

蓝桥杯全国软件和信息技术专业人才大赛组委会

第十五届蓝桥杯全国软件和信息技术专业人才大赛 项目实战赛——人工智能科目竞赛规则及说明

1. 参赛资格

人工智能科目分为大学组和职业院校组。具有正式学籍的在校全日制研究生、本科生(以报名时状态为准)只可报名参赛大学组；高职高专学生(以报名时状态为准)只可报名参赛职业院校组。

2. 竞赛赛程

本次比赛设立全国赛，比赛时间和个人赛（软件赛和电子赛）省赛及决赛时间不冲突。

比赛时长：4 小时。

详细赛程安排以组委会公布信息为准。

3. 竞赛形式

本次比赛为团队赛，每支参赛队须由 3 名选手（设队长 1 名）组成。

采用线下比赛的方式。一人一机，全程机考。

选手机器通过局域网连接到各个赛场的竞赛服务器。

选手答题过程中无法访问互联网，也不允许使用本机以外的资源（如 USB 连接）。

竞赛系统以“服务器-浏览器”方式发放试题、回收选手答案。

团队成员可自行分配并共同完成试题中的任务，每位成员答完题后可将答案通过 FTP 发送给队长，队长回收所有团队成员答案后一起提交至比赛系统，超过比赛时间将无法提交。

4. 参赛选手机器环境

选手机器配置：

X86 兼容机器，Intel Core i7 8 核 16 线程 CPU 处理器、内存不小于 16G，固态硬盘不小于 500G，操作系统：Windows10 及以上。

编程环境：

序号	软件名称及版本号
1	Chrome 浏览器 v90 以上版本

2	Visual Studio Code 软件版本: v1.76 扩展: Chinese、Python、Jupyter
3	Anaconda Anaconda3-2023.03
4	Python 3.8
5	Notebook 6.5.2
6	Numpy 1.24.4
7	Tensorflow 2.10.0
8	Keras 2.10.0
9	Torch 2.0.1
10	Scikit-learn 1.3.0
11	Pandas 1.3.5
12	Matplotlib 3.3.4
13	Scipy 1.9.3
14	Opencv-python 4.2.0.32
15	Transformers 2.6.0
16	Jieba 0.42.1
17	Gensim 4.3.1
18	Pillow 9.4.0
19	Requests 2.28.2
20	Beautifulsoup4 4.11.1
21	Imbalanced-learn 0.11.0
22	Scikit-image 0.16.2
23	Six 1.16.0
24	Gradio 3.43.2
25	Onnx 1.12.1
26	Onnxruntime 1.12.0
27	Flask 2.2.3

5. 试题形式

试题全部为实操任务。选手根据任务要求,通过编写程序代码、完善程序代码的形式完成预期需求。题目均包含完整的题面 PDF 文档和基础源代码压缩包。题面文档中会详细说明题目的背景、需求、目标。选手需认真读题,实现最终目标。

6. 试题范围

本次比赛主要考察机器学习、深度学习、自然语言处理、计算机视觉等人工智能技术应用相关领域职业能力。要求选手根据竞赛题目进行业务需求分析,基于人工智能实训环境,完成数据预处理、模型训练与评估、模型应用部署等工作任务。考查的技术技能包括但不限于以下内容:

模块	技术技能	描述
一	数据预处理	数据（数值、文本、图像等）清洗、异常值检测与处理、数据转换、数据标准化/归一化、数据不均衡处理、特征提取等数据处理技术。
二	模型训练与评估	线性回归、逻辑回归、决策树、朴素贝叶斯、随机森林等常见的传统机器学习算法应用；卷积神经网络（Convolutional Neural Networks, CNN）、循环神经网络（Recurrent Neural Networks, RNN）、长短期记忆网络（Long Short-Term Memory, LSTM）、自编码器（AutoEncoders）、生成对抗网络（Generative Adversarial Networks, GAN）、等深度学习模型应用；Scikit-learn 机器学习库应用、Tensorflow、Keras、Pytorch 深度学习框架应用；模型效果评估，包括准确率、精确率、召回率、F1 分数、R-squared、均方误差等模型性能评估指标计算方法。
三	模型应用部署	ONNX 模型转换、TensorFlow Lite 模型转换；Flask 框架应用；Gradio 模型部署；模型压缩；模型量化；模型输出后处理。

7. 答案提交

选手只有在比赛时间内提交的答案是有效的，比赛之后的任何提交均无效。

每个参赛队提交答案前，需要通过 WinSCP 工具以 FTP 形式合并各队员的答案。每个参赛队只提交一份最终答案，可重复提交，以最后一次提交的答案为准并作为评测的依据。

选手须使用比赛指定的方式来提交答案，任何其他方式的提交（如邮件、U 盘）都不作为评测依据。

比赛过程中，试题分数不会显示给选手，选手应当在没有反馈的情况下自行设计数据调试自己的程序。

选手须仔细阅读并严格遵守试题指定的答案文件格式或内容。

8. 样题

竞赛说明

一、背景描述

从党的十八届三中全会提出推动媒体融合发展重大任务，到“十四五”规划建议中明确提出推进媒体深度融合、实施全媒体传播工程、做强新型主流媒体、建强用好县级融媒体中心；从中央办公厅、国务院办公厅印发《关于推动传统媒体和新兴媒体融合发展的指导意见》，到印发《关于加快推进媒体深度融合发展的意见》，媒体融合发展成为国家战略。推进传统媒体和新兴媒体融合发展，技术创新是驱动力。人工智能、大数据、云计算、5G、物联网、区块链等核心技术发挥重要作用。

在移动互联网时代，信息不再稀缺。以新媒体平台为例，随着平台普及和用户规模的不断扩大，大量的数据源源不断地产生和积累，利用人工智能技术来打造智能生产和传播平台已成为一种必要和重要的需求。新媒体平台内容一般呈现出多模态的特点，包括文本、图像、视频等形式。不同模态的数据可以提供丰富的信息表达方式。图像和视频可以传达更直观、生动的视觉信息，而文本则可以提供更加详细和抽象的语义信息。基于人工智能技术，可以充分挖掘数据中的潜在信息，提供更深入的分析和洞察；可以提供更高效的创作者内容生产工具；可以自动检测信息传播的准确性；可以全方位地分析用户的行为、需求、兴趣和情感状态，为企业、组织带来更精准的决策支持和业务价值。

现在，你们是某新媒体平台的研发团队成员。为了使平台提供更加智能化、个性化的用户体验，你们将结合自然语言处理、计算机视觉等领域的人工智能技术，以 Python 作为基础开发语言，利用传统机器学习和深度学习算法进行新媒体数据的相关处理、分析、应用。你们作为该团队的核心成员，请按照下面任务完成本次工作。

二、成果物提交

人工智能应用开发赛项参赛选手按照各模块的任务要求完成对应的成果物，将各模块的成果物统一压缩为“成果物.zip”进行提交，本赛项基于选手提交的竞赛成果物，进行机器自动评分。所有提交内容的文件命名务必与任务中给出的文件命名保持一致，否则会导致判分异常。

需提交的成果物如下：

模块	任务	成果物
模块 1	任务 1	songs_processed.csv 文件
	任务 2	2.py 文件
	任务 3	3.py 文件
模块 2	任务 1	songs_testout.csv 文件
	任务 2	image_classify.h5 文件
	任务 3	pred_test.txt 文件
模块 3	任务 1	1.py 文件

	任务 2	2.py 文件、text_classifier.onnx 文件
	任务 3	3.py 文件

模块一：数据预处理

任务 1：缺陷数据预处理

【介绍】

在当今数字音乐时代，大量的歌曲被发布，但只有少数能够进入热门排行榜。对于音乐产业而言，能够准确预测一首歌曲的热度具有重要的商业意义。本任务提供了一个关于音乐的数据集，它包含了大量的歌曲信息以及与之相关的各种特征，基于这些特征，我们能够使用机器学习算法来预测歌曲的热度。但是，这个数据集存在缺陷，需要你们研发小组对这个数据集做一些预处理。`songs_origin.csv`，是本任务提供的数据集。数据集中包含了丰富的音乐元数据和特征，如声学特性、舞蹈性、能量等。

【目标】

请按以下要求实现对数据集的预处理。

- ① 处理数据集中的缺失值，对于数据集中的缺失值，以其所在列的均值进行填充。
- ② 处理数据集中的异常值，对于 `acousticness_yr` 列的值大于 1 或小于 0 的行进行删除。
- ③ 处理数据集中的重复行，对于数据集中出现多行的相同数据，只保留一行，删除其余重复行。

请在模块一文件夹下 1.py 文件中编写代码，针对数据集 `songs_origin.csv` 完成以上数据处理，并将处理后的结果保存在模块一文件夹下，命名为 `songs_processed.csv`。正确实现以上对数据集的处理，即完成目标。

任务 2：图像数据预处理

【介绍】

在新媒体平台上，用户经常分享各种图片，这些图片反映了他们的兴趣和喜好。因此这些图片就可以用来制作后续图像分析要用的图像数据集。在本任务中，我们收集并整理了一批图像数据，你们研发小组将对该数据集进行必要的图像预处理。`Imagedata.zip` 是本任务提供的数据集。解压后，你将看到一个 `image_class_index.json` 文件和一个 `images` 文件夹。

其中, `image_class_index.json` 文件是类别索引映射文件, `images` 文件夹包含了若干文件夹, 每个文件夹下是同一类别的图像。

【目标】

请按要求编写以下函数代码。

`img_processor` 函数

✓ 函数功能

对输入图像进行预处理, 包括图像缩放、中心裁剪和标准化。

✓ 参数

`data_path`: 字符串, 指定图像文件的路径。

`dst_size`: 元组, 指定最终输出图像的目标尺寸, 默认为 `(224, 224)`。

✓ 返回值

`image_src`: `numpy.ndarray`, 原始图像。

`image`: `numpy.ndarray`, 预处理后的图像, 尺寸为 `dst_size`。

`(startx, starty)`: 元组, 中心裁剪的起始坐标, 表示裁剪区域在原始图像中的起始位置。

请在模块一文件夹下 `2.py` 文件中 `#TODO` 处补充代码, 确保实现以下目标。

① 图像缩放: 将图像大小调整为 `(256, 256)`。

② 中心裁剪: 计算裁剪的起始坐标 `startx` 和 `starty`, 从图像中裁剪出大小为 `(224, 224)` 的中心区域。

③ 标准化: 使用代码中设置的 `_mean` 和 `_std` 对图像像素进行标准化。

任务 3: 数据增强

【介绍】

数据增强是提升新媒体平台中自然语言处理 (NLP) 模型性能的关键技术。在这些平台上, 用户生成的文本数据多样且丰富, 但也往往包含有限的语义表达方式。通过数据增强, 我们可以创造出各种训练样本的变体, 这不仅扩大了训练数据集, 还增强了模型对新颖表达方式的**理解能力, 从而提升了模型在实际应用中的泛化能力。同义词替换就是一种有效的数据增强方法, 它通过使用同义词典替换训练数据中的词汇, 以生成语义相似但表达不同的训练样本。在本任务中, 考生需要在 PyTorch 的 Dataset 类中实现这样一个数据增强方法, 借助提供的同义词典, 实现同义词替换功能。这种方法尤其适用于新媒体平台, 可以帮助模型更好地适应并理解用户的多样化语言表达, 进而提高新媒体平台上各类 NLP 应用的性能。

`data.csv` 是训练数据, 包含 160 条样本, 其中 `text` 列表示文本内容, `text_id` 列表示文本对

应的 `id.loc.txt` 和 `per.txt` 是本任务提供的地名和称谓增强字典,共包含了 273 个地名和 66 个家庭称谓。

【目标】

请按要求编写以下函数代码。

`augment` 函数

✓ 函数功能

对输入的文本进行数据增强。

从地名/称谓增强词典中，找出所有出现在输入文本中的地名/称谓。

对于每个找到的地名/称谓，从增强词典中随机选择一个新的地名/称谓作为替代，并将文本中的原地名/称谓替换为这个新地名/称谓。

替换信息记录在 `aug_info['locs']` 和 `aug_info['pers']` 列表中，包括原始地名/称谓和替换后的地名/称谓。

✓ 参数

`text` 输入文本

✓ 返回值

增强后的文本和增强信息。

`aug_info['locs']` 和 `aug_info['pers']` 为嵌套字典列表，其中列表中每个元素及其说明见下表。

字段名	释义	数值类型
<code>original</code>	原始实体	字符串
<code>replacement</code>	替换后的实体	字符串

请在模块一文件夹下 `3.py` 文件中 `#TODO` 处补充函数代码，并运行，确保能够实现以下目标：

- ① 正确地返回增强后的文本和增强信息。
- ② 除了对增强词典中的替换，其他文本保持不变。

模块二：模型训练与评估

任务 1：回归预测模型训练

【介绍】

音乐对于许多人来说是一种情感表达和情绪释放的方式。音乐流媒体平台为用户提供了广泛的音乐内容和社交互动功能。通过分析歌曲的特征和历史数据，可以估计哪些歌曲可能受用户欢迎、应该最大程度地提高哪些歌曲的曝光度。本任务提供了一个关于音乐的数据集，

它包含了大量的歌曲信息以及与之相关的各种特征。目前，我们已经对这个数据集做了一些预处理和特征选择。请你们研发小组基于这些特征，使用机器学习算法来实现自动预测歌曲的热度。`songs_train.csv` 是本任务提供的训练集。其中，`popularity` 列是目标变量，其余是有关音乐的特征，如声学特性、舞蹈性、能量等。`songs_test.csv` 是本任务提供的测试集，与训练集同分布。其中，`popularity` 列是目标变量（该列数值为空），其余是有关音乐的特征，如声学特性、舞蹈性、能量等。

【目标】

请在模块二文件夹下 `1.py` 文件中编写代码，按以下要求实现对歌曲热度的自动预测。

- ① 选择合适的回归算法，构建一个回归模型。
- ② 使用训练好的回归模型预测测试集 `songs_test.csv` 的目标变量。
- ③ 将输出结果保存在模块二文件夹下，命名为 `songs_testout.csv`。注意：结果文件中保留测试集 `songs_test.csv` 的原始数据，仅填充 `popularity` 列即可。
- ④ `songs_testout.csv` 结果得分不低于 0.8，视为实现目标。

提示：判题过程中的得分计算，以回归模型的 R 平方（决定系数）为准。

任务 2：图像分类模型训练

【介绍】

随着新媒体平台上图像内容的爆炸性增长，为了帮助平台更好地组织和推荐内容，你们研发小组定制训练了一套图像分类模型。为验证该模型在新媒体环境中的适应性和结构是否正确，你们将采用一种常见的方法，即使用少量数据对其进行训练，并观察模型是否能够过拟合这些数据。当模型过拟合，即表示它能完全匹配到训练数据，验证了网络结构在实际应用中的有效性。接下来你们要设计一个图像分类模型，并使用提供的少量数据进行训练，直至模型过拟合。`dataset` 是本任务提供的训练数据，一共 10000 张图片，10 个类别，图片大小为 (32, 32, 3)。

【目标】

请按要求编写以下函数代码。

`build_model_and_train` 函数

✓ 函数功能：

构建一个 10 分类的图像分类模型，输入尺寸为 (batch_size, 32, 32, 3)。

定义模型的损失函数，可以是任意的分类损失函数。

训练模型，直至模型的训练损失小于 $1e-3$ ，训练集准确率（正确分类数量/数据集总数 * 100%）为 100%。

保存模型到模块二文件夹下，命名为 **image_classify.h5**，该文件可以直接被 `tf.keras.models.load_model` 方法读取。

请在模块二文件夹下 **2.py** 文件中 #TODO 处补充函数代码，确保能够实现以下目标：

- ① 正确训练模型直到模型性能指标达到任务给定的要求。
- ② 正确保存模型。

提示：考虑到训练资源和实际应用场景，建议考生设计一个相对轻量的图像分类模型，避免在检测时出现资源超限的问题。

任务 3：虚假信息识别模型训练

【介绍】

新媒体作为信息传播的重要渠道，需确保传播的信息可信可靠。因此，平台需要采取假新闻过滤机制限制虚假信息的传播，提高内容质量。在本任务中，你将基于处理好的数据集，设计并构建一个针对新闻文本的真假分类模型。**news_train.txt** 是本任务提供的训练集文本。文件中每行是一个样本，即经过分词处理后的句子。**label_newstrain.txt** 是本任务提供的训练集标签。文件中每行是一个 0/1 数值标签，与 **news_train.txt** 文件逐行一一对应。0 表示文本真实，1 则表示虚假。**news_test.txt** 是本任务提供的测试集文本。文件中每行是一个样本，即经过分词处理后的句子。测试集的 F1 值不低于 0.9，即视为实现目标 1。

【目标】

请在模块二文件夹下 **3.py** 文件中编写代码，并按以下要求实现对新闻文本的真假进行自动预测。

- ① 选择合适的分类算法，构建一个分类模型。
- ② 使用训练好的分类模型预测测试集 **news_test.txt** 的类别标签。
- ③ 将输出结果保存在模块二文件夹下，命名为 **pred_test.txt**。注意：结果文件中只保存 0/1 标签，每行一个数值，第一行的数值对应 **news_test.txt** 中第一行文本的真假类别，以此类推。
- ④ **pred_test.txt** 结果准确率不低于 0.9，视为实现目标。

提示：判题过程中的准确率计算示例：真实标签为 [0, 1, 0, 1, 1]，预测标签为 [0, 1, 1, 1, 0]，准确率为 0.6，表示 60% 的样本被正确预测。

模块三：模型应用部署

任务 1：模型融合

【介绍】

你们研发小组已经完成了一个音乐热度预测模型的训练。在上线之前，一般需要对模型的泛化能力进行测试。模型融合可能会比单个模型有更好的泛化能力。在本任务中，你们要对多个训练好的回归模型完成模型融合，结合多个回归模型的预测结果，以生成最终的预测结果。

【目标】

`model1.pkl`、`model2.pkl`、`model3.pkl` 是本任务提供的基于 `sklearn` 训练的回归模型。请按要求编写以下函数代码。

`predictY` 函数

✓ 函数功能

加载提供的 `model1.pkl`、`model2.pkl`、`model3.pkl`，分别对给定的测试数据进行预测。

基于指定权重对三个模型的预测结果进行加权平均，得到模型融合后的预测结果。

其中 `model1.pkl`、`model2.pkl`、`model3.pkl` 分别对应的权重为 0.3, 0.1, 0.6。

✓ 参数

`test_data`，字典，`key` 表示特征名，`value` 表示特征值。

✓ 返回值

`output_ensemble`，浮点型数值，表示模型融合后的预测结果。

请在模块三文件夹下 `1.py` 文件中 `#TODO` 处补充函数代码，确保实现以下目标：

① 正确输出模型推理结果。

任务 2：模型转换

【介绍】

在前面的任务中，我们搭建并训练完成了文本分析模型。在实际应用中，还需要将训练后的模型转换为其他格式，以支持不同的推理引擎。ONNX (Open Neural Network Exchange) 提供了一个开放的源模型格式，支持多种深度学习框架之间的模型转换。本任务要求你们研发小组将提供的基于 `PyTorch` 的文本分类模型转换为 ONNX 格式，并实现一个使用 ONNX 模型进行推理的简单应用。`model.pt` 是本任务提供的 `PyTorch` 模型权重文件。

【目标】

请按要求编写以下函数代码。

convert 函数

✓ 函数功能

将本任务提供的 `pt` 模型文件转换为 ONNX 格式。

ONNX 模型文件保存在模块三文件夹下，命名为 `text_classifier.onnx`。

inference 函数

✓ 函数功能

读取 ONNX 模型文件。

使用 ONNX 模型进行推理。

返回推理结果。

✓ 参数

`model_path`，字符串类型，ONNX 模型文件的绝对路径。

`input`，整数列表类型，为待预测样本，示例如：[101, 304, 993, 1008, 102]。

✓ 返回值

`result`，浮点数列表类型，为 ONNX 文件推理结果，示例如：[[0.53419, 0.44313]]。

请在模块三文件夹下 `2.py` 文件中 `#TODO` 处补充函数代码，并执行 `main()` 函数，确保能够实现以下目标：

- ① 正确转换模型文件。
- ② 正确输出推理结果。

任务 3：模型部署

【介绍】

为了充分挖掘新媒体平台上积累的用户文本数据中的命名实体信息并为其提供应用服务，我们需要将该模型部署到线上环境。一般来说，在自然语言处理应用的开发中，我们会使用 Web 框架（如 `Flask`）部署模型，并为外部提供 API 接口以支持在线预测。在本任务中，你们研发小组将接手一个部分完成的 `Flask` 项目。该项目已经包括了命名实体识别（NER）模型的载入、运行以及接收 `tokenize` 后的数据的代码。考生的核心任务是将模型的整数向量输出转换为一个结构化的实体标注列表，并通过 `Flask API` 返回这个列表。此步骤不仅实现了自动化的实体识别，更使得新媒体平台可以实时、高效地获取文本中的实体信息，为后续的数据分析和运营决策提供有力支撑。`ner.pt` 是本任务提供的 `PyTorch` 模型权重文件。

【目标】

请按要求编写以下函数代码。

`process` 函数

✓ 函数功能

将本任务提供的请求数据使用提供的模型进行推理。

将推理结果转换为结构化的实体标注列表。

对于不合理的序列，如单字实体 ['O', 'B-LOC', 'O'] 和以 I- 起始的实体 ['O', 'I-LOC', 'I-LOC']，将其作为非实体序列处理。

实体标注列表为一个嵌套字典列表，其中列表中每个元素及其说明见下表。

字段名	释义	数值类型
start	实体的起始位置	数值
end	文本的结束位置	数值
label	实体类型	字符串

请在模块三文件夹下 **3.py** 文件中 **#TODO** 处补充函数代码，并运行代码块，确保能够实现以下目标：

- ① 正确地使用模型进行推理。
- ② 返回正确的结构化结果。

9. 评分

全部题目将使用机器自动评分。评分标准如下：

模块	任务	评分细则	总分值
模块 1	任务 1	完成目标得 6 分。	20
	任务 2	完成目标得 7 分。	
	任务 3	完成目标得 7 分。	
模块 2	任务 1	完成目标得 15 分。	50
	任务 2	完成目标得 20 分。	
	任务 3	完成目标得 15 分。	
模块 3	任务 1	完成目标得 10 分。	30
	任务 2	完成目标得 10 分。	
	任务 3	完成目标得 10 分。	

10. 其他注意事项

（1）选手必须符合参赛资格，不得弄虚作假。资格审查中一旦发现问题，则取消其报名资格；竞赛过程中发现问题，则取消竞赛资格；竞赛后发现问题，则取消竞赛成绩，收回获奖证书及奖品等，并在大赛官网上公示。

（2）参赛选手应遵守竞赛规则，遵守赛场纪律，如有任何违规，将会被组委会取消报名或竞赛资格。

（3）竞赛采用机器阅卷+少量人工辅助。选手要特别注意提交答案的形式。必须仔细阅读题目的要求和示例，不得随意添加不需要的内容。

（4）大赛组委会将于赛前两周在大赛官网发布比赛手册，请参赛选手须按照比赛手册中的要求进行备赛。

（5）从第十三届蓝桥杯大赛全国总决赛开始，历届违纪作弊选手三年内将被禁止参加蓝桥杯大赛任何赛项，其他要求见大赛官网蓝桥杯大赛比赛管理办法

<https://dasai.lanqiao.cn/notices/844>。