机器学习导论习题六

编程大作业

2024 年 6 月 26 日

作业注意事项

1. 作业所需的 Python 环境配置要求请参考[Link];
2. 请在 IPython Notebook 第一个单元格中填写个人的学号, 姓名, 邮箱;
3. 本次作业需提交的文件与对应的命名方式为:
   (a) 编程题代码文件 — `UCI-HAR.ipynb`;
   (b) 由 (b) 导出得到的 HTML 文件 — `UCI-HAR.html`;
   (c) Kaggle 比赛代码文件 — `Kaggle.ipynb`;
   (d) 由 (c) 导出得到的 HTML 文件 — `Kaggle.html`;

   请确保 IPython Notebook 和 HTML 都包含运行代码输出和你的回答; 请将以上文件打包为 "`学号_姓名.zip`" (例如 "`221300001_张三.zip`") 后提交; **注意不要提交数据集**;
4. 若多次提交作业, 则在命名 zip 文件时加上版本号, 例如 "`221300001_张三_v1.zip`" (批改时以版本号最高, 提交时间最新的文件为准);
5. 本次作业提交截止时间为 **7 月 10 日 23:59:59**. 未按照要求提交作业, 提交作业格式不正确, 或作业命名不规范, 将会被扣除部分作业分数; 除特殊原因 (如因病缓交, 需出示医院假条) 逾期未交作业, 本次作业记 0 分; **如发现抄袭, 抄袭和被抄袭双方成绩全部取消**;
6. 学习过程中, 允许参考 ChatGPT 等生成式语言模型的生成结果, 但必须在可信的信息源处核实生成结果的真实性; **不允许直接使用模型的生成结果作为作业的内容, 否则将视为作业非本人完成并取消成绩**;
7. 本次作业提交地址为[Link], 请大家预留时间提前上交, 以防在临近截止日期时, 因网络等原因无法按时提交作业.
8. 本次作业卷面总分为 **110 分**, 超出 100 分的部分**正常计入平时成绩**。

作业内容介绍

(i) **UCI-HAR 数据集**　UCI-HAR 数据集收集自三星 Galaxy S II 智能手机的传感器, 30 名 19 至 48 岁的志愿者按照要求执行六种动作. 论文文件链接为[Link], 数据集网页链接为[Link]. 原始标签从 1 到 6 依次是: 步行, 上楼, 下楼, 坐着, 站立, 躺下; 原始特征 561 维, 由 33 物理量的各自的 17 个统计量构成, 物理量包括三个方向的加速度及其傅里叶变换等, 统计量包括均值和方差等.

(ii) **运行环境**　本次作业需要的第三方库详见 `./requirements.txt`, 其中 `torch` 只需安装 CPU 版本. 可以使用 Jupyter Lab, 或者 VS Code 和 PyCharm 等 IDE 调试和运行 IPython Notebook 文件. **请在 IPython Notebook 的 Markdown 单元里回答问题和汇报结果; 你需要固定随机数种子以确保实验结果可以复现.**

(iii) **Kaggle 比赛**　Kaggle 比赛的数据集包括三部分, 训练集, 公榜测试集和私榜测试集. 参赛者只能获取训练集数据的标签, 需要据此训练模型, 最终提交模型预测的测试集的标签. 比赛结束前, 参赛者只能看到自己提交的结果在公榜测试集上的准确率; 比赛结束后, 最终成绩以私榜测试集上的准确率为准. 注意, Kaggle 平台限制单日提交次数.

(iv) **MindSpore** MindSpore 是由华为于 2019 年开源的人工智能计算框架, 同时兼容 CPU, 支持 CUDA 的 GPU 和华为自研 Ascend 芯片. 本次作业的附加题 2(4) 将把 `torch` 实现的神经网络模型迁移到 `mindspore` 上. `mindspore` 仅用于附加题 2(4), 安装参考链接[Link], 推荐使用版本 `3.8.*` 的 `python`.

(v) **学件市场**　一个训练好的机器学习模型及其**规约 (specification)** 称作**学件 (learnware)**. 规约能够让用户在不获得训练数据和模型权重的情况下得知哪些模型可以满足用户需求. **学件市场 (learnware market)** 中有大量的学件, 用户只需生成自身需求对应的规约, 并把规约提交给学件市场, 学件市场就能自动返回能够复用于用户需求的模型. **北冥坞**是学件市场的开源实现, 提供了上传学件和查搜学件的接口. 本次作业的附加题 6 将在北冥坞上传学件, 并从北冥坞中获得可用于 UCI-HAR 的模型. 北冥坞主页链接[Link], 北冥坞文档链接[Link].

1 [20pts] 处理数据

(1) **[5pts] 加载数据** 从 `./data/` 中加载数据, 特征数据加载为 `np.float64`, 标签数据加载为 `np.int64`. 注意, 原始数据的标签从 `1` 开始, 你需要转换成从 `0` 开始.

(2) **[5pts] 检查数据** 分析并回答如下问题:

- 数据中是否存在缺失值?
- 是否存在类别不平衡的问题?
- 数据属性取值是否需要归一化?

(3) **[5pts] 可视化属性分布** 填充缺失值并且归一化属性取值之后, 选择方差最大的特征, 绘制小提琴图, 可视化对比各个类别的样本在该属性上取值分布. 绘图参考图1.

(4) **[5pts] 可视化类别相关性** 绘制热力图, 可视化前 51 个属性两两之间的 Pearson 相关系数. 绘图参考图2.

2 [15pts + 附加 5pts] 分类模型

(1) **[5pts] 调用 `sklearn` 实现基线模型** 固定超参数, 汇报如下基线模型的运行时间和准确率: k 近邻, 高斯核支持向量机 (高斯核又称径向基核), 随机森林.

(2) **[5pts] 调用 `xgboost` 实现 Boosting 模型** 固定超参数, 汇报 `xgboost` 的运行时间和准确率.

(3) **[5pts] 基于 `torch` 训练神经网络模型** 每遍历一轮训练数据, 就在 `./ckpt/` 中保存当前模型权重, 固定超参数, 绘制神经网络在训练集和测试集上的准确率随训练轮数变化的折线图. 绘图参考图3.

(4) **[附加 5pts] 基于 `mindspore` 训练神经网络模型** 使用国产化软件复现 (3) 的结果, 并比较二者在效率等方面的差异.

3 [15pts] 参数调优

(1) **[5pts] 5 折交叉验证** 调用 `sklearn` 实现, 为 k 近邻选择最优的邻居数量 k, 汇报在训练集上选出的 k 及其 5 折交叉验证准确率, $k \in \{1, \cdots, 16\}$.

(2) **[5pts] 多进程并行加速** 为高斯核支持向量机选择最优的正则化系数 C, 汇报在训练集上选出的 C 及其 5 折交叉验证准确率, 同时汇报使用多进程并行加速后的总用时, $C \in \{0.01, 0.1, 1.0, 10.0, 100.0\}$.

(3) **[5pts] 搜索超参数** 使用 `optuna` 搜索 `xgboost` 的超参数, 汇报在训练集上选出的超参数及其 5 折交叉验证准确率 (更换随机数种子不低于 93.0%).

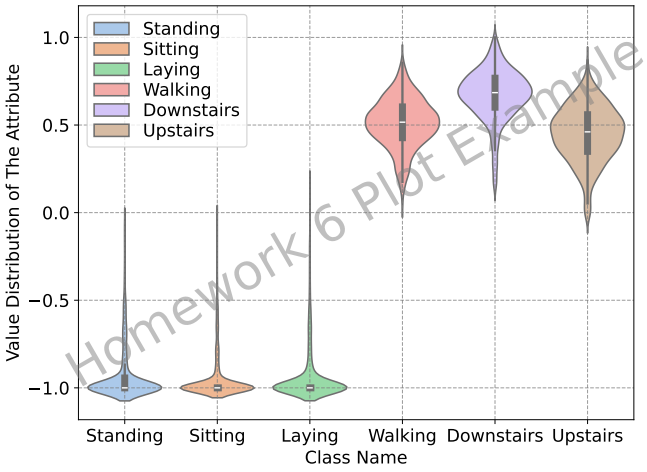

图 1: 小提琴图示例

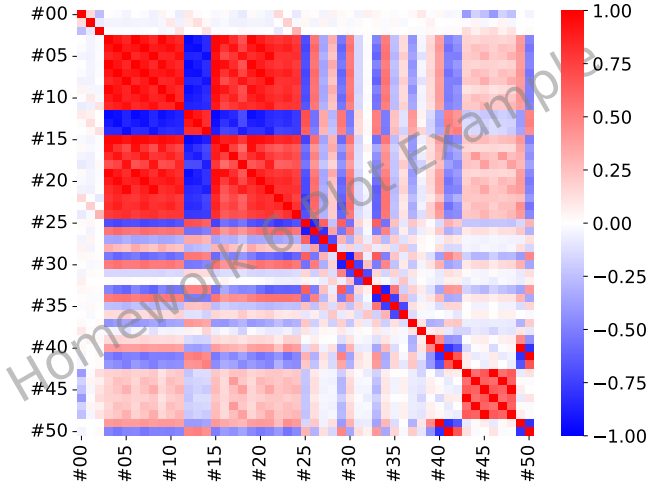

图 2: 热力图示例

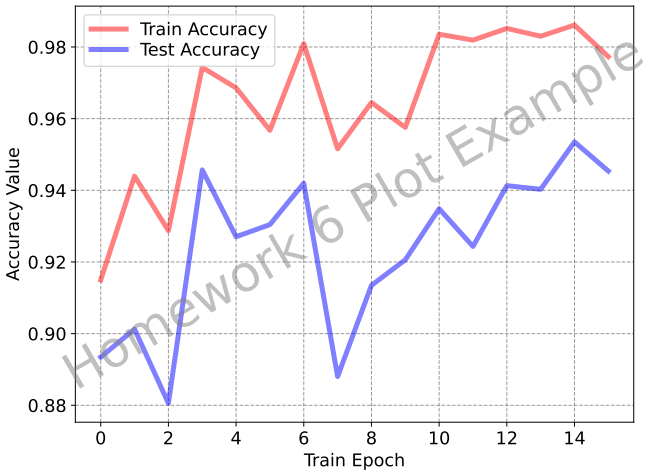

图 3: 折线图示例

4 [15pts] 集成模型

(1) **[5pts] 简单多数投票** 采用简单多数投票法集成之前题目调过参的分类器, 至少包括: k 近邻, 高斯核支持向量机, `xgboost`. 汇报测试集上准确率的提升.

(2) **[5pts] Stacking** 改用 Stacking 集成上一问中的分类器, 通过 5 折交叉验证训练 Stacking 模型. 汇报测试集上准确率的提升.

(3) **[5pts] 探索其他集成方式** 以下方式任选其一: 把神经网络最后一个隐层的输出作为新的特征, 训练一个根据样本决定采用哪个模型的路由模型, 或者提出你自己的集成方式并给出清晰的说明. 汇报测试集上准确率的提升.

5 [35pts] Kaggle 比赛

(1) **[10pts] 成功参赛** 成功提交私榜不低于 Baseline 1 的预测结果, **队伍名称必须包含学号**. 比赛链接为[Link], 比赛时间为 **6 月 25 日至 7 月 10 日**.

(2) **[25pts] 比赛排名** 预测结果在私榜不低于 Baseline 2 或者 Baseline 3 任一的前提下, 按照下式赋分: $10 + 15 \times \left(1 - \frac{\text{你的排名}-1}{\text{达到 baseline 的总人数}}\right)$. **请确保 `Kaggle.ipynb` 能够复现你的结果.**

6 [附加 5pts] 学件市场

将你在 UCI-HAR 上的最优模型上传学件市场, 并展示你生成的规约查搜到的学件. **并在 IPython Notebook 和 HTML 中给出北冥坞的注册邮箱和上传的学件的 ID, 并附上查搜的截图**.