机器学习第三次作业

汪隽立 计 14 2021012957

2023 年 12 月 18 日

解答 2.1.1.

在第一次选择划分特征时，因为 x_2，x_3 的信息增益都为 0，故选择 x_1。第一次划分以后，对于左节点，其经验熵为

$$H(D) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$$

对于左节点的第二步划分，选择 x_2，x_3 的信息增益都为

$$IG = H(D) - \left[\frac{2}{3}(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}) + \frac{1}{3}\times 0\right] = 0.252$$

也就是说，第二步无论选择 x_2 还是 x_3 作为划分特征，总会有一个叶节点包括两个 y 不同的样本，也即这里会有一个样本被分类错误，即 $\epsilon \geq \frac{1}{4}$。

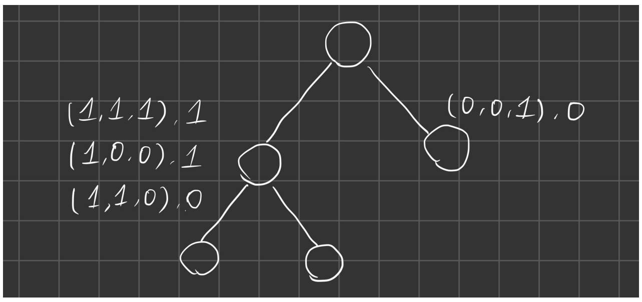

图 1: Sol 2.1.1

解答 2.1.2.

如图 2 所示。

解答 2.2.1.

t 个决策树，每个决策树都不选该特征的概率为 $(1-\frac{1}{d})^t$。

解答 2.2.2.

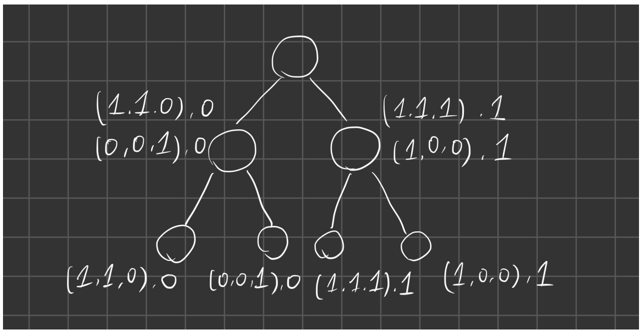

图 2: Sol 2.1.2

t 个决策树，每个决策树独立自助 m 个样本。某个样本从未被选中的概率为 $(1-\frac{1}{n})^{mt}$。

解答 2.3.5.

实验结果见图 3。

实验过程中，可调整的超参数有 min_sample 和 max_depth。在给出的结果中，可以看出当 max_depth 为 4 时，模型就已经较好地学习到了数据的分布。这时，随着 max_depth 的增加，模型的泛化能力不再增加，对于过拟合的现象不能很好地解决。

而适当增加 min_sample，可以增加模型的泛化能力。例如当 min_sample 变为 30 时，分类问题的结果变成了图 4:

解答 3.1.1.

首先证明 G 在 $x=0$ 处可导。$\lim_{x\to 0^+}\frac{G(x)-1}{x}=1$，$\lim_{x\to 0^-}\frac{G(x)-1}{x}=\lim_{x\to 0^-}\frac{e^x-1}{x}=1$。故 $G'(0)=1$。又 G 在除了 $x=0$ 处处可导，故 G 处处可导。

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

因为 G' 单调递增，故 G 是 $\mathbb{R}$ 上的凸函数。

解答 3.1.2.

定义 $D_t(i)Z_t=G(-y_i\sum_{j=1}^{N}\bar{\alpha_{t,j}}h_j(x_i))$，$Z_t$ 是归一化因子。那么:

$$\epsilon_t(h)=\sum_{i=1}^{m}D_t(i)\mathbb{1}_{[h(x_i)\neq y_i]}$$

。以下证明其合理性:

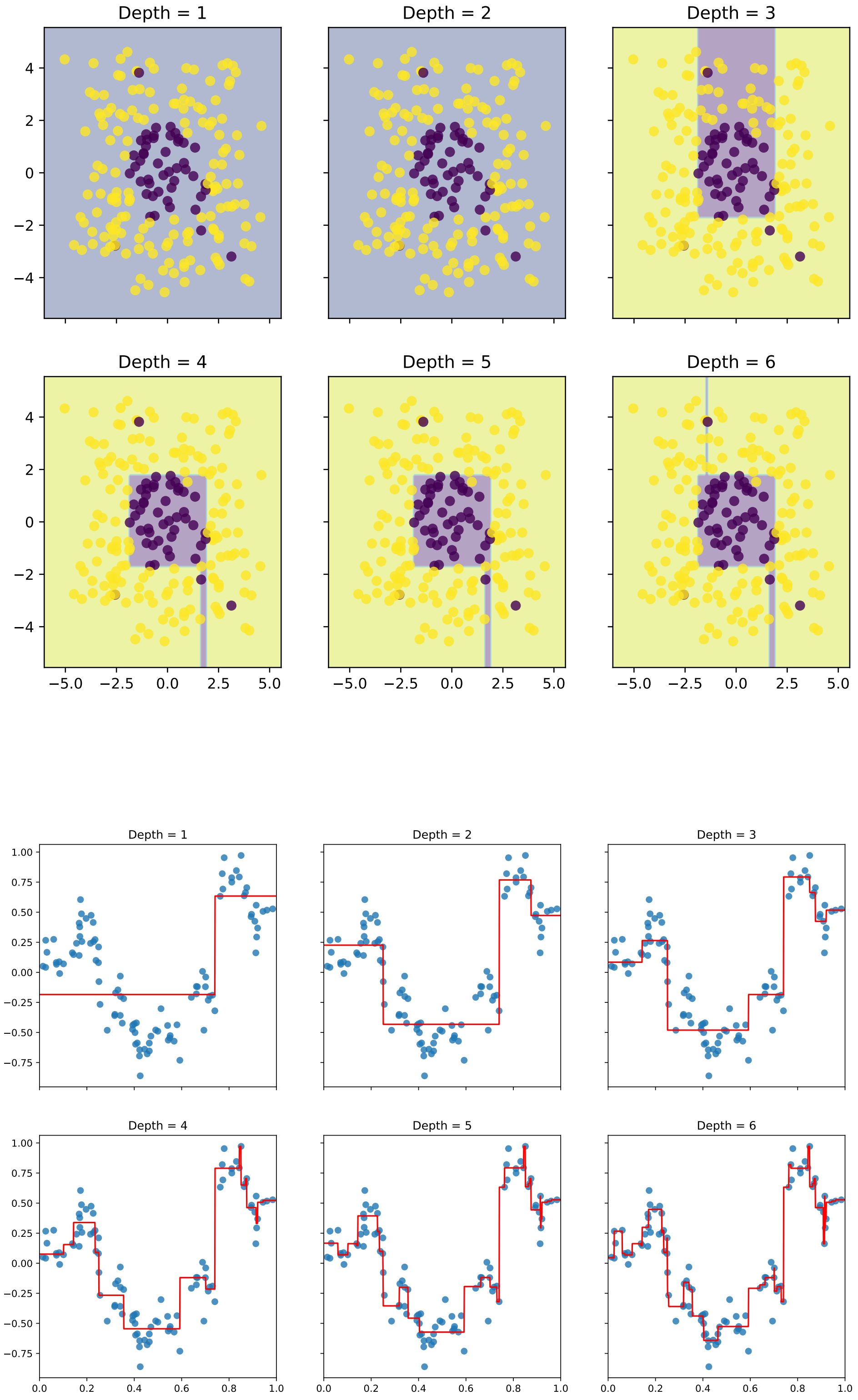
Depth = 1
Depth = 2
Depth = 3
Depth = 4
Depth = 5
Depth = 6
4
2
0
−2
−4
−5.0
−2.5
0.0
2.5
5.0
Depth = 1
Depth = 2
Depth = 3
Depth = 4
Depth = 5
Depth = 6
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0.0
0.2
0.4
0.6
0.8
1.0

图 3: Sol 2.3.5

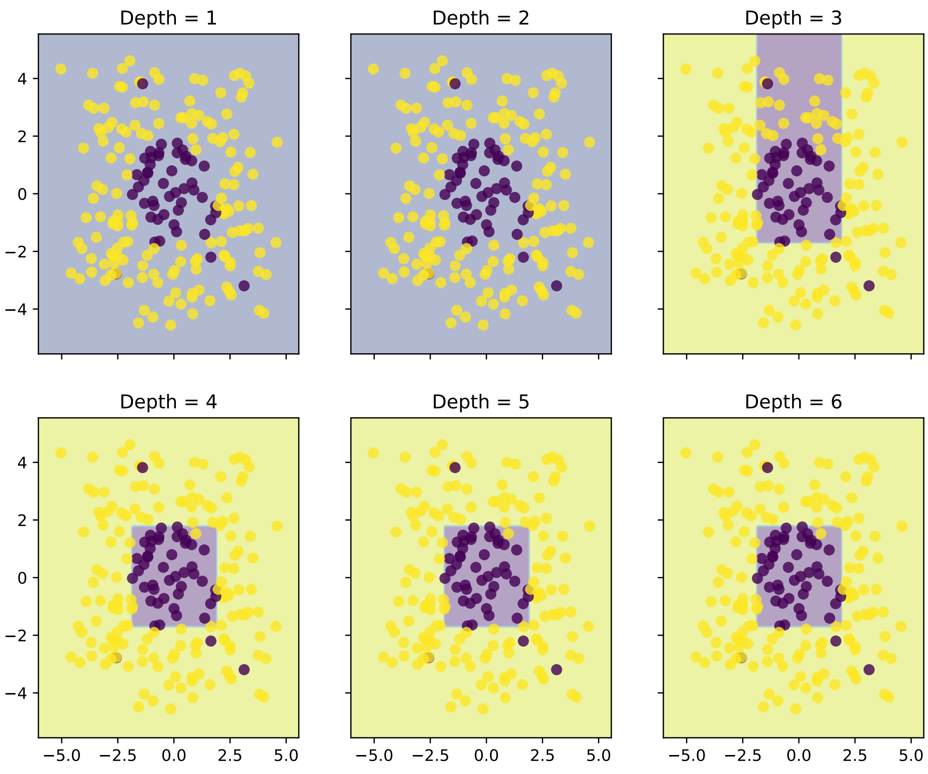

图 4: min_sample=30

利用坐标下降的方法，我们可以类似地得到：

$$
\begin{aligned}
F'(\bar{\alpha_t}, e_k) =& \lim_{\eta \to 0} \frac{F(\bar{\alpha_t} + \eta e_k) - F(\bar{\alpha_t})}{\eta} \\
=& \frac{1}{m} \sum_{i=1}^{m} y_i h_k(x_i) G(-y_i \sum_{j=1}^{N} \bar{\alpha_{t,j}} h_j(x_i)) \\
=& -\frac{Z_t}{m} \left[\sum_{i=1}^{m} D_t(i) \mathbb{1}_{[y_i h_k(x_i)=+1]} - \sum_{i=1}^{m} D_t(i) \mathbb{1}_{[y_i h_k(x_i)=-1]} \right] \\
=& (2\epsilon_{t,k} - 1) \frac{Z_t}{m}
\end{aligned}
$$

也就是说，在新的 Boosting 算法中，我们找的依然是使得坐标下降最快的那个 h_k。

解答 3.2.

$$
\begin{aligned}
\sum_i D_{t+1}(i) \mathbb{1}_{[y_i \neq h_t(x_i)]} =& \sum_i \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t} \mathbb{1}_{[y_i \neq h_t(x_i)]} \\
=& \frac{\exp(\alpha_t)}{Z_t} \epsilon_t \qquad (\text{在上一行中只取 } y_i \neq h_t(x_i) \text{ 的项}) \\
=& \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \frac{1}{2\sqrt{\epsilon_t(1-\epsilon_t)}} \epsilon_t \\
=& \frac{1}{2}
\end{aligned}
$$

因此，如果第 $t+1$ 步选取的弱分类器和第 t 步的相同，那么说明第 $t+1$ 步时，**$\mathcal{H}$ 中的所有分类器至少拥有 $\frac{1}{2}$ 的误差**，这与 Weak-Learnable 的假设矛盾。

解答 3.3.

根据 Adaboost 的 Empirical Bound:

$$
\hat{R}(h) \leq \exp(-2\gamma T^2)
$$

当 $T > \frac{\log m}{2\gamma^2}$ 时，有 $\hat{R}(h) < \frac{1}{m}$。又 $\hat{R}(h) = \{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$，故这时有 $\hat{R}(h) = 0$。

解答 3.4.

选择

$$
\alpha = \frac{1}{2} \log \frac{1+2\gamma}{1-2\gamma}
$$

根据 Adaboost 的 Empirical Bound，有:

$$
\begin{aligned}
Z_t \leq & (1-\epsilon_t)\exp(-\alpha_t)+\epsilon_t\exp(\alpha_t) \\
= & (1-\epsilon_t)\sqrt{\frac{1-2\gamma}{1+2\gamma}}+\epsilon_t\sqrt{\frac{1+2\gamma}{1-2\gamma}} \\
= & (\sqrt{\frac{1+2\gamma}{1-2\gamma}}-\sqrt{\frac{1-2\gamma}{1+2\gamma}})\epsilon_t+\sqrt{\frac{1-2\gamma}{1+2\gamma}} \\
\leq & (\sqrt{\frac{1+2\gamma}{1-2\gamma}}-\sqrt{\frac{1-2\gamma}{1+2\gamma}})(\frac{1}{2}-\gamma)+\sqrt{\frac{1-2\gamma}{1+2\gamma}} \\
= & \sqrt{1-4\gamma^2}
\end{aligned}
$$

再利用 $\hat{R}(h) \leq \prod_{t=1}^{T} Z_t$，即得 $\hat{R}(h) \leq (1-4\gamma^2)^{\frac{T}{2}}$。

解答 3.5.1.

$$h_t = \arg\min_{h\in\mathcal{F}} ||h_t+g_t||^2$$

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

解答 3.5.2.

$$g_t = f_{t-1}(x)-y$$

$$h_t = \arg\min_{h\in\mathcal{F}} ||h_t+f_{t-1}(x)-y||^2$$

解答 3.5.3.

$$g_t = \left(\frac{-y_i e^{-y_i f_{t-1}(x_i)}}{1+e^{-y_i f_{t-1}(x_i)}}\right)_{i=1}^{n} = \left(\frac{-y_i}{e^{y_i f_{t-1}(x_i)}+1}\right)_{i=1}^{n}$$

$$h_t = \arg\min_{h\in\mathcal{F}} ||h_t+g_t||^2$$

解答 3.5.7.

实验结果见图 5。可以看到，在二分类问题上，Logistic 回归比 GBM 更好地解决了过拟合的问题。而在 Regression 问题上，可以看出当迭代次数为 30 时，模型有比较好的泛化能力，而当迭代次数再增大时，模型出现了过拟合的现象。

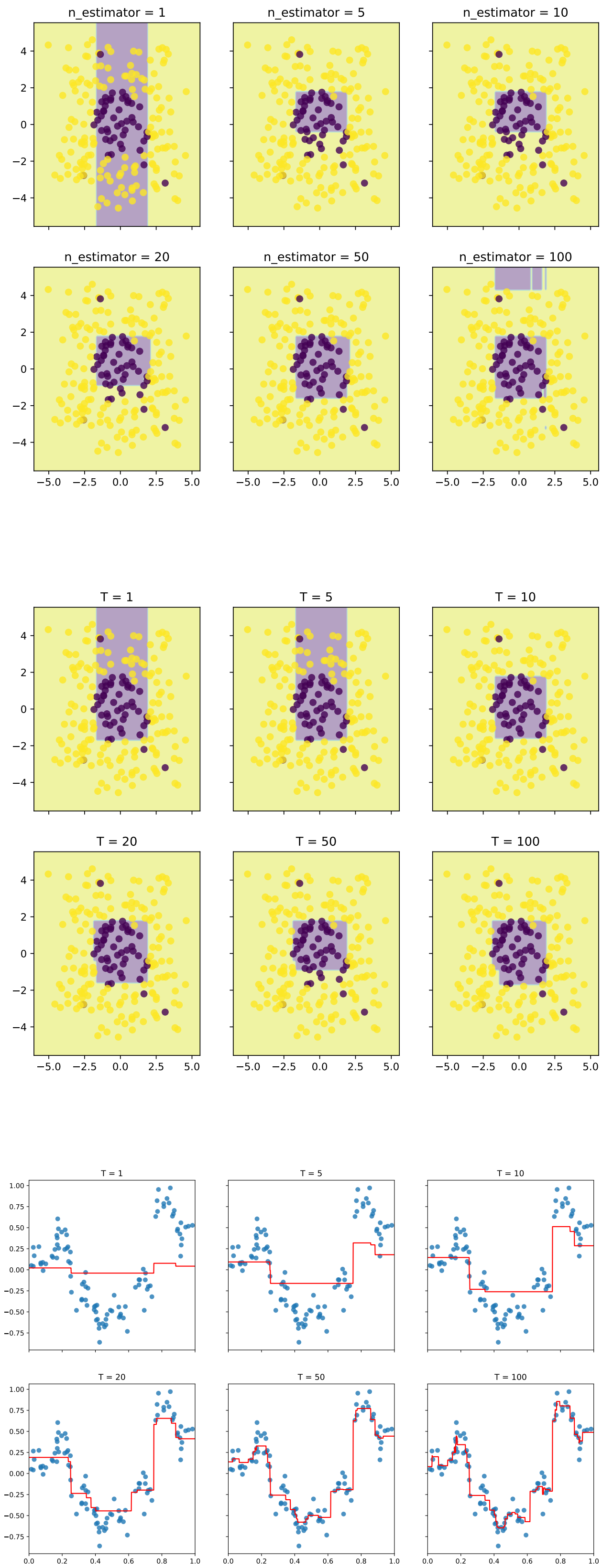
n_estimator = 1
n_estimator = 5
n_estimator = 10
n_estimator = 20
n_estimator = 50
n_estimator = 100
T = 1
T = 5
T = 10
T = 20
T = 50
T = 100
T = 1
T = 5
T = 10
T = 20
T = 50
T = 100

图 5: Sol 3.5.7