例题

简介

比赛时组委会将为每位参赛选手提供一个 账号用于登录算力平台，比赛相关训练、调试代码及数据预先保存在云端。

比赛云平台开发环境如下：

名称	版本
Ubuntu	18.04.6 LTS

名称	版本
Python	3.10.13
jupyter lab	版本 4.2.5
sklearn	1.5.2
LLaMA-Factory	0.9.1.dev0
pytorch	2.1.0
transformers	4.46.1

- NPU：910B（1块）
  参赛选手可以提前访问网站查看使用说明。
- 代码编辑与交互
  参赛可在 jupyter lab环境中，选择notebook方式、py程序和终端执行方式

任务1：数据挖掘建模（50分）

1.1、题目描述

- 参赛选手通过数据挖掘技术对鸢尾花数据集（Iris）进行分析建模。参赛选手需要完成数据预处理、模型训练和模型推理，并提交推理结果。
- 训练数据集存放在 `./data/iris_train.csv` 您需要根据这份数据集训练模型。
- 测试数据集存放在`./data/iris_test.csv` 您需要根据这份数据集的数据特征用训练的模型推理出鸢尾花分类标签

1.2、任务描述

1) 数据预处理（10分）

- **数据集描述**：`iris_train.csv`文件，是一份鸢尾花数据，集每条记录包含4个特征（花萼长度、花萼宽度、花瓣长度、花瓣宽度）和1个标签（鸢尾花的种类：Setosa、Versicolor、Virginica）。
- **预处理步骤**：
  - **数据清洗**：检查数据集中是否存在缺失值或异常值，并进行相应的处理（如删除、填充等）。
  - **特征选择**：根据需要选择合适的特征进行建模。
  - **数据标准化/归一化**：对特征进行标准化或归一化处理，以提高模型训练的稳定性。
- **保存预处理后的数据**：将清洗后的数据集保存到`output`目录下，文件格式为CSV，文件名为`cleaned_train_data.csv`。
-

2) 建模及模型训练（10分）

- **环境准备**：组委会提供Jupyter lab环境和基础算法包scikit-learn，选手可*自行安装其它算法库*

- **模型选择**：选择合适的分类模型进行训练，例如K近邻（KNN）、支持向量机（SVM）、决策树、随机森林等。
- **模型训练**：使用清洗后的训练集对选择的模型进行训练。
- **保存代码在云端**：将模型代码或笔记(.ipynb)保存在云端，需要确保过程可复现

3) 模型推理和评估（10分）

- **模型推理**：使用训练后的模型在测试集上进行推理，预测标签数据。
- **结果保存**：将推理后的结果保存在云端，待本模块任务结束后，由裁判监督统一下载拷贝上交。
- **结果格式**：推理结果应包含测试集的样本ID和对应的预测标签，格式如下：

```
Sample_ID, Predicted_Species
1, Flag_Type_1
2, Flag_Type_2
3, Flag_Type_3
...
```

- **保存推理结果在云端**：将推理结果保存到`output`目录下，文件名为`test_data_predictions.csv`。

4) 模型推理和评估（20分）

- **模型准确性评分**：根据模型在测试集上的准确性评分，最高20分。

1.3、评分标准

- **数据预处理**（10分）：数据清洗、特征选择、数据分割、标准化/归一化、编码等步骤的合理性和有效性，以及预处理后数据文件的保存。
- **模型选择与训练**（10分）：模型选择的合理性、训练过程的正确性。
- **模型推理与结果提交**（10分）：推理结果的准确性、结果保存的格式正确性，
- **模型准确性评分以及模型准确性评分。(20分)

> 模型推理结果输出 10分， 准确性排名10分

任务2：大模型微调与应用（50分）

2.1、题目描述

组委会提供 Qwen1.5-1.8B-Chat 大语言模型，以及节选的MRPC数据集，需要参赛选手通过微调技术对大模型进行微调，使微调后的大模型在提供的测试数据集上取得更优的精度。参赛选手需要完成数据预处理、模型微调和模型推理，并提交推理结果。

模型 `Qwen1.5-1.8B-Chat` 模型路径 `/home/public/data/Model/Qwen1.5-1.8B-Chat`

数据集 MRPC

MRPC（Microsoft Research Paraphrase Corpus）是一个用于自然语言处理（NLP）任务的数据集，主要用于句子对相似性判断任务。该数据集由微软研究院（Microsoft Research）发布，旨在帮助研究人员开发和评估句子对相似性检测模型。

数据集概述

MRPC 数据集包含大量的句子对，每个句子对都有一个标签，表示这两个句子是否是同义句（即是否表达了相同的意思）。数据集的主要特点如下：

- **句子对**：数据集中的每个样本由两个句子组成。
- **标签**：每个句子对都有一个二进制标签（0 或 1），表示这两个句子是否是同义句。
  - `1`：表示两个句子是同义句。
  - `0`：表示两个句子不是同义句。

数据集示例

文件的格式如下：

```
# 标签 #1 String #2 String
1 This is a sentence. This is another sentence.
0 This is a different sentence. This is a completely different sentence.
```

以下是 MRPC 数据集中的一个示例：

```
1 The company announced its earnings report. The firm released its financial results.
0 The company announced its earnings report. The firm released its annual report.
```

在这个示例中：

- 第一个句子对是同义句，因为“earnings report”和“financial results”表达了相同的意思。
- 第二个句子对不是同义句，因为“earnings report”和“annual report”表达了不同的意思。

数据集文件

- ./data/msr_paraphrase_train.tsv. 训练集
- ./data/msr_paraphrase_train.tsv 测试集

2.2、任务描述

1) 数据预处理（10分）

- **数据集描述**：根据`训练`数据集，结合任务目标生成Alpaca json格式训练数据
- **保存预处理后的数据**：将生成的训练数据保存为json文件，结果保存在云端。

2) 模型微调（20分）

- **环境准备**：组委会提供LLamaFactory大模型训练框架
- **模型选择**：使用主办方提供的的大模型为基座模型。
- **模型训练**：使用清洗后的训练集对指定的大模型进行微调。
- **保存训练指令在云端**：将微调指令保存在云端

3) 模型推理和评估（20分）

- **模型推理**：使用训练后的模型用`测试集`进行推理，预测标签数据。
- **结果保存**：将推理后的结果保存在云端，待本模块任务结束后，由裁判监督统一下载拷贝上交。
- **结果格式**：推理结果应包含测试集的样本ID和对应的预测标签，格式如下：

```
ID, preQuality
1, 0
2, 1
3, 1
...
```

- **保存推理结果在云端**：保存在`output`目录下，名称为`test_data_llm_predictions.csv`

2.3、评分标准

- **数据集转换**（10分）：根据数据集，结合任务目标生成正确的Alpaca json格式训练数据。
- **模型选择与训练**（20分）：训练脚本正确。
- **模型推理与结果提交**（10分）：推理结果的准确性、结果保存的格式正确性，
- **模型准确性评分以及模型准确性评分。(20分)

> 模型推理结果输出 10分， 准确性排名10分

注意事项

- 参赛选手需确保数据预处理和模型训练的代码可复现，以便裁判进行验证。
- 推理结果需在规定时间内提交，逾期提交将不予评分。
- 选手需要在output目录下输出的内容

文件名	说明
cleaned_train_data.csv	鸢尾花清洗后的数据集
test_data_predictions.csv	鸢尾花推理结果
MRPC_train_data.json	MRPC数据集转换成LLM训练结构
train.sh	模型微调命令

文件名	说明
test_data_llm_predictions.csv	MRPC 测试集推理结果
*.ipynb or *.py	处理程序