机器学习 HW3: 决策树与提升算法

姓名：聂礼昂
学号：2022012097

2024 年 12 月 4 日

1 介绍

本次作业是清华大学软件学院，机器学习, 2024 年秋季学期课程第三次作业。

2 决策树

2.1 ID3 决策树算法

2.1.1 1.

1. 计算数据集的熵

$$H(D) = -0.7\log_2(0.7) - 0.3\log_2(0.3) \approx 0.78 + 0.52 = 0.881$$

2. 计算每个特征的条件熵

晴天和非晴天的条件熵为：

$$\begin{aligned} H(D_{\text{晴天}}) &= \frac{34}{50} \times \left(-\frac{28}{34}\log_2\frac{28}{34} - \frac{6}{34}\log_2\frac{6}{34} \right) \\ &\quad + \frac{16}{50} \times \left(-\frac{7}{16}\log_2\frac{7}{16} - \frac{9}{16}\log_2\frac{9}{16} \right) \\ &= 0.457 + 0.316 \\ &= 0.773 \end{aligned}$$

有雪和无雪的条件熵为：

$$
\begin{aligned}
H(D_{\text{晴天}}) &= \frac{18}{50} \times \left(-\frac{16}{18} \log_2 \frac{16}{18} - \frac{2}{18} \log_2 \frac{2}{18} \right) \\
&\quad + \frac{32}{50} \times \left(-\frac{19}{32} \log_2 \frac{19}{32} - \frac{13}{32} \log_2 \frac{13}{32} \right) \\
&= 0.181 + 0.624 \\
&= 0.805
\end{aligned}
$$

3. 计算信息增益

   对“天气是否晴天”特征的信息增益

$$
IG(\text{晴天}) = H(D) - H(D_{\text{晴天}}) = 0.881 - 0.773 = 0.108
$$

   对“是否有雪”特征的信息增益

$$
IG(\text{有雪}) = H(D) - H(D_{\text{有雪}}) = 0.881 - 0.805 = 0.076
$$

4. 选择信息增益最大特征

   根据计算的结果，“天气是否晴天”的信息增益是 0.108，而“是否有雪”的信息增益是 0.076。因此，ID3 决策树的根节点应该选择 **“天气是否晴天”** 作为划分特征。

2.1.2 2.

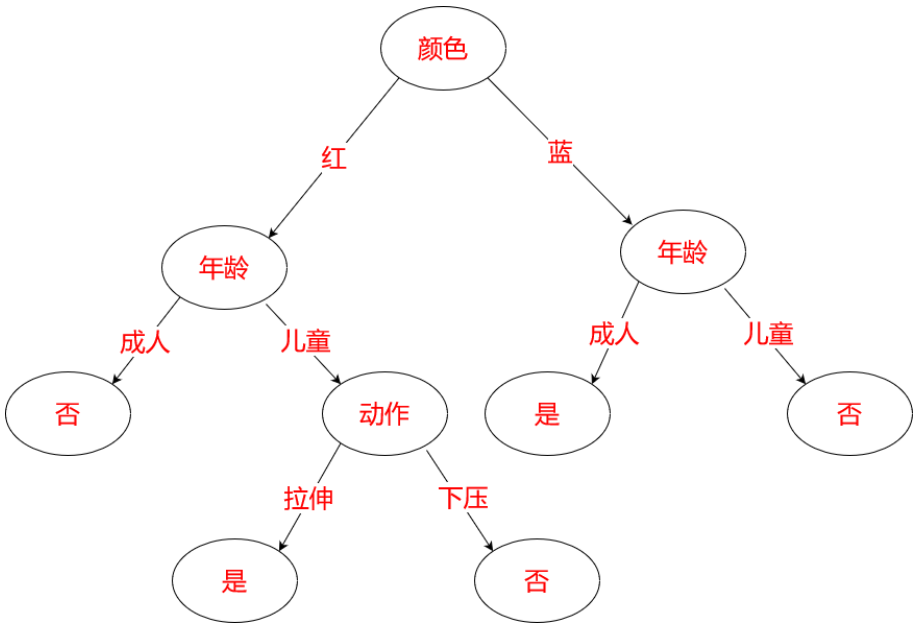

图 1: ld3 tree with MinError

1. 第一步：计算根节点的 *MinError*

$$MinError(\text{根节点}) = \min\left(\frac{11}{24}, \frac{13}{24}\right) = \frac{11}{24} = 0.458$$

2. 第二步：按特征划分，计算每个子节点的 *MinError*

   (a) 按“颜色”划分

$$MinError(\text{颜色}) = \frac{15}{24} \cdot \frac{5}{15} + \frac{9}{24} \cdot \frac{3}{9} = \frac{1}{3} = 0.333$$

   (b) 按“大小”划分

$$MinError(\text{大小}) = \frac{18}{24} \cdot \frac{7}{18} + \frac{6}{24} \cdot \frac{2}{6} = \frac{3}{8} = 0.375$$

   (c) 按“动作”划分

$$MinError(\text{动作}) = \frac{12}{24} \cdot \frac{5}{12} + \frac{12}{24} \cdot \frac{6}{12} = \frac{11}{24} = 0.458$$

   (d) 按“年龄”划分

$$MinError(\text{年龄}) = \frac{12}{24} \cdot \frac{6}{12} + \frac{12}{24} \cdot \frac{5}{12} = \frac{11}{24} = 0.458$$

   这四种分类的 *MinError* 都小于根节点，其中按” 颜色” 划分的 *MinError* 最小，因此选择颜色作为当前节点的分裂特征，将根节点分为红蓝两个子节点

3. 递归对子节点进行分裂

   计算第一层右节点各个特征的 *MinError*，年龄对应的 *MinError* 为 0，可以完全分开。

   计算第一层左节点各个特征的 *MinError*

$$MinError(\text{根节点}) = MinError(\text{大小}) = MinError(\text{动作}) = 0.333$$

$$MinError(\text{年龄}) = 0.267$$

   因此按年龄进行分裂。

   第二层左节点已完全分开，计算第二层右节点各个特征的 *MinError*，动作对应的 *MinError* 为 0，可以完全分开。

   因此对第二层右节点按照动作分裂。至此已完全分开，决策树构建成功。

2.1.3 ID3 算法是否能保证生成“最优”的决策树？

回答：不能保证生成最优的决策树。

1. 原因分析

   1. **局部最优选择**：ID3 算法使用贪心策略，每次划分选择信息增益最大的特征，而不考虑全局最优划分。因此，可能导致子树结构次优，无法得到整体最优树。

   2. **过拟合问题**：如果数据中存在噪声或小样本特征，ID3 可能会过度拟合训练数据，导致在测试集上的性能下降。

   3. **未考虑特征组合**：ID3 算法只能基于单一特征进行划分，无法捕捉特征之间的交互作用。

2. 反例假设数据中两个特征 A 和 B 单独来看信息增益都较低，但它们的组合（如 $A \wedge B$）可以完美划分数据。ID3 算法不会优先考虑这种组合特征，可能选择其他次优特征作为划分点，最终生成的决策树不是最优的。

2.2 代码部分

1-4 见 tree.py

5 运行结果如下

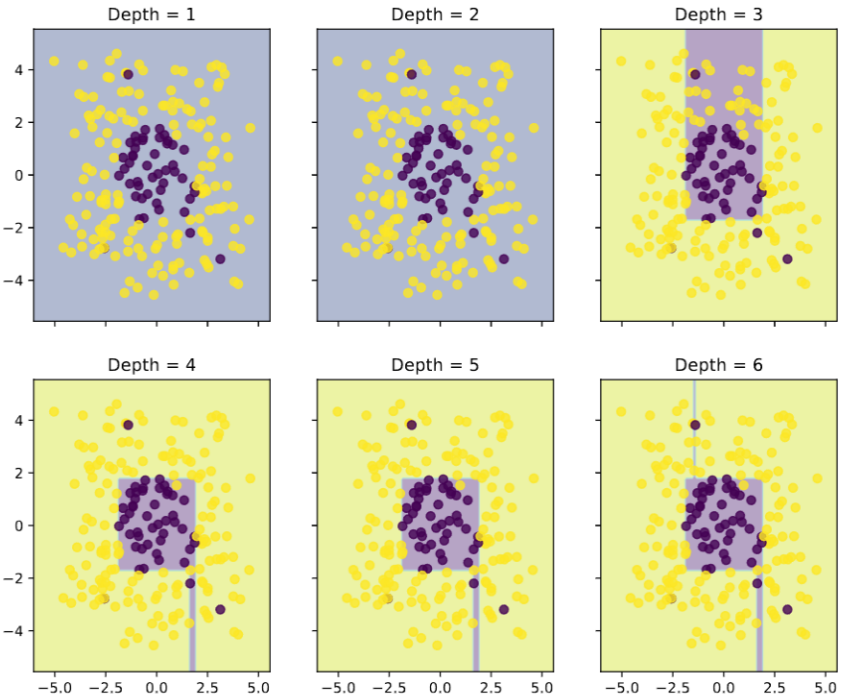

图 2: DT_entropy

通过以上两张图，可以看到

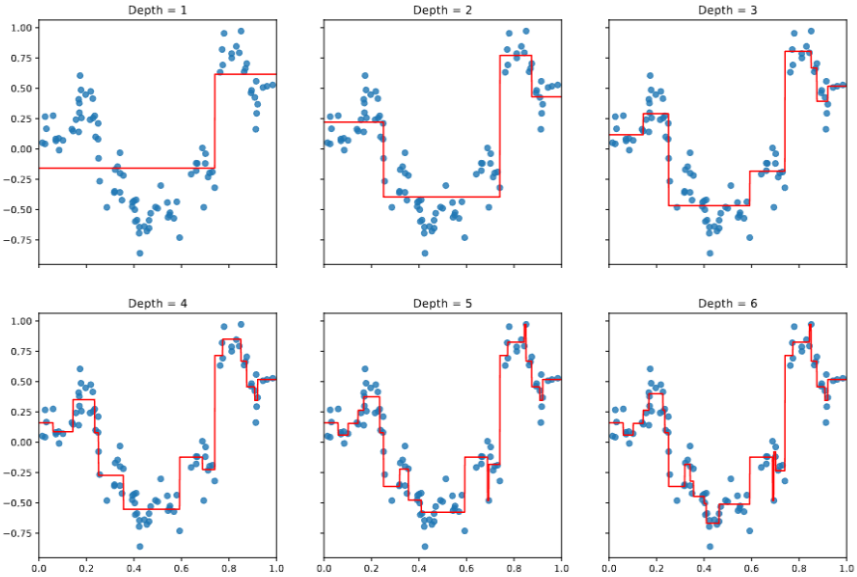

图 3: DT_regression

1. 随着树的最大深度的增大，对训练样本的拟合程度越好。
2. 限制树的最大深度可以防止过拟合，但过小的深度可能导致欠拟合
3. 较深的树可能会捕捉到数据中的细微模式，但也更容易学习到噪声。

3 提升算法

3.1 弱分类器的更新保证

问题 1: 证明 $Z_t = \sum_{i=1}^{n} D_t(i) \exp(-\alpha_t y_i h_t(x_i)) = 2\sqrt{\epsilon_t(1-\epsilon_t)}$

证明: 根据 Z_t 的定义:

$$Z_t = \sum_{i=1}^{n} D_t(i) \exp(-\alpha_t y_i h_t(x_i))$$

由于 $y_i h_t(x_i)$ 的取值为 1 或 -1，可分情况讨论:

- 当 $y_i h_t(x_i) = 1$ 时，$\exp(-\alpha_t y_i h_t(x_i)) = \exp(-\alpha_t)$;
- 当 $y_i h_t(x_i) = -1$ 时，$\exp(-\alpha_t y_i h_t(x_i)) = \exp(\alpha_t)$。

因此，Z_t 可分为两部分:

$$Z_t = \exp(-\alpha_t) \sum_{i=1}^{n} D_t(i) \mathbb{1}[h_t(x_i) = y_i] + \exp(\alpha_t) \sum_{i=1}^{n} D_t(i) \mathbb{1}[h_t(x_i) \neq y_i]$$

记 $\epsilon_t = \sum_{i=1}^{n} D_t(i) \mathbb{1}[h_t(x_i) \neq y_i]$，$1 - \epsilon_t = \sum_{i=1}^{n} D_t(i) \mathbb{1}[h_t(x_i) = y_i]$，代入后:

$$Z_t = \exp(-\alpha_t)(1-\epsilon_t) + \exp(\alpha_t)\epsilon_t$$

将 $\alpha_t = \frac{1}{2}\ln\frac{1-\epsilon_t}{\epsilon_t}$ 代入，计算 $\exp(\pm\alpha_t)$：

$$\exp(\alpha_t) = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}, \quad \exp(-\alpha_t) = \sqrt{\frac{\epsilon_t}{1-\epsilon_t}}$$

因此：

$$Z_t = \sqrt{\frac{\epsilon_t}{1-\epsilon_t}}(1-\epsilon_t) + \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\epsilon_t$$

化简后可得：

$$Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

问题 2: 证明 h_t 关于分布 D_{t+1} 的错误率为 $\frac{1}{2}$

分布 D_{t+1} 的定义为：

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

因此，h_t 关于分布 D_{t+1} 的错误率为：

$$\text{错误率} = \sum_{i=1}^{n} D_{t+1}(i)\mathbb{1}[h_t(x_i) \neq y_i]$$

代入 $D_{t+1}(i)$ 的表达式：

$$\text{错误率} = \frac{1}{Z_t}\sum_{i=1}^{n} D_t(i)\exp(-\alpha_t y_i h_t(x_i))\mathbb{1}[h_t(x_i) \neq y_i]$$

当 $h_t(x_i) \neq y_i$ 时，$\exp(-\alpha_t y_i h_t(x_i)) = \exp(\alpha_t)$，因此：

$$\text{错误率} = \frac{\exp(\alpha_t)}{Z_t}\sum_{i=1}^{n} D_t(i)\mathbb{1}[h_t(x_i) \neq y_i]$$

由于 $\epsilon_t = \sum_{i=1}^{n} D_t(i)\mathbb{1}[h_t(x_i) \neq y_i]$，可得：

$$\text{错误率} = \frac{\exp(\alpha_t)\epsilon_t}{Z_t}$$

将 $Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$ 和 $\exp(\alpha_t) = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$ 代入：

$$\text{错误率} = \frac{\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\epsilon_t}{2\sqrt{\epsilon_t(1-\epsilon_t)}}$$

化简后得到：

$$\text{错误率} = \frac{\epsilon_t(1-\epsilon_t)}{2\epsilon_t(1-\epsilon_t)} = \frac{1}{2}$$

1. $Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$；2. h_t 关于分布 D_{t+1} 的错误率为 $\frac{1}{2}$。

由此可知，$t+1$ 轮选取的弱分类器 h_{t+1} 与 h_t 不会相同，因为每一步都会基于更新的分布 D_t 重新选择最优的弱分类器，重新选出的 h_{t+1} 的在 D_{t+1} 的错误率一定会小于 $\frac{1}{2}$，因此 h_{t+1} 与 h_t 不会相同。

3.2 替换目标函数

3.2.1 确定满足条件的函数

首先，我们需要检查每个函数是否满足题面中对 ϕ 的假设条件：

- $\phi_1(x) = 1_{x\geq 0}$：这是一个指示函数，当 $x \geq 0$ 时 $\phi_1(x) = 1$，否则 $\phi_1(x) = 0$。这个函数不满足 $\forall x < 0, \phi(x) > 0$ 的条件，因为它在 $x < 0$ 时 $\phi_1(x) = 0$。

- $\phi_2(x) = (1 + x)^2$：这个函数是凸函数，且 $\forall x \geq 0, \phi_2(x) \geq 1$，但是 $\phi_2(x)$ 在 $x < 0$ 时可以小于 1，不满足 $\forall x < 0, \phi(x) > 0$ 的条件。

- $\phi_3(x) = \max\{0, 1 + x\}$：这个函数是凸函数，且 $\forall x \geq 0, \phi_3(x) \geq 1$，但是 $\forall x < -1, \phi_3(x) = 0$，不满足条件。

- $\phi_4(x) = \log_2(1+e^x)$：这个函数是凸函数，且 $\forall x \geq 0, \phi_4(x) \geq 1$，同时 $\forall x < 0, \phi_4(x) > 0$，也满足所有条件。

因此，只有 $\phi_4(x)$ 满足题面中的假设条件。

使用 $\phi_4(x)$ 时，$D_t(i)$ 的表达式：

由于 $\phi_4(x) = \log_2(1 + e^x)$，其导数 $\phi_4'(x) = \frac{e^x}{(1+e^x)\ln(2)}$。因此，$D_t(i)$ 可以表示为：

$$D_t(i) = \frac{\phi_4'(-y_i f_t(x_i))}{Z_t} = \frac{e^{-y_i f_t(x_i)}}{(1 + e^{-y_i f_t(x_i)})\ln(2) Z_t}$$

3.2.2 确定最优步长 β

对于 $\phi_4(x)$，我们有：

$$\frac{dL(\alpha + \beta e_t)}{d\beta} = -\sum_{i=1}^{n} y_i h_t(x_i) \frac{e^{-y_i f_t(x_i)}}{(1 + e^{-y_i f_t(x_i)})\ln(2)}$$

令 $\frac{dL(\alpha+\beta e_t)}{d\beta} = 0$，我们得到：

$$\sum_{i=1}^{n} y_i h_t(x_i) \frac{e^{-y_i f_t(x_i)}}{(1 + e^{-y_i f_t(x_i)})\ln(2)} = 0$$

即

$$\sum_{i=1}^{n} y_i h_t(x_i) e^{-y_i f_t(x_i)} = 0$$

由此可以确定最优步长 β。

3.3 带未知标签的 Boosting 算法

3.3.1 问题 1: 用 ϵ_t 和 α_t 表示 Z_t

根据 Adaboost 的框架:

$$\begin{aligned}
Z_t &= \mathbb{E}_{i\sim D_t}[e^{-\alpha_t y_i h_t(x_i)}] \\
&= \sum_{i=1}^{n} e^{-\alpha_t y_i h_t(x_i)} \\
&= \sum_{i=1}^{n} e^{-\alpha_t}\mathbb{1}[h_t(x_i)y_i = 1] + \sum_{i=1}^{n} e^{\alpha_t}\mathbb{1}[h_t(x_i)y_i = -1] + \sum_{i=1}^{n} 1\mathbb{1}[h_t(x_i)y_i = 0] \\
&= \epsilon_t^{+} e^{-\alpha_t} + \epsilon_t^{-} e^{\alpha_t} + \epsilon_t^{0}.
\end{aligned}$$

展开得:

$$Z_t = \epsilon_t^{+}\exp(-\alpha_t) + \epsilon_t^{-}\exp(\alpha_t) + \epsilon_t^{0},$$

3.3.2 问题 2: 计算 $F(x_{t-1}, e_t)$ 并指出优化目标

损失函数 F 的定义:

$$\begin{aligned}
F'(\alpha_{t-1}, e_k) &= \lim_{\eta\to 0}\frac{F(\overline{\alpha}_{t-1} + \eta e_k) - F(\overline{\alpha}_{t-1})}{\eta} \\
&= -\frac{1}{n}\sum_{i=1}^{n} y_i h_k(x_i) e^{-y_i\sum_{j=1}^{N}\alpha_{t-1,j}h_j(x_i)} \\
&= -\frac{1}{n}\sum_{i=1}^{n} y_i h_k(x_i)\overline{D}_t(i)\overline{Z}_t \\
&= -\left[\sum_{i=1}^{n}\overline{D}_t(i)\mathbb{1}[y_i h_k(x_i) = +1] - \sum_{i=1}^{n}\overline{D}_t(i)\mathbb{1}[y_i h_k(x_i) = -1] + 0\mathbb{1}[y_i h_k(x_i) = 0]\right] \\
&= -\left[(\overline{\epsilon}_{t,k}^{+}) - \overline{\epsilon}_{t,k}^{-}\right]\frac{\overline{Z}_t}{n} = \left[(\overline{\epsilon}_{t,k}^{-}) - \overline{\epsilon}_{t,k}^{+}\right]\frac{\overline{Z}_t}{n}
\end{aligned}$$

优化目标: 在第 t 步中，我们选择一个合适的 k 以最小化

$$\left[(\overline{\epsilon}_{t,k}^{-}) - \overline{\epsilon}_{t,k}^{+}\right]\frac{\overline{Z}_t}{n}$$

同时还要寻找一个合适的步长，由下一题计算。

3.3.3 问题 3: 解 $\frac{\partial F}{\partial \alpha_t} = 0$，并给出 α_t 的更新公式

计算导数： 对 $F(x_{t-1}, e_t) = \epsilon_t e^{-\alpha_t} + (1-\epsilon_t)e^{\alpha_t}$ 求导：

$$\begin{aligned}
&\frac{\partial F(\bar{\alpha}_{t-1} + \eta e_k)}{\partial \eta} = 0 \\
&\qquad \Leftrightarrow -\frac{1}{n}\sum_{i=1}^{n} y_i h_k(x_i) e^{-y_i \sum_{j=1}^{N} \bar{\alpha}_{t-1,j} h_j(x_i) - \eta y_i h_k(x_i)} = 0 \\
&\qquad \Leftrightarrow \sum_{i=1}^{n} y_i h_k(x_i) \bar{D}_t(i) \bar{Z}_t e^{-\eta y_i h_k(x_i)} = 0 \\
&\qquad \Leftrightarrow \sum_{i=1}^{n} \bar{D}_t(i) y_i h_k(x_i) e^{-\eta y_i h_k(x_i)} = 0 \\
&\qquad \Leftrightarrow \sum_{i=1}^{n} \bar{D}_t(i) 1_{[y_i h_k(x_i) = +1]} e^{-\eta} - \sum_{i=1}^{n} \bar{D}_t(i) 1_{[y_i h_k(x_i) = -1]} e^{\eta} = 0 \\
&\qquad \Leftrightarrow [\bar{\epsilon}^{+}_{t,k} e^{-\eta} - \bar{\epsilon}^{-}_{t,k} e^{\eta}] = 0 \\
&\qquad \Leftrightarrow \eta = \frac{1}{2} \log \frac{\bar{\epsilon}^{+}_{t,k}}{\bar{\epsilon}^{-}_{t,k}}
\end{aligned}$$

因此 α_t 的更新公式为

$$\alpha_t = \frac{1}{2} \log \frac{\epsilon_t^{+}}{\epsilon_t^{-}}$$

3.3.4 问题 4: 证明训练误差 $\hat{\epsilon} = \frac{1}{n}\sum_{i=1}^{n} I_{y_i f(x_i) < 0} \leq \prod_{t=1}^{T} Z_t$

证明： 首先我们有

$$D_{t+1}(i) = \frac{1}{n} \frac{e^{-y_i f_t(x_i)}}{\prod_{s=1}^{t} Z_s}.$$

$$\begin{aligned}
\hat{\varepsilon}(h) &\leq \frac{1}{n}\sum_{i=1}^{n} 1[y_i f(x_i) \leq 0] \leq \frac{1}{n}\sum_{i=1}^{n} e^{-y_i f(x_i)} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left[n \prod_{t=1}^{T} Z_t \right] D_{T+1}(i) = \prod_{t=1}^{T} Z_t
\end{aligned}$$

Z_t 是一个归一化因子

$$\begin{aligned}
Z_t &= \sum_{i=1}^{n} D_t(i) e^{-\alpha_t y_i h_t(x_i)} \\
&= \sum_{i: y_i h_t(x_i) \geq 0} D_t(i) e^{-\alpha_t} + \sum_{i: y_i h_t(x_i) < 0} D_t(i) e^{\alpha_t} \\
&= \epsilon_t^{+} e^{-\alpha_t} + \epsilon_t^{-} e^{\alpha_t} + \epsilon_t^{0}
\end{aligned}$$

现在我们有 $\hat{\varepsilon}(h) \leq \prod_{t=1}^{T} Z_t$ 和 $Z_t = \epsilon_t^{+} e^{-\alpha_t} + \epsilon_t^{-} e^{\alpha_t} + \epsilon_t^{0} e^{\alpha_t}$.

对 Z_t 求导得到 α_t

$$\frac{\partial z_t}{\partial \alpha_t} = 0 \Rightarrow \alpha_t = \frac{1}{2}\log\frac{\epsilon_t^+}{\epsilon_t^-}.$$

计算最小值

$$Z_t = \epsilon_t^+ e^{-\alpha_t} + \epsilon_t^- e^{\alpha_t} + \epsilon_t^0 = 2\sqrt{\epsilon_t^+\epsilon_t^-} + \epsilon_t^0.$$

$\hat{\mathcal{E}}(h) \le \prod_{t=1}^T Z_t \overset{\min}{\Leftrightarrow} \prod_{t=1}^T \left[2\sqrt{\epsilon_t^+\epsilon_t^-} + \epsilon_t^0\right]$

$$= \prod_{t=1}^T \sqrt{1 - \left(\frac{\epsilon_t^+ - \epsilon_t^-}{\sqrt{1-\epsilon_t^0}}\right)^2}$$

有不等式 $-x \le e^{-x}$. 可以得出

$$\le \prod_{t=1}^T \exp\left[-\frac{1}{2}\left(\frac{\epsilon_t^+ - \epsilon_t^-}{\sqrt{1-\epsilon_t^0}}\right)^2\right] = \exp\left[-\frac{1}{2}\sum_{t=1}^T\left(\frac{\epsilon_t^+ - \epsilon_t^-}{\sqrt{1-\epsilon_t^0}}\right)^2\right]$$

因此，训练误差 $\hat{\epsilon}(h)$ 满足上届

$$\hat{\epsilon}(h) \le \exp\left[-\frac{1}{2}\sum_{t=1}^T\left(\frac{\epsilon_t^+ - \epsilon_t^-}{\sqrt{1-\epsilon_t^0}}\right)^2\right]$$

如果对于每个弱学习器的误差都满足

$$\frac{\epsilon_t^+ - \epsilon_t^-}{\sqrt{1-\epsilon_t^0}} \ge \gamma \ge 0$$

带入到上面训练误差的上届

$$\hat{\epsilon}(h) \le \exp\left(-\frac{\gamma^2 T}{2}\right).$$

因此

$$\frac{1}{n}\sum_{i=1}^n 1_{y_i f(x_i)<0} \le \exp\left(-\frac{\gamma^2 T}{2}\right).$$

3.4 Gradient Boosting Machines

3.4.1 算法流程

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

1. 令 $f_0(x) = 0$。
2. For $t = 1$ to T:

(a) 计算在各个数据点上的梯度 $g_t = \left(\frac{\partial}{\partial \widehat{y}_i} \ell(y_i, \widehat{y}_i) \Big|_{\widehat{y}_i = f_{t-1}(x_i)} \right)_{i=1}^{n}$。

(b) 根据 $-g_t$ 拟合一个回归模型, $h_t = \arg\min_{h \in \mathcal{F}} \sum_{i=1}^{n} (g_{t,i} + h(x_i))^2$。

(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, $f_t(x) = f_{t-1}(x) + \alpha_t h_t(x)$。

3.4.2 回归问题

考虑回归问题，假设损失函数 $\ell(y, \widehat{y}) = \frac{1}{2}(y - \widehat{y})^2$。第 t 轮迭代时的 g_t 以及 h_t 的表达式为：

- 梯度 g_t 的计算公式为 $g_{t,i} = \frac{\partial}{\partial \widehat{y}_i} \ell(y_i, \widehat{y}_i) \Big|_{\widehat{y}_i = f_{t-1}(x_i)} = f_{t-1}(x_i) - y_i$。
- 回归模型 h_t 可以通过拟合 g_t 来得到，即 $h_t(x) = \arg\min_{h \in \mathcal{F}} \sum_{i=1}^{n} (f_{t-1}(x_i) - y_i + h(x_i))^2$。

3.4.3 二分类问题

考虑二分类问题，假设损失函数 $\ell(y, \widehat{y}) = \ln(1 + e^{-y\widehat{y}})$。第 t 轮迭代时的 g_t 以及 h_t 的表达式为：

- 梯度 g_t 的计算公式为 $g_{t,i} = \frac{\partial}{\partial \widehat{y}_i} \ell(y_i, \widehat{y}_i) \Big|_{\widehat{y}_i = f_{t-1}(x_i)} = -\frac{y_i e^{-y_i f_{t-1}(x_i)}}{1 + e^{-y_i f_{t-1}(x_i)}}$。
- 回归模型 h_t 可以通过拟合 g_t 来得到，即 $h_t(x) = \arg\min_{h \in \mathcal{F}} \sum_{i=1}^{n} (-\frac{y_i e^{-y_i f_{t-1}(x_i)}}{1 + e^{-y_i f_{t-1}(x_i)}} + h(x_i))^2$。

3.4.4 fit 函数

见 boosting.py

3.4.5 predict 函数

见 boosting.py

3.4.6 gradient_logistic 函数

见 boosting.py

3.4.7 运行的结果

根据实验结果可知：

1. 迭代次数 T 的选择很重要。过高的迭代次数会导致过拟合，而过低则可能导致欠拟合。

2. 选择不同的损失函数对模型结果也有不同的影响。

   (a) 使用 L2 损失函数，梯度更新更直接，收敛更快，但对异常值比较敏感，因为平方项会放大大误差的影响，可能会导致模型在包含异常值的数据集上过拟合。

   (b) 使用 Logistic 损失函数，需要更多的迭代来达到相同的性能水平，因为其梯度更新在预测值接近 0 或 1 时会减慢。由于 Logistic 损失对异常值不太敏感，它在包含异常值的数据集上可能比 L2 损失更鲁棒。

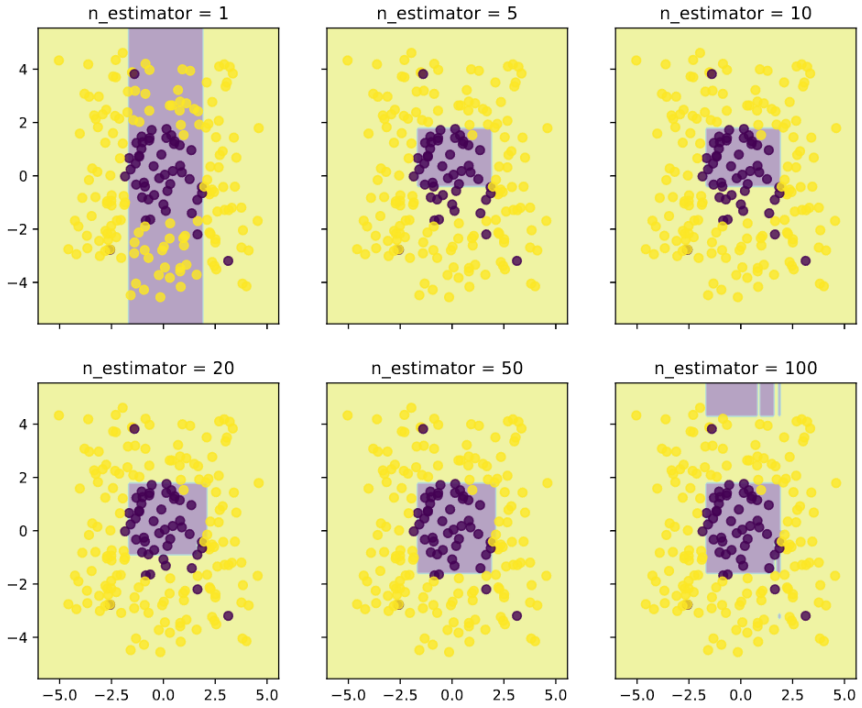

图 4: GBM_l2

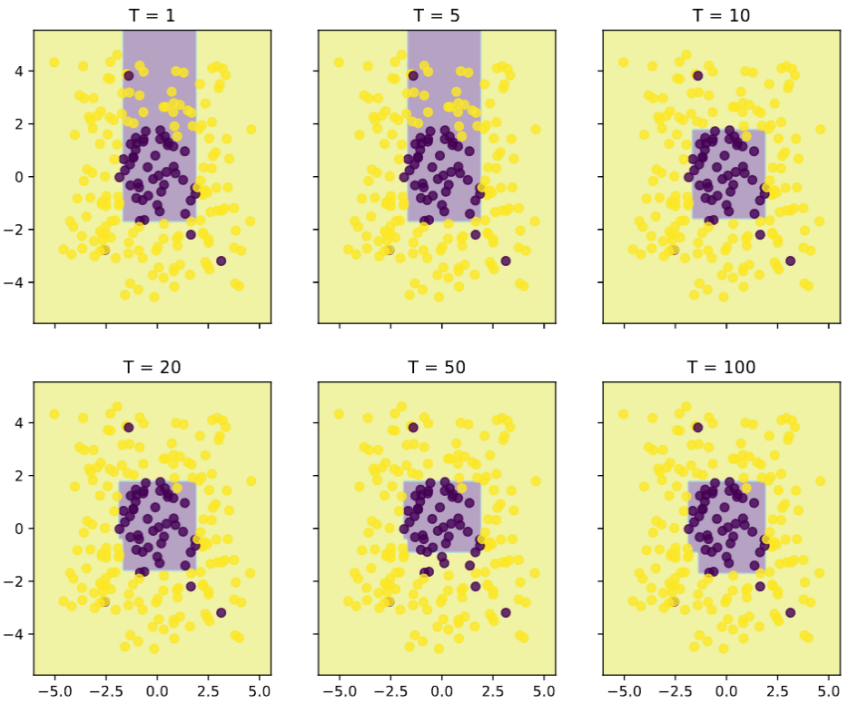

图 5: GBM_logistic

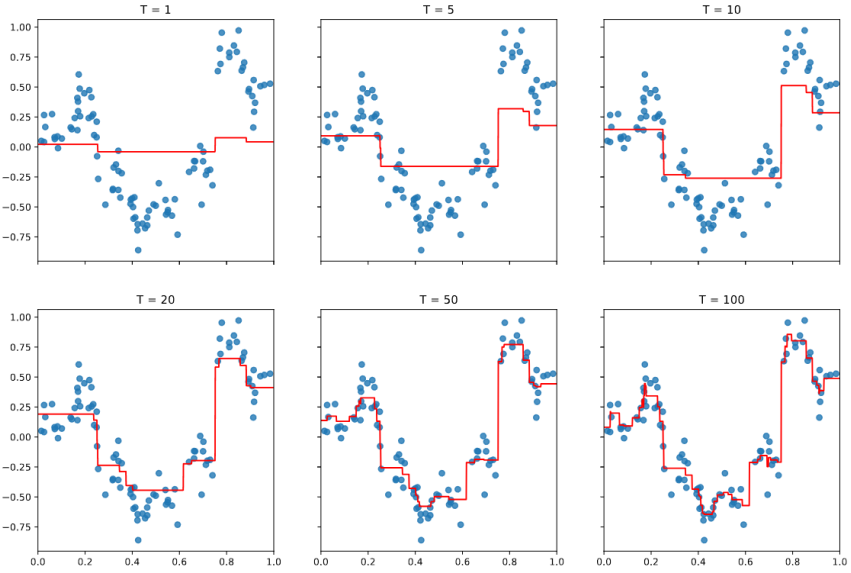

图 6: GBM_regression