



燕山大学

Python 机器学习三级项目指导书

Python Machine Learning Three-level Project Instruction

教 务 处

2024 年 5 月

目 录

一、三级项目分组	1
二、三级项目选题	1
(一) 简历增值类赛题	1
飞桨大赛: https://aistudio.baidu.com/aistudio/competition/1/1	1
千言数据集评测: https://aistudio.baidu.com/aistudio/competition/4/1	1
文心智能体大赛: https://aistudio.baidu.com/competition/5/1	1
(二) 学习检验类赛题	2
1. 基于 caltech101 的图像分类问题	2
2. 安全帽检测	3
3. 新冠 X 射线图像分类	5
4. 中文新闻文本标题分类	6
5. 基于集成学习的 Amazon 用户评论质量预测	8
三、三级项目实施	10
四、三级项目答辩	10
1. 答辩形式	10
2. 答辩时长	11
3. PPT 内容	11
4. 提交材料	11

一、三级项目分组

三级项目以真实竞赛题目进行练习。

- “简历增值类”赛题：由学生自愿选题并报名，需要根据比赛要求的人数来组队，最多不超过 5 人。
- “学习检验类”赛题：先由学生自行分组，原则上每组 4 人，最多上限 5 人；再由老师随机指定题目给各队伍。

注意：

- (1) 组队仅限本班，不可跨班组队。
- (2) 每组选 1 人做小组负责人（组长），分配组员工作任务，分工合作完成三级项目。
- (3) 根据老师发放的分组模板，各班班长统计本班分组情况，认真填写小组信息，不要有信息错误。班长统计人员齐全并整理好后，统一交给老师。

二、三级项目选题

（一）简历增值类赛题

此类赛题由学生自愿选择，并报名参赛。

适宜学生：适合想要提高简历含金量的同学，或想保研加分的同学（仅限一类赛赛题）。

可选题目范围：

飞桨大赛：<https://aistudio.baidu.com/aistudio/competition/1/1>

千言数据集评测：<https://aistudio.baidu.com/aistudio/competition/4/1>

文心智能体大赛：<https://aistudio.baidu.com/competition/5/1>

注意事项：

- (1) 选题时要注意所选比赛的竞赛时间，尤其关注提供资源的时间，一定要选择在三级项目 18 周进行项目验收前可完成的赛题。
- (2) 选题中有一些是我校认定的一类赛，报名参赛若获奖，可以保研加分，想在做三级项目的同时参赛的同学，可以选择其中的一类赛赛题。
- (3) 涉及硬件支撑的，学校不提供。
- (4) 有的赛题提供了基线代码，必须有明显的改进与提升。

（二）学习检验类赛题

老师随机从以下 5 个题目中任选其一分配给此类的各组。

题目位于 AI Studio 平台“Python 机器学习（2024）22 级”课程中“比赛”一栏。

1. 基于 caltech101 的图像分类问题

（1）任务描述

基于 Caltech101 数据集的图像分类，Caltech101 包含 101 种类别的物体，每种类别大约 40 到 800 个图像，本次练习赛选取了其中 16 个类别，需要根据图片特征，用算法从中识别该图像属于哪一个类别。

（2）数据说明

任务所使用图像数据集，包含 1567 张图片，被分为 16 类，每个类别图片超过 80 张。16 个类别分别为：ak47、binoculars、boom-box、calculator、cannon、computer-keyboard、computer-monitor、computer-mouse、doorknob、dumb-bell、flashlight、head-phones、joy-stick、palm-pilot、video-projector、washing-machine。

已将训练集按照“图片路径+\t+标签”的格式抽取出来，可以直接进行图像分类任务，希望答题者能够给出自己的解决方案。

训练集格式：图片路径+\t+标签

测试集格式：图片路径

（3）提交答案

需要提交模型代码项目版本和结果文件。结果文件为 TXT 文件格式，命名为 result.txt，文件内的字段需要按照指定格式写入。

结果文件要求：

- 1) 每个类别的行数和测试集原始数据行数应一一对应，不可乱序。
- 2) 输出结果应检查是否为 205 行数据，否则成绩无效。
- 3) 输出结果文件命名为 result.txt，一行一个类别。

样例如下：

...

2

10

1

5

9

12

15

3

12

3

2

6

7

8

9

3

5

15

0

3

2

...

2. 安全帽检测




(1) 任务描述

工地、工厂等场所在进行安全生产时，要求所有进入场地的人员都佩戴安全帽，每年都会有因为不按要求佩戴安全帽而产生安全事故。

要求参赛者给出一个算法或模型，对于给定的图片，检测出图片中的人员佩戴安全帽的情况。给定图片数据，选手据此训练模型，为每张测试数据预测出最正确的类别。

(2) 数据说明

安全帽检测数据集共包括 40000 张训练图像和 1000 张测试图像，每张训练图像对应 xml 标注文件。

 test	2021/5/19 14:25	文件夹	
 train	2021/5/19 14:25	文件夹	
 test_name.txt	2021/5/18 20:56	文本文档	21 KB



```

<annotation>
  <folder>images</folder>
  <filename>hard_hat_workers0.png</filename>
  <size>
    <width>416</width>
    <height>416</height>
    <depth>3</depth>
  </size>
  <segmented>0</segmented>
  <object>
    <name>helmet</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <occluded>0</occluded>
    <difficult>0</difficult>
    <bndbox>
      <xmin>357</xmin>
      <ymin>116</ymin>
      <xmax>404</xmax>
      <ymax>175</ymax>
    </bndbox>
  </object>
  <object>
    <name>helmet</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <occluded>0</occluded>
    <difficult>0</difficult>
    <bndbox>

```

共包含 3 类：0:'head'，1:'helmet'，2:'person'。

(3) 提交答案

最终提交名为 pred_result.txt 的文件：每一行代表一个目标，每一行内容分别表示
图像名 置信度 xmin ymin xmax ymax 类别。文件内的字段需要按照指定格式写入。

```

hard_hat_workers4511 0.882 41 77 430 255 2
hard_hat_workers4511 0.682 32 108 234 211 1
hard_hat_workers3186 0.788 78 99 123 156 0

```

3. 新冠 X 射线图像分类

(1) 任务描述

新冠一般指新型冠状病毒肺炎。新型冠状病毒肺炎（Corona Virus Disease 2019，COVID-19），简称“新冠肺炎”，世界卫生组织命名为“2019 冠状病毒病”，是指 2019 新型冠状病毒感染导致的肺炎。2019 年末 2020 年初，新冠大爆发，为全世界人民带来了巨大的灾难。

患新型冠状病毒肺炎的患者，肺部 x-射线影像中会有病毒感染的特征性表现，因此，准确识别出 x-射线影像中的新冠阳性影像，具有重要的现实意义。

本比赛使用科学家们收集并开源的新冠肺炎阳性患者的胸部 x 线影像数据库（包括正常、病毒性肺、新冠阳性三种影像），旨在通过深度学习建模，准确识别出其中的新冠阳性影像。

(2) 数据说明

数据集包括：

- 训练数据：包含标签。
- 测试数据：不包含标签，需要参赛者通过模型预测，给出结果并上传。

（3）提交内容及格式

可参照下列代码，替换相应的模型参数，生成比赛指定格式的提交文件

提交格式： 图片路径\t 类型**（注意 tab 符号分割，图片路径仅保持图片最后一级的路径，不带文件夹路径）**

如：

0.png Viral Pneumonia

1.png NORMAL

15.png Viral Pneumonia

4. 中文新闻文本标题分类

（1）任务描述

基于 THUCNews 数据集的文本分类，THUCNews 是根据新浪新闻 RSS 订阅频道 2005~2011 年间的历史数据筛选过滤生成，包含 74 万篇新闻文档，参赛者需要根据新闻标题的内容用算法来判断该新闻属于哪一类别。

（2）数据说明

THUCNews 是根据新浪新闻 RSS 订阅频道 2005~2011 年间的历史数据筛选过滤生成，包含 74 万篇新闻文档（2.19 GB），均为 UTF-8 纯文本格式。在原始新浪新闻分类体系的基础上，重新整合划分出 14 个候选分类类别：财经、彩票、房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游戏、娱乐。

已将训练集按照“标签 ID+\t+标签+\t+原文标题”的格式抽取出来，可以直接根据新闻标题进行文本分类任务，希望答题者能够给出自己的解决方案。

训练集格式 标签 ID+\t+标签+\t+原文标题 测试集格式 原文标题

（3）提交答案

考试提交，需要提交模型代码项目版本和结果文件。

结果文件为 TXT 文件格式，命名为 result.txt，文件内的字段需要按照指定格式写入。

①每个类别的行数和测试集原始数据行数应一一对应，不可乱序。

②输出结果应检查是否为 83599 行数据，否则成绩无效。

③输出结果文件命名为 result.txt，一行一个类别。

样例如下：

...

游戏

财经

时政

股票

家居

科技

社会

房产

教育

星座

科技

股票

游戏

财经

时政

股票

家居

科技

社会

房产

教育

...

5. 基于集成学习的 Amazon 用户评论质量预测

(1) 任务描述

本案例中我们将基于集成学习的方法对 Amazon 现实场景中的评论质量进行预测。需要大家完成两种集成学习算法的实现 (Bagging、AdaBoost.M1)，其中基分类器使用 SVM 和决策树两种，对结果进行对比分析。

①案例背景

随着电商平台的兴起，以及疫情的持续影响，线上购物在我们的日常生活中扮演着越来越重要的角色。在进行线上商品挑选时，评论往往是我们十分关注的一个方面。然而目前电商网站的评论质量参差不齐，甚至有水军刷好评或者恶意差评的情况出现，严重影响了顾客的购物体验。因此，对于评论质量的预测成为电商平台越来越关注的话题，如果能自动对评论质量进行评估，就能根据预测结果避免展现低质量的评论。本案例中我们将基于集成学习的方法对 Amazon 现实场景中的评论质量进行预测。

②任务

本案例中需要完成两种集成学习算法的实现 (Bagging、AdaBoost.M1)，其中基分类器要求使用 SVM 和决策树两种，因此，一共需要对比四组结果 (AUC 作为评价指标)：

- Bagging + SVM
- Bagging + 决策树
- AdaBoost.M1 + SVM
- AdaBoost.M1 + 决策树

注意集成学习的核心算法需要手动进行实现，基分类器可以调库。

③基本要求

- 根据数据格式设计特征的表示
- 汇报不同组合下得到的 AUC
- 结合不同集成学习算法的特点分析结果之间的差异

④扩展要求

- 尝试其他基分类器（如 k-NN、朴素贝叶斯）
- 分析不同特征的影响
- 分析集成学习算法参数的影响
- 尝试各种方法提升排行榜上预测性能

（2）数据说明

①数据描述

本次数据来源于 Amazon 电商平台，包含超过 50,000 条用户在购买商品后留下的评论，各列的含义如下：

- reviewerID: 用户 ID
- asin: 商品 ID
- reviewText: 英文评论文本
- overall: 用户对商品的打分（1-5）
- votes_up: 认为评论有用的点赞数（只在训练集出现）
- votes_all: 该评论得到的总评价数（只在训练集出现）
- label: 评论质量的 label, 1 表示高质量, 0 表示低质量（只在训练集出现）

评论质量的 label 来自于其他用户对评论的 votes, $\text{votes_up}/\text{votes_all} \geq 0.9$ 的作为高质量评论。此外测试集包含一个额外的列 Id, 标识了每一个测试的样例。

②文件说明

- train.csv: 训练集。
- test.csv: 测试集, 用户和商品保证在训练集中出现过, 没有关于 votes 和 label 的列。

文件使用 \t 分隔, 可以使用 pandas 进行读取:

```
import pandas as pd

train_df = pd.read_csv('train.csv', sep='\t')
```

（3）提交答案

提交文件需要对测试集中每一条评论给出预测为高质量的概率, 每行包括一个 Id (和测试集对应) 以及预测的概率 Predicted (0-1 的浮点数), 用逗号分隔。

示例提交格式如下：

```
Id,Predicted
0,0.9
1,0.45
2,0.78 ...
```

提交文件需要命名为 result.csv。

三、三级项目实施

三级项目共 12 学时，前 8 学时进行实践，后 4 学时进行答辩。

1. 三级项目题目发布在 AI Studio 平台中。

2. 两类赛题的分组与选题：

（1）“简历增值类”赛题，小组需要先进行赛事报名，再共同开展三级项目；

（2）“学习检验类”赛题，每位成员需要通过“比赛”一栏进入相应的题目，共同开展三级项目实施。

注意：AI Studio 平台环境仅支持百度的 PaddlePaddle（飞浆）框架。如果想使用其他框架（如 Pytorch、TensorFlow 等框架），需要在希冀平台登录(<http://222.30.145.84>)，在“机器学习”课程中，通过老师发布的三级项目进入环境去编辑调试程序，按要求输出结果文件后，上传到 AI Studio 平台中进行测评即可。

3. 实时提交，AI Studio 平台实时显示该三级项目提交的结果排名。

4. 凡是平台提供 baseline 基线代码的，三级项目要求必须超过基线模型才有成绩。

5. 禁止抄袭，一经发现，取消小组全体成员成绩。

6. 在指定时间内完成项目，撰写三级项目报告，制作答辩汇报 PPT，最后现场答辩。

四、三级项目答辩

1. 答辩形式

（1）每个小组采用 PPT 讲解 + 程序演示的形式进行答辩。

答辩时以 1 人为主进行讲解，其他人补充说明。

(2) 每组必须全员参加答辩。

每名同学均有可能被提问，所以每个人都需要做好充分的答辩准备。

2.答辩时长

每个小组答辩时间不超过 20 分钟：

- 学生讲解和演示总时长不超过 10 分钟；
- 评审提问问答不超过 10 分钟。

3.PPT 内容

(1) 封面要有选题题目、班级及相应小组序号、小组成员。

(2) PPT 要呈现项目的思路、核心算法、解决过程、调试过程、运行结果（含调参结果对比）、**项目在比赛中的最好排名截图（截止到答辩当日）**。涉及基线代码的，要有基线代码与改进后代码的程序改进对比以及成绩比较。

(3) PPT 要呈现**小组分工**及**组内自评成绩**，**每一档不能超过 2 人**。

4.提交材料

每个小组分别提交以下材料：三级项目报告（电子版+纸质版）、答辩 PPT、完整项目。

(1) **每个文件的命名格式：班级-小组-选题**，如：22-1-第 1 组-垃圾分类。

(2) 提交时间：截止到答辩当天。

(3) 提交方式及内容：

①在学习通上《2024 年 Python 机器学习 实验课（22 级）》课程的三级项目提交处，把所有三级项目电子版材料交齐，具体包括**三级项目报告电子版、答辩 PPT、项目程序文件、结果文件等**。

②注意事项：

- 每个小组由组长提交一份即可。
- 纸质版三级项目报告需要在答辩的同时交给答辩验收的老师。

如果三级项目报告撰写不符合要求，老师有权要求学生修改并重新提交。

封面设计： 贾丽

地 址： 中国河北省秦皇岛市河北大街 438 号

邮 编： 066004

电 话： 0335-8057068

传 真： 0335-8057068

网 址： <http://jwc.ysu.edu.cn>