[Problem 4] 编程题说明

在本题中，我们尝试使用 AdaBoost 与 Random Forest 这两种经典的集成学习的方法进行分类任务。

本次实验使用的数据集为 UCI 数据集 Adult，是一个二分类数据集，具体信息可见：Link

为了便于使用，我们已经将下载好的数据集划分为训练集与测试集，放置在 adult_dataset 文件夹中，其中：

- 训练集包含 32561 条样本，测试集包含 16281 条样本
- 每条样本的特征为 14 维，标签为 0/1 值

由于 Adult 是一个类别不平衡数据集，我们选用 AUC 作为分类器性能的评价指标，允许调用 sklearn 中进行 AUC 指标的计算。

我们提供如下的三个脚本模板，你可以在模板的基础上进行编程：

1. `AdaBoost.py`：AdaBoost 分类器的实现
2. `RandomForest.py`：RandomForest 分类器的实现
3. `main.py`：执行入口脚本，调用两种分类器进行数据集的分类与测试

[4.1 - 10 pts] 实现 AdaBoost 与 Random Forest

请参考《机器学习》中对 AdaBoost 与 Random Forest 的介绍，实现 AdaBoost 分类器与 Random Forest 分类器。**具体的，你需要在代码模板中实现 `AdaBoostClassifier` 与 `RandomForestClassifier` 的 `fit(X, y)` 与 `predict_proba(X)` 方法。**

请使用**决策树分类器**作为这两种方法的基分类器。你可以直接调用 sklearn 中的决策树分类器实现，也可以手动实现你自己的版本。

> **在实现过程中，你可能需要注意以下几点：**
>
> 1. 为了减小计算量，突出集成学习方法对模型性能的提升作用，**在本次作业中，我们限制所有决策树基学习器的最大深度不超过 4。**请确保你的代码满足该要求。
> 2. Adult 中使用的是 `0/1` 标签，而《机器学习》中描述 AdaBoost 算法时采用的是 `-1/+1` 标签。你可能需要对此进行必要的适配。
> 3. 根据《机器学习》中的描述，Random Forest 的基决策树每次会先从属性集合中随机选择一个大小为 k 的子集，再从子集中选取一个最优属性作为划分。请确保你的代码满足该要求。

我们将会使用如下的代码测试你的实现：

```
from RandomForest import RandomForestClassifier
from AdaBoost import AdaBoostClassifier

X_train, y_train, X_test, y_test = load_dataset()

rf_clf = RandomForestClassifier(T=10)
result = rf_clf.evaluate(X_train, y_train, X_test, y_test)

ad_clf = AdaBoostClassifier(T=10)
result = ad_clf.evaluate(X_train, y_train, X_test, y_test)
```

只要你的实现正确，通过我们的精度测试，即可获得本题的全部分数。

[4.2 - 10 pts] 模型评估与超参数调整

请结合上述 AdaBoost 与 Random Forest 的实现，研究基学习器数量对分类器训练效果的影响。

请在 `main.py` 的 `make_plot(X_train, y_train)` 函数中实现以下功能：

1. 分别使用 AdaBoost 与 Random Forest 分类器，设置基分类器数量为 1~20，利用 5 折交叉验证得到分类器在训练数据集上的 AUC 指标。
2. 绘制上述 AUC 指标的折线图，其中横轴为基分类器的数量（1~20），纵轴为对应情况下 5 折交叉验证的平均 AUC 指标。图中应当有两条折线，分别对应 AdaBoost 与 Random Forest 分类器，且标注清晰横、纵轴的含义及两条折线分别表示哪种模型。
    - 请确保执行 `make_plot(X_train, y_train)` 函数后，会将生成的图像保存为当前目录下的 `evaluation.png` 文件。

> **在本题的实现过程中，你需要注意以下几点：**
>
> 1. 你可以调用 sklearn 中的方法来减小你的工作量，包括但**不限于**：`cross_val_score`，`KFold` 等
> 2. 由于 Adult 是一个类别不平衡数据集，在使用交叉验证划分 KFold 时，你可能需要注意使用分层采样（stratified sampling），确保各个 Fold 中含有正/负样本的比例与原始分布中基本一致。
> 3. 我们会执行你的 `make_plot(X_train, y_train)` 函数，检查生成的图像是否与实验报告中的一致。请确保你的代码可以正常执行。

在解答的 PDF 文件中，请汇报以下内容：

1. 绘制得到的折线图 `evaluation.png`
2. 简要分析基分类器数量与两种分类器分类效果的关系（几句话即可）

[4.3 - 5pts] 模型测试

基于上一小题的超参数搜索结果，对 `AdaBoostClassifier` 与 `RandomForestClassifier` 选取最好的基分类器数目，并在测试集上评估模型的表现。

请在实验报告中分别汇报这两种模型的最优超参数（基分类器数量），以及在训练集上训练后，在测试集上测试的 AUC 指标。