技术的发展，预训练大模型（例如 Equiformer V2）已成为预测材料性能的重要工具。这些模型利用了大量材料数据，通过深度学习技术能够有效预测诸如氢吸附能、形成能等关键材料性能参数。在本任务中，你将学习如何结合 Pymatgen 和预训练大模型（如 Equiformer V2）进行 HER 催化剂的筛选与设计。

任务目标

1. 合金结构获取：

使用 Pymatgen 工具，从 Materials Project 数据库中获取相关合金材料的晶体结构数据。你将选择几种不同的元素形成的单元素金属或合金作为潜在的 HER 催化剂候选，并且筛选出其中稳定的材料(Energy above hull \approx 0eV, 或者是属性的 predicted stable).获取的晶体结构数据将用于后续的表面切割和吸附位点生成。



ZnCu
mp-987

TABLE OF CONTENTS

[Summary](#)
[Crystal Structure](#)
[Properties](#)
[Contributed Data](#)
[Literature References](#)
[External Links](#)
[More](#)
[Related Materials](#)

Thermodynamic StabilityAqueous Stability

Thermodynamic Stability ⓘ

DataMethodsAPI

Predicted Stable	✓
Energy Above Hull	0.000 eV/atom
Predicted Formation Energy	-0.091 eV/atom
Calculation method	GGA / GGA+U / R2SCAN
Decomposition Path	Not predicted to decompose
Amorphous Limit	Not calculated

2. 低指数晶面切面与吸附位点生成：

使用 Pymatgen 对获取的合金结构进行低指数晶面(通常 h,k,l 小于 1)的切割，生成催化剂表面。对切割后的晶面进行氢吸附位点的生成。

3. 结构弛豫与形成能计算：

使用 Equiformer V2 等预训练大模型，对生成的吸附体系进行结构弛豫（geometry optimization），从而获得体系的最低能量构型。通过参考资料，将能量后处理为计算体系的吸附能，并根据氢吸附能的大小筛选出具有最佳 HER 性能的

合金催化剂。

4. 催化剂筛选与设计:

根据计算结果,筛选出具有最优 H^* 吸附能的合金催化剂,并讨论其潜在的 HER 性能。可以结合文献资料,验证计算结果的合理性,并提出进一步的优化方案或新材料设计思路。

指导建议

Pymatgen 的使用:

- 了解如何使用 Pymatgen 从 Materials Project 获取合金结构数据。你可以利用 MPRester 类来检索特定合金的晶体结构。
- 通过 Pymatgen 中的 SlabGenerator 或 adsorption 模块,进行晶面切割与吸附位点的生成。具体的内容可以参考系列教程 (<https://blog.shishiruqi.com/2019/06/08/pymatgen-adsorption/>)

Equiformer V2 的使用:

- Equiformer V2 是一种基于深度学习的预训练大模型,专门用于材料科学中的性能预测任务。为了方便的使用,请使用 fairchem 提供的模型与预训练 checkpoint 进行结构优化: <https://github.com/FAIR-Chem/fairchem>
- 使用 OC20 系列的预训练模型时,OCPCalculator 中得到的能量就是吸附能,但是使用 OC22 系列模型时可能不一样,需要自行查阅里面的文档查看是否需要处理后。

数据分析与筛选:

- 根据计算得到的 H^* 吸附能和形成能,分析不同合金材料的 HER 性能。
- 使用 Python 的 Pandas 和 Matplotlib 等库对结果进行可视化和数据分析,帮助你做出筛选决策。

进一步讨论与优化:

- 在得到筛选结果后,结合文献资料讨论这些合金催化剂的实际应用潜力。
- 你可以考虑通过掺杂等方式进一步优化催化剂性能,并设计新的实验或计算方案。

7.数学建模

背景:

大学生数学建模竞赛(CUMCM)是我国规模最大、影响最广泛的数学竞赛之一,吸引了大量热爱数学与应用科学的大学生参与。竞赛旨在培养学生运用数学方法解决实际问题的能力,提升他们的逻辑思维与团队协作精神。参赛者通常需要在规定时间内,基于一个或多个实际问题,建立数学模型、进行数据分析、提出解决方案,并撰写详细的分析报告。

任务描述

1. 竞赛题目选择与理解:

- 从提供的链接中选择一个近几年的大学生数学建模竞赛题目,或自行搜索其他著名竞赛(如美赛 MCM/ICM)的题目进行分析。

- 仔细阅读并理解题目的背景、问题陈述及要求。明确竞赛题目希望解决的问题类型，并思考可能的建模方法。
2. **数学模型建立：**
 - 根据题目要求，建立合理的数学模型。你需要考虑使用哪些数学方法和理论来描述和解决问题，例如微积分、线性代数、概率论与统计学、优化算法等。
 - 模型的建立应简洁且具备可操作性，同时能够充分反映问题的本质特征。
 3. **数据收集与处理：**
 - 如果设计到相关的内容，使用 Python 的 Pandas 等工具对数据进行清洗、整理和初步分析，确保数据符合建模要求。
 4. **模型求解与分析：**
 - 对建立的数学模型进行求解，可以使用数值方法、优化算法或计算机模拟等手段。
 - 对求解结果进行分析与讨论，评估模型的有效性及其对实际问题的解释能力。可以使用 Matplotlib 或 Seaborn 等工具进行结果可视化。
 5. **结果汇报与讨论：**
 - 按照竞赛要求，撰写详细的结果分析报告。报告应包括问题陈述、模型建立过程、数据分析、模型求解、结果讨论与结论等部分。
 - 在报告中，你需要清晰地表达建模思路和计算结果，并提出可能的改进方向或后续研究建议。

参考链接： 近几年的高教社杯数学竞赛题

http://www.mcm.edu.cn/html_cn/node/c74d72127066f510a5723a94b5323a26.html
http://www.mcm.edu.cn/html_cn/node/388239ded4b057d37b7b8e51e33fe903.html
http://www.mcm.edu.cn/html_cn/node/90d223833c1eb50f899aa096a66c6896.html
http://www.mcm.edu.cn/html_cn/node/10405905647c52abfd6377c0311632b5.html
http://www.mcm.edu.cn/html_cn/node/b0ae8510b9ec0cc0deb2266d2de19ecb.html
http://www.mcm.edu.cn/html_cn/node/7cec7725b9a0ea07b4dfd175e8042c33.html

8. 机器学习入门项目：

背景：

随着数据科学和人工智能的快速发展，机器学习已成为解决各种现实问题的强大工具。对于初学者来说，参与一些经典的入门竞赛项目不仅能够加深对机器学习基本概念的理解，还能通过实战提升数据分析和建模的技能。Kaggle 平台是全球最大的机器学习和数据科学社区，提供了丰富的竞赛项目，供学习者和专业人士挑战自我，提升技能。

在本项目中，你将参与两个经典的 Kaggle 入门竞赛项目：**Titanic - Machine Learning from Disaster**（分类问题）和 **House Prices - Advanced Regression Techniques**（回归问题）。通过对这些数据集的深入分析和建模，你将学会如何处理真实世界中的数据，并运用机器学习算法进行预测。

任务描述:

项目 1: Titanic - Machine Learning from Disaster

项目简介

泰坦尼克号的沉没是历史上最著名的海难之一。在这场灾难中，不同乘客的生存几率受到多种因素的影响，如性别、年龄、舱位等。Kaggle 的 Titanic 竞赛要求你根据乘客的个人信息，预测他们在这次灾难中的生存几率。这个项目是经典的二分类问题，非常适合作为机器学习的入门练习。

项目要求

1. 数据预处理:

- 下载并加载 Titanic 数据集。可以使用 Pandas 库对数据进行探索性分析 (EDA)，如查看缺失值、数据类型和数据分布等。
- 处理缺失值 (如年龄、船票价格等)，并进行适当的数据转换 (如将分类变量转换为数值变量)。

2. 数据可视化分析:

- 使用 Matplotlib 或 Seaborn 等可视化工具，对影响生存率的关键因素 (如性别、年龄、舱位) 进行数据可视化分析。
- 绘制乘客年龄分布、不同舱位乘客的生存率对比、性别对生存的影响等图表，深入理解数据背后的规律。

3. 模型选择与训练:

- 尝试使用多种分类算法 (如逻辑回归、随机森林、K-近邻、支持向量机等) 进行建模，比较不同模型的性能。
- 使用训练集对模型进行训练，并通过交叉验证评估模型的表现。

4. 模型评估与优化:

- 使用测试集对模型进行评估，计算模型的准确率、精确率、召回率、F1-score 等性能指标。
- 尝试通过特征选择、超参数调优、集成学习等方式提高模型的预测效果。

5. 结果分析与报告撰写:

- 分析模型的预测结果，讨论哪些因素对生存率影响最大，并解释模型的决策逻辑。
- 撰写一份报告，报告内容应包括数据探索与可视化分析、模型选择与训练、结果分析与讨论等部分。

项目 2: House Prices - Advanced Regression Techniques

项目简介

房价预测是经济学和房地产市场中的一个重要问题。Kaggle 的 House Prices 竞赛要求你根据多个影响房价的特征 (如房屋面积、房龄、地理位置等)，预测房屋的最终销售价格。这是一个典型的回归问题，非常适合初学者学习如何在机器学习中处理连续变量的预测任务。

项目要求

1. 数据预处理:

- 下载并加载 House Prices 数据集，进行数据探索性分析 (EDA)，如查看缺失值、数据类型和数据分布等。
- 处理数据中的缺失值和异常值，并对数据进行适当的转换，如对数变

换、标准化和分类变量编码。

2. 数据可视化分析:

- 使用 Matplotlib 或 Seaborn 等工具, 对房价与各特征(如房屋面积、房龄、地理位置等)的关系进行可视化分析。
- 绘制散点图、热力图、分布图等, 分析不同特征对房价的影响。

3. 模型选择与训练:

- 尝试使用多种回归算法(如线性回归、岭回归、Lasso 回归、决策树、随机森林、XGBoost 等)进行建模, 并比较它们的表现。
- 使用训练集对模型进行训练, 并通过交叉验证评估模型的表现。

4. 模型评估与优化:

- 使用测试集对模型进行评估, 计算模型的均方误差(MSE)、均方根误差(RMSE)等性能指标。
- 尝试通过特征工程、超参数调优、集成学习等方式优化模型的预测效果。

5. 结果分析与报告撰写:

- 分析模型的预测结果, 讨论哪些特征对房价影响最大, 并解释模型的预测逻辑。
- 撰写一份报告, 内容包括数据探索与可视化分析、模型选择与训练、结果分析与讨论等部分。

参考链接:

Titanic - Machine Learning from Disaster

<https://www.kaggle.com/competitions/titanic>

House Prices - Advanced Regression Techniques

<https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques>

9.Fashion-MNIST 数据集识别

背景:

Fashion-MNIST 是一个由 Zalando Research 发布的图像数据集, 用于替代传统的 MNIST 手写数字数据集。与 MNIST 不同, Fashion-MNIST 包含了各种时尚产品的灰度图像, 涵盖 10 个不同类别的服装和配饰(如 T 恤、裤子、包、鞋等)。每个类别包含 7000 张 28x28 像素的图片, 其中训练集和测试集各占 60000 张和 10000 张。

Fashion-MNIST 数据集的设计初衷是为了提供一个比手写数字更具挑战性的数据集, 供研究人员和工程师测试机器学习和深度学习模型的性能。由于时尚产品图像具有较高的多样性和复杂性, 使用该数据集进行分类任务, 不仅可以帮助我们了解模型在图像识别任务中的表现, 也可以推动视觉识别领域的研究与应用。

在本项目中, 你将使用 Fashion-MNIST 数据集, 训练并测试一个分类模型。你可以选择不同的机器学习或深度学习算法, 并对模型的性能进行优化和评估。本任务旨在帮助你掌握图像数据处理、模型训练与优化的基本技能, 并加深你对卷积神经网络(CNN)等深度学习技术的理解。

任务描述:

1. 根据如下教程和代码, 搭建基础的 PyTorch 环境, 实现 Fashion-MNIST 识别程序的运行:

https://pytorch.org/tutorials/beginner/basics/quickstart_tutorial.html

https://zh.d2l.ai/chapter_linear-networks/image-classification-dataset.html

2. 在附录中的探索方向任选几样深度学习技巧进行探索, 汇报这些技巧是否可行/发现了什么结果, 并尝试搜索相关资料进行解释。

附录: 机器学习探索附录

这里介绍一些机器学习的常见技巧和值得探索实验的方向:

1 模型扩展与优化

- 多种模型比较: 尝试不同的机器学习模型 (例如支持向量机、随机森林) 以及深度学习模型 (例如不同架构的现代卷积神经网络, 可参阅 https://zh.d2l.ai/chapter_convolutional-modern/index.html), 并对比这些模型在手写数字识别任务中的表现。
- 超参数调优: 使用网格搜索或随机搜索进行超参数调优, 找出最优的模型参数组合。
- 模型集成: 尝试模型集成 (如投票分类器或堆叠模型), 提升模型性能。

2 数据增强与处理

- 数据增强: 对 Fashion-MNIST 数据集进行数据增强 (如旋转、缩放、翻转等), 并观察其对模型准确率的影响。
- 不平衡数据处理: 制造一个不平衡的子集, 设计和实现处理不平衡数据的策略, 如过采样、欠采样、或使用平衡损失函数。

3 迁移学习

- 迁移学习: 使用预训练的模型 (在更大的数据集如 ImageNet 上预训练的 CNN 模型), 然后对 Fashion-MNIST 数据集进行微调, 从而实现迁移学习。

4 跨域应用

- 跨域数据集测试: 将训练好的模型应用到其他类似的数据集 (如 EMNIST) 上, 观察模型的泛化能力。

5 可解释性与模型评价

- 模型可解释性: 使用可视化工具 (如 Grad-CAM、SHAP) 来解释模型的决策过程, 探讨模型如何做出决策。
- 模型评价指标: 除了准确率, 计算并分析其他评价指标 (如 F1-score、ROC 曲线、AUC 值等), 观察不同模型其他指标与准确率之间的关系。

6 实际应用模拟

- 噪声与干扰: 模拟实际应用中可能存在的噪声和干扰 (如图像模糊、背景杂物), 要求学生的模型能够在这些情况下仍然有效地识别目标类型。

10.其他自选题目：

如果感觉有其他的感觉比较合适的选题，可以向老师申请。