作业 3：决策树与提升算法

清华大学软件学院

机器学习, 2024 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量都不超过 10 行。
- 若题目没有特殊要求，可以直接使用课上已经证明的结论，但请注明出处。
- 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（40pt）

2.1 ID3 决策树算法（15pt）

1. 考虑下表中的数据集，该数据集记录了一家人是否外出的情况，包含了天气是否是晴天与地面是否有雪两个二值特征，以及特征和标签的组合在数据集中出现的次数。

是否外出	是否晴天	次数
是	是	28
是	否	7
否	是	6
否	否	9

是否外出	是否有雪	次数
是	是	16
是	否	19
否	是	2
否	否	13

虽然表中没有列出数据集中的具体实例，但已经足以构建一棵 ID3 决策树的根节点（决策树桩）。请通过计算说明应该如何构建。

2. 熵是决策树中用于决定分裂特征的常用统计指标之一。在本题中，你需要使用另一种统计指标为下表的气球数据集构建一棵决策树。

颜色	大小	动作	年龄	是否充气	次数
红色	大	拉伸	成人	否	6
红色	大	拉伸	儿童	是	5
红色	大	下压	儿童	否	2
蓝色	大	下压	成人	是	2
蓝色	大	拉伸	儿童	否	1
蓝色	大	下压	儿童	否	2
红色	小	下压	儿童	否	2
蓝色	小	下压	成人	是	4

该数据集包含 4 个二值特征（颜色、大小、动作、年龄）以及一个二值标签（是否充气）。

这次你需要使用一种新的统计指标，称为最小错误（MinError），定义如下：

$$MinError(S) = \min\{p_-, p_+\}$$

最小错误对决策树某节点中的剩余实例进行计算。例如，如果某节点中有 10 个实例，其中 8 个是正类，则 $MinError = \min\{0.2, 0.8\} = 0.2$ 。与熵类似，最小错误也可以用来量化节点的“纯净度”。

使用最小错误替换 ID3 算法信息增益公式中的熵函数，构建一棵完整的决策树（展现过程）。

3. ID3 算法是否总能保证生成一颗“最优”的决策树？这里，“最优”的定义是：决策树能够最好的拟合训练数据，且深度最小。若能，请说明原因；若不能，请举出反例。

2.2 代码实验 (25pt)

在本题中，你将使用决策树解决二分类问题和回归问题。

1. 根据熵的定义，完成 tree.py 中 compute_entropy 函数。
2. 根据基尼系数的定义，完成 tree.py 中 compute_gini 函数。
3. 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。
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 (60pt+10pt)

3.1 弱分类器的更新保证 (10pt)

AdaBoost 算法中，每轮迭代时分布 D 会按照如下方式更新：

$$\alpha_t \leftarrow \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{1}$$

$$Z_t \leftarrow 2[\epsilon_t(1-\epsilon_t)]^{\frac{1}{2}} \tag{2}$$

$$D_{t+1}(i) \leftarrow \frac{D_t(i)\exp(-\alpha_t y_i h_t(\boldsymbol{x}_i))}{Z_t} \tag{3}$$

其中，$\epsilon_t = Pr_{i\sim D_t}[h_t(\boldsymbol{x}_i) \neq y_i]$。这样的更新机制能够让第 t 轮中被错误分类的数据在 $t+1$ 轮拥有更大的权重，从而让下一轮的学习器更关注这些数据。为了更好地理解这一点，请完成以下题目：

1. 通过计算证明：$Z_t \doteq \sum_{i=1}^{n} D_t(i)\exp(-\alpha_t y_i h_t(\boldsymbol{x}_i)) = 2[\epsilon_t(1-\epsilon_t)]^{\frac{1}{2}}$。
2. 证明：h_t 关于分布 D_{t+1} 的错误率正好为 $1/2$，即对任意 $1 \leq t < T$：

$$\sum_{i=1}^{n} D_{t+1}(i)\mathbb{1}_{[y_i \neq h_t(\boldsymbol{x}_i)]} = 1/2 \tag{4}$$

并据此说明，对任意 $1 \leq t < T$，$t+1$ 步选取的弱分类器 h_{t+1} 不会与 h_t 相同。

3.2 替换目标函数 (10pt)

AdaBoost 中使用了指数函数作为目标函数，我们也可以将其替换成其它函数。假设函数 $\boldsymbol{\phi}: \mathbf{R} \to \mathbf{R}$ 是单调递增且处处可微的凸函数，并且满足 $\forall x \geq 0, \boldsymbol{\phi}(x) \geq 1$ 且 $\forall x < 0, \boldsymbol{\phi}(x) > 0$，我们定义新的优化目标为 $L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \boldsymbol{\phi}(-y_i f(\boldsymbol{x}_i))$，其中 f 仍然定义为 $f = \sum_{t=1}^{T} \alpha_t h_t$。我们可以得到：

$$\left.\frac{dL(\boldsymbol{\alpha}+\beta\boldsymbol{e}_t)}{d\beta}\right|_{\beta=0} = -\sum_{i=1}^{n} y_i h_t(\boldsymbol{x}_i)\boldsymbol{\phi}'(-y_i f_t(\boldsymbol{x}_i)) \tag{5}$$

$$\propto -\sum_{i=1}^{n} y_i h_t(\boldsymbol{x}_i)\frac{\boldsymbol{\phi}'(-y_i f_t(\boldsymbol{x}_i))}{Z_t} \tag{6}$$

$$\propto -\sum_{i=1}^{n} y_i h_t(\boldsymbol{x}_i) D_t(i) \tag{7}$$

$$= 2\epsilon_t - 1 \tag{8}$$

其中，$D_t(i) = \frac{\boldsymbol{\phi}'(-y_i f_t(\boldsymbol{x}_i))}{Z_t}$。可以看到，若和原始的 AdaBoost 算法类似地使用坐标下降法优化新的目标函数 $L(\boldsymbol{\alpha})$，每轮的弱学习器仍然最小化分布 D_t 下的期望损失 $\epsilon_t = Pr_{i\sim D_t}[h_t(\boldsymbol{x}_i) \neq y_i]$。基于以上分析，回答下面的题目：

1. 考虑如下的函数：(1) $\boldsymbol{\phi}_1(x) = \mathbb{1}_{x\geq 0}$; (2) $\boldsymbol{\phi}_2(x) = (1+x)^2$; (3) $\boldsymbol{\phi}_3(x) = \max\{0, 1+x\}$; (4) $\boldsymbol{\phi}_4(x) = log_2(1+e^x)$。哪一个函数满足题面中对于 $\boldsymbol{\phi}$ 的假设条件？请求出使用该函数时，$D_t(i)$ 的表达式。
2. 上文的分析只确定了坐标下降中每次选择的最优坐标方向；对于前一问中选择的函数 $\boldsymbol{\phi}$，请通过求解 $\frac{dL(\boldsymbol{\alpha}+\beta \boldsymbol{e}_t)}{d\beta} = 0$ ，进一步确定最优步长 β。（写出 β 应满足的方程并适当化简即可，无需求解方程）

3.3 带未知标签的 Boosting 算法（附加题，10pt）

在实际实践中，有些情况下我们收集到的训练数据中可能包含未标记的数据。对于这种情形，我们考虑一个二分类问题的变种：数据标签的取值除了 $+1$ 和 -1 外，还可以取 0 表示数据分类未知。我们仍然定义函数 $f : \mathcal{X} \to \mathbf{R}$ 对于数据 $(x, y) \in \mathcal{X} \times \mathcal{Y}$ 的损失函数为 $\mathbb{1}_{yf(x)<0}$，其中 $\mathcal{X} \subset \mathbf{R}^d$, $\mathcal{Y} = \{-1, 0, +1\}$。

此时，我们考虑在这个问题上构造一个类似 Adaboost 的算法。设训练数据为 $S = ((\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, v_n)) \in (\mathcal{X} \times \mathcal{Y})^n$，假设集合 H 为一系列 $\mathcal{X}$ 到 $\mathcal{Y}$ 的映射；对于任意 $h_t \in H$ 和分布 D_t，定义 $\epsilon_t^s = \mathbb{E}_{i\sim D_t}[\mathbb{1}_{y_i h_t(\boldsymbol{x}_i)=s}]$，其中 s 可以取 -1, 0, $+1$（简记为 $\epsilon_t^-, \epsilon_t^0, \epsilon_t^+$）。

我们假设算法的其它细节都与 AdaBoost 类似。例如，目标函数仍为 $F = \frac{1}{m}\sum_{i=1}^n \exp(-y_i f(\boldsymbol{x}_i))$, D_t, Z_t 的计算方式也与 AdaBoost 一致。每一个弱学习器也会有一个权重 α_t，最终的假设为 $f(\boldsymbol{x}) = \sum_t \alpha_t h_t(\boldsymbol{x})$。

在上述假设下，我们尝试通过坐标下降的视角，推导出一个**带未知标签的 Boosting 算法**。请完成以下题目：

1. 用 ϵ_t^s 和 α_t 表示 Z_t。
2. 计算 $F'(\bar{\alpha}_{t-1}, \boldsymbol{e}_k)$，并指出第 t 步时弱分类器的优化目标。（结果用含 ϵ_t^s 的表达式表示）
3. 解 $\frac{\partial F\bar{\alpha}_{t-1}+\eta \boldsymbol{e}_k)}{\partial \eta} = 0$，并给出 α_t 的更新公式。（结果用含 ϵ_t^s 的表达式表示）
4. 在 AdaBoost 中，我们证明了训练误差 $\hat{\mathcal{E}} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{y_i f(\boldsymbol{x}_i)<0} \leq \prod_{t=1}^T Z_t \leq \prod_{t=1}^T 2\sqrt{\epsilon_t(1-\epsilon_t)}$。类似地，请求出带未知标签的 Boosting 算法中，用含 ϵ_t^s 的表达式表示的训练误差上界，并证明若每个弱学习器的误差满足 $\frac{\epsilon_t^+ - \epsilon_t^-}{\sqrt{1-\epsilon_t^0}} \geq \gamma > 0$，则有：

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{y_i f(\boldsymbol{x}_i)<0} \leq \exp\left(-\frac{\gamma^2 T}{2}\right)$$

其中，T 是弱学习器的个数。

提示：

(a) 当 $\lambda \in [0, 1]$，下述不等式成立：

$$\lambda\sqrt{x} + (1-\lambda)\sqrt{y} \leq \sqrt{\lambda x + (1-\lambda)y}$$

(b) $1 + x \leq \exp(x)$.

3.4 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 $\ell(\mathbf{y}, \widehat{\mathbf{y}}) = \ln(1 + e^{-\mathbf{y}\widehat{\mathbf{y}}})$。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）。