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

2021 年 3 月

目　录

一、三级项目分组

三级项目分组进行，每组最多 4 人。

组内选出 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。

根据老师在群中发放的分组模板，各组组长认真填写小组信息。

二、三级项目选题

老师随机从以下 6 个题目中任选其一分配给每组。

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.猫十二分类问题

（1）任务描述

利用训练的模型来预测数据所属的类别。

（2）数据说明

本数据集包含 12 种类的猫的图片。

整个数据将被分为训练集与测试集。在训练数据中，我们提供彩色的图片，如图所示。

训练集：在训练集中，我们将提供高清彩色图片以及图片所属的分类。

测试集：在测试数据集中，我们仅仅提供彩色图片。

（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 CSV 文件格式，命名为 result.csv，文件内的字段需要按照指定格式写入。

文件格式：cat_12_test/WMgOhwZzacYO23lCusqnBxIdibpkT5GP.jpg 0

其中，前半部分为【图片路径】，后半部分为【类别编号】。

提交的预测结果要与我们提供的 label 图像名字与格式 保持完全一致，否则上传无法通过格式检查。

3.垃圾图片分类问题

（1）任务描述

近年来，随着人工智能的发展，其在语音识别、自然语言处理、图像与视频分析等诸多领域取得了巨大成功。随着政府对环境保护的呼吁，垃圾分类成为一个亟待解决的问题，本次竞赛将聚焦在垃圾图片的分类，利用人工智能技术，对居民生活垃圾图片进行检测，找出图片中有哪些类别的垃圾。 要求参赛者给出一个算法或模型，对于给定的

图片，检测出图片中的垃圾类别。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

本竞赛所用训练和测试图片均来自生活场景。总共四十个类别，类别和标签对应关系在训练集中的 dict 文件里。图片中垃圾的类别，格式是“一级类别/二级类别”，二级类别是具体的垃圾物体类别，也就是训练数据中标注的类别，比如一次性快餐盒、果皮果肉、旧衣服等。一级类别有四种类别：可回收物、厨余垃圾、有害垃圾和其他垃圾。

数据文件包括训练集(有标注)和测试集(无标注)，训练集的所有图片分别保存在 train 文件夹下面的 0-39 个文件夹中，文件名即类别标签，测试集共有 400 张待分类的垃圾图片在 test 文件夹下，testpath.txt 保存了所有测试集文件的名称，格式为：name+\n。

（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 model_result.txt，文件内的字段需要按照指定格式写入。

提交结果的格式如下：

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。

输出结果应检查是否为 400 行数据，否则成绩无效。

输出结果文件命名为 model_result.txt，一行一个类别标签（数字）

样例如下：

```
・・・
35
3
2
37
10
3
26
4
34
21
・・・
```

4.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。

（2）任务介绍

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

（3）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，votes_up/votes_all ≥ 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。

5.中文新闻主题分类

（1）任务描述

近年来，随着人工智能的发展，其在语音识别、自然语言处理、图像与视频分析等诸多领域取得了巨大成功。本次竞赛将聚焦在新闻文本的分类，利用人工智能技术，对新闻文本进行分类。要求参赛者给出一个算法或模型，对于给定的文本，检测出文本所属类别。给定文本数据，选手据此训练模型，为每个测试数据预测出最正确的类别。文本分类任务，参赛者根据原始文本数据判断其所属类别。

（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。 格式介绍： 为了使参赛者快速进入比赛核心阶段，我们已将训练集按照“标签 ID+\t+标签+\t+原文标题”的格式抽取出来，参赛者可以直接根据新闻标题进行文本分类任务，希望参赛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

试题提交，需要提交模型代码项目版本和结果文件。

要求提交结果的格式如下：

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总行数为 83599，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

6.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

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

...

三、三级项目实施

实施平台：AI Studio。

实施方案：

1.三级项目题目发布在AI Studio平台《Python机器学习（2021）》课程的“**比赛**”一栏中。

2.根据教师指定好的选题，每组成员进入相应的题目中进行三级项目练习。

3.每组的每个成员均可无限次提交代码。

4.实时提交，平台实时显示该三级项目题目的结果排名。

5.禁止抄袭，一经发现，取消小组全体成员成绩。

四、三级项目答辩

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.答辩形式

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.答辩时长

每个小组答辩时间不超过 10 分钟：

- 学生讲解和演示总时长不超过 6 分钟；
- 评审提问问答不超过 4 分钟。

3.提交材料

每个小组分别提交以下材料：一份三级项目报告、一份答辩 PPT、一份完整的项目。

（1）每个文件的命名格式：班级-小组-选题，如：18-5-第 2 组-猫十二分类。

（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《2021Python 机器学习实验课》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答辩时，提交纸质版三级项目报告给答辩验收老师。如果撰写不符合要求，老师有权要求学生修改并重新打印提交。

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