作业 3：决策树与提升算法

清华大学软件学院

机器学习, 2023 年秋季学期

1 介绍

本次作业需要提交说明文档（PDF 形式）。注意事项如下：

- 本次作业总分值为 110 分，若得分超过 100 分，则按照 100 分截断。
- 作业按点给分，因此请在说明文档中按点回答，方便助教批改。
- 友情提示：每个算法的主要代码已经实现，因此每一小题的代码都不大（不超过 10 行）。
- 不要使用他人的作业，也不要向他人公开自己的作业，否则处罚很严厉，会扣至-100（倒扣本次作业的全部分值）。
- 统一文件的命名：{学号}_{姓名}_hw3.zip

2 决策树与随机森林（30pt）

2.1 ID3 算法无法找到最优解的一个情形 (5pt)

考虑如下的训练集，其中 $\mathcal{X} = \{0,1\}^3$，$\mathcal{Y} = \{0,1\}$:

$$((1,1,1),1)$$
$$((1,0,0),1)$$
$$((1,1,0),0)$$
$$((0,0,1),0)$$

我们使用该训练集构建一棵深度为 2 的决策树 (也即，对于每一个输入向量，我们根据两个维度上的特征来给出其标签)。

1. 假设我们使用 ID3 算法为题中给出的训练集构建一棵深度为 2 的决策树。我们每次选取用于划分当前节点的特征时，使用信息增益 (Information Gain) 作为标准，且当两个特征的信

息增益一致时，随机选取其中一个特征用于划分节点。证明使用 ID3 算法得到的决策树至少有 $\frac{1}{4}$ 的训练误差。

2. 给出一棵深度为 2 的且训练误差为 0 的决策树。

2.2 随机森林 (5pt)

有 n 个特征为 d 维的样本，构建随机森林对其进行分类，随机森林中有 t 个二叉决策树，每个决策树有 h 个内部节点，使用 m 个自助采样（Bootstrap）得到的样本，Breiman 算法每次选择的特征数 $K = 1$，请分别简要说明：

1. 任意一个特定的特征从未被选中分割的概率。
2. 任意一个特定的样本从未在任何一棵树中被考虑的概率。

2.3 代码实验 (20pt)

在本题中，你将使用决策树解决二分类问题和回归问题。

1. 补全 tree.py 中 DecisionTree 类的 fit 函数。提示：递归调用决策树的构造与 fit 函数。
2. 根据决策树熵的定义，完成 tree.py 中 compute_entropy 函数。
3. 根据基尼系数的定义，完成 tree.py 中 compute_gini 函数。
4. 完成 tree.py 中 mean_absolute_deviation_around_median 函数。

5. 运行 tree.py，在实验文档中记录决策树在不同数据集上运行的结果，包括

   (a) DT_entropy.pdf，使用决策树在二分类问题上的结果。

   (b) DT_regression.pdf，使用决策树在回归问题上的结果。

   并简要描述实验现象（例如超参数对于决策树的影响）。

3 提升算法（80pt）

3.1 噪音不敏感的 AdaBoost 算法 (10pt)

AdaBoost 在存在噪声的情况下可能会过度拟合，部分原因是对错误分类样本施加了较高的惩罚。为了减小这种影响，可以使用以下目标函数。

$$F = \frac{1}{m}\sum_{i=1}^{m} G(-y_i f(x_i))$$

其中函数 G 为

$$G(x) = \begin{cases} e^x & \text{if } x \leq 0 \\ x+1 & \text{if } x > 0 \end{cases}$$

1. 证明函数 G 是 $\mathbf{R}$ 上的凸函数且处处可导。
2. 对 AdaBoost 算法而言，$F(\bar{\alpha}) = \frac{1}{m}\sum_{i=1}^{m} e^{-y_i f(x_i)} = \frac{1}{m}\sum_{i=1}^{m} e^{-y_i \sum_{j=1}^{N} \bar{\alpha}_j h_j(x_i)}$。课件中已经证明了，AdaBoost 的执行过程等价于使用坐标下降 (Coordinate Descent) 法优化该函数。对于题中给出的目标函数，也即 $F(\bar{\alpha}) = \frac{1}{m}\sum_{i=1}^{m} G(-y_i f(x_i)) = \frac{1}{m}\sum_{i=1}^{m} G(-y_i \sum_{j=1}^{N} \bar{\alpha}_j h_j(x_i))$，同样使用坐标下降法优化该函数，类比 AdaBoost 算法，请给出第 t 步时弱分类器 h_t 应当优化的损失函数 ϵ_t 的表达式 (假设初始化时，所有样本的权重系数相同)。

3.2 探究 AdaBoost 算法是否能使用完全相同的弱分类器 (10pt)

在第 $t+1$ 轮 Adaboost 算法的迭代中，样本 (x_i, y_i) 的权重值 $D_{t+1}(i)$ 取决于前 t 个弱分类器的准确率，$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(\boldsymbol{x}_i))}{Z_t}$，其中 α_t 表示弱分类器 h_t 的权重，$Z_t = \sum_i D_t(i)\exp(-\alpha_t y_i h_t(\boldsymbol{x}_i))$ 表示归一化因子。

试证明 $\sum_i D_{t+1}(i)\mathbb{1}_{[y_i \neq h_t(x_i)]} = \frac{1}{2}$。并由此证明，第 $t+1$ 步选取的弱分类器 h_{t+1} 不会与 h_t 相同。

3.3 AdaBoost 的训练误差 (10pt)

给定包含 m 条数据的训练集，假设 AdaBoost 算法中，基分类器 h_t 的误差 ϵ_t 的上界为 $1/2 - \gamma$, 其中 $\gamma > 0$。请证明当 $T > \dfrac{\log m}{2\gamma^2}$ 时，AdaBoost 的训练误差可以达到 0。

3.4 简化版本的 AdaBoost(10pt)

假设弱学习条件成立，即对于某个已知的正数 γ，有每个弱分类器的训练误差 $\epsilon_t = \Pr_{i\sim D_t}[h_t(x_i) \neq y_i] \leq 1/2 - \gamma$，这个条件在 Boosting 算法**开始前**已知。假设我们修改 AdaBoost 算法的执行过程，将每一轮迭代中得到的弱分类器的权重系数固定为一个常数 $\alpha_t \equiv \alpha$, 那么算法最终得到的分类器是未加权的基分类器的多数投票 $f(x_i) = \mathrm{sgn}(\sum_{i=1}^{T} \alpha h_t(x_i))$。试找出一个合理的 α 的取值，不改变算法的其余部分，使得最终得到的分类器的训练误差不多于 $(1-4\gamma^2)^{T/2}$。

3.5 Gradient Boosting Machines(40pt)

总结课件中的 Gradient Boosting Machine 的算法流程如下：

1. 令 $f_0(\mathbf{x}) = 0$。
2. For t=1 to T：
   (a) 计算在各个数据点上的梯度 $\mathbf{g}_t = \left(\frac{\partial}{\partial \widehat{\mathbf{y}}_i}\ell(\mathbf{y}_i, \widehat{\mathbf{y}}_i)|_{\widehat{\mathbf{y}}_i = f_{t-1}(\mathbf{x}_i)}\right)_{i=1}^{n}$。
   (b) 根据 $-\mathbf{g}_t$ 拟合一个回归模型，$h_t = \arg\min_{h\in\mathcal{F}}$ _____。
   (c) 选择合适的步长 α_t, 最简单的选择是固定步长 $\eta \in (0, 1]$。
   (d) 更新模型，$f_t(\mathbf{x}) =$ _____。

请完成以下题目：

1. 完成上述算法中的填空。
2. 考虑回归问题，假设损失函数 $\ell(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{1}{2}(\mathbf{y} - \widehat{\mathbf{y}})^2$。直接给出第 t 轮迭代时的 $\mathbf{g}_t$ 以及 h_t 的表达式。(使用 f_{t-1} 表达)。

3. 考虑二分类问题，假设损失函数 $\ell(\mathbf{y}, \hat{\mathbf{y}}) = \ln(1 + e^{-\mathbf{y}\hat{\mathbf{y}}})$。直接给出第 t 轮迭代时的 $\mathbf{g}_t$ 以及 h_t 的表达式。(使用 f_{t-1} 表达)。

4. 完成 boosting.py 中 GradientBoosting 类的 fit 函数。
5. 完成 boosting.py 中 GradientBoosting 类的 predict 函数。
6. 完成 boosting.py 中函数 gradient_logistic。
7. 运行 boosting.py，在实验文档中记录 GBM 在不同数据集上运行的结果，包括
   (a) GBM_l2.pdf，使用 L2 loss 在二分类问题上的结果。
   (b) GBM_logistic.pdf，使用 logistic loss 在二分类问题上的结果。
   (c) GBM_regression.pdf，使用 L2 loss 在回归问题上的结果。

   并简要描述实验现象（例如超参数对于 GBM 的影响、损失函数对于 GBM 的影响等)。