机器学习 作业3

1 介绍

2 决策树与随机森林

2.1 ID3 算法无法找到最优解的一个情形

2.1.1 ID3训练误差

以下为各节点的ID3算法运行情况：

- 根节点

首先计算Parent Entropy:

$$H_{parent} = -\frac{1}{2}log\frac{1}{2} - \frac{1}{2}log\frac{1}{2} = \log 2 = 1$$

分别采用三个维度特征分类：

$$H_{x_1} = \frac{3}{4}(-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}) + \frac{1}{4}(-1\log 1) = 0.689$$
$$H_{x_2} = \frac{1}{2}(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}) + \frac{1}{2}(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}) = 1$$
$$H_{x_3} = \frac{1}{2}(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}) + \frac{1}{2}(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}) = 1$$

故以第一个维度特征进行分类，$\{((1,1,1),1),((1,0,0),1),((1,1,0),0)\}$进入左孩子，$\{((0,0,1),0)\}$进入右孩子

- 根节点左孩子

首先计算Parent Entropy:

$$H_{parent} = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$$

分别采用第二、三个维度特征分类：

$$H_{x_2} = \frac{2}{3}(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}) + \frac{1}{3}(-1\log 1) = 0.667$$
$$H_{x_3} = \frac{2}{3}(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}) + \frac{1}{3}(-1\log 1) = 0.667$$

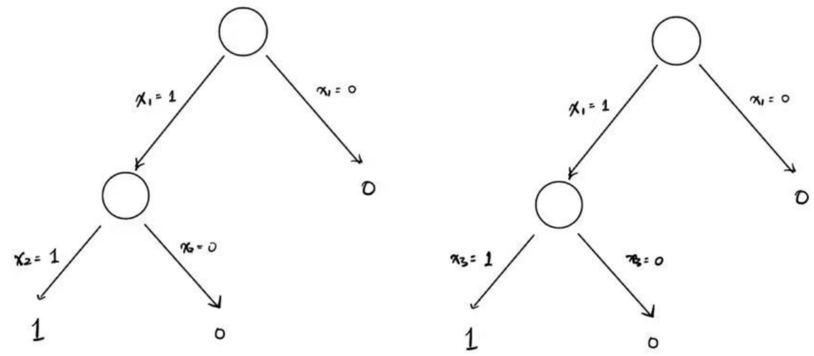
故随机采用x_2或x_3进行分类

- 根节点右孩子

仅有一个样本，不可再分

- 最终决策树

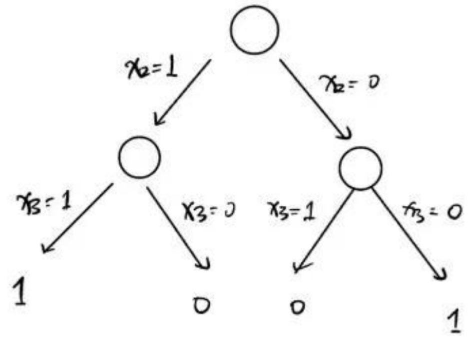
如下图所示：



故4个样本中总有一个会被分错，至少有1/4的训练误差

2.1.2 训练为0的决策树

如下图所示：



2.2 随机森林

2.2.1 任意一个特定的特征从未被选中分割的概率

在随机森林的构建当中，总共要进行$t \cdot h$次分割，每次抽取1个特征，故每一次分割未被选中的概率为

$$\frac{d-1}{d}$$

而每次分割选取特征是独立的，故总概率为

$$(\frac{d-1}{d})^{th}$$

2.2.2 任意一个特定的样本从未在任何一棵树中被考虑的概率

一共要构建t个二叉树，而每棵二叉树要抽取m个自助采样的样本，每次采样未被选中的概率为

$$(\frac{n-1}{n})^{m}$$

而每棵树的构建是独立的，故总概率为

$$(\frac{n-1}{n})^{tm}$$

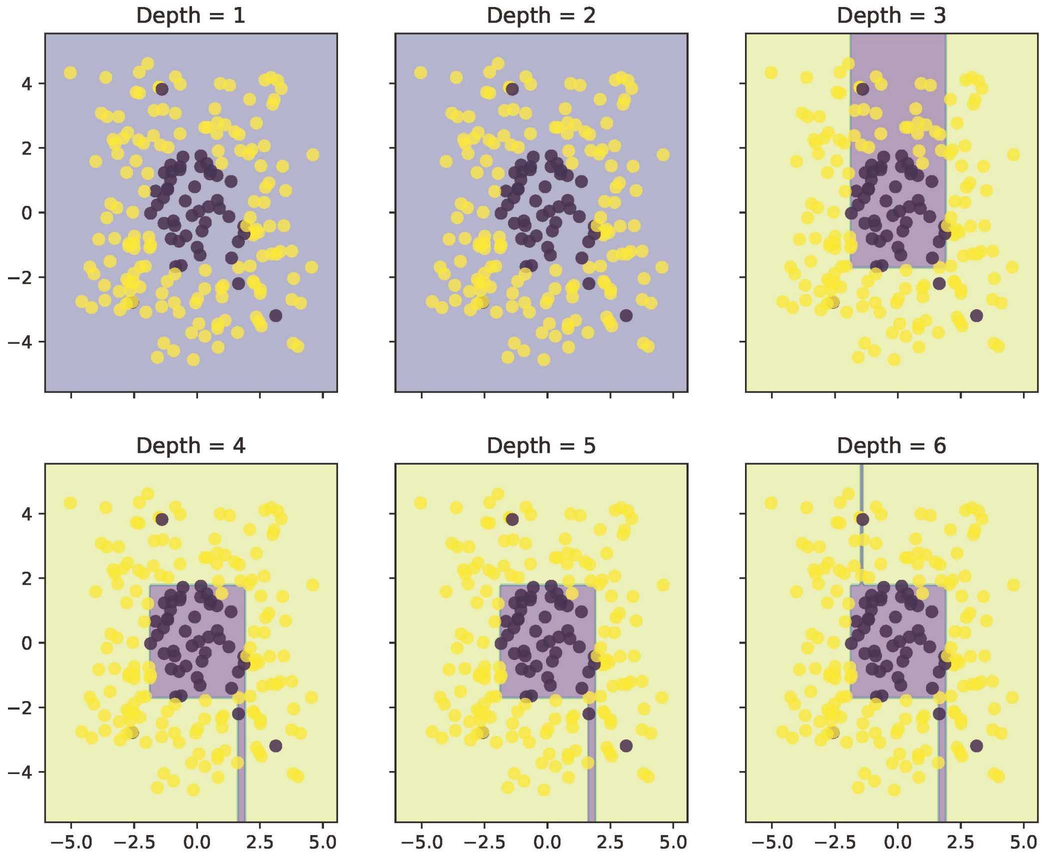
2.3 代码实验

2.3.1 1-4

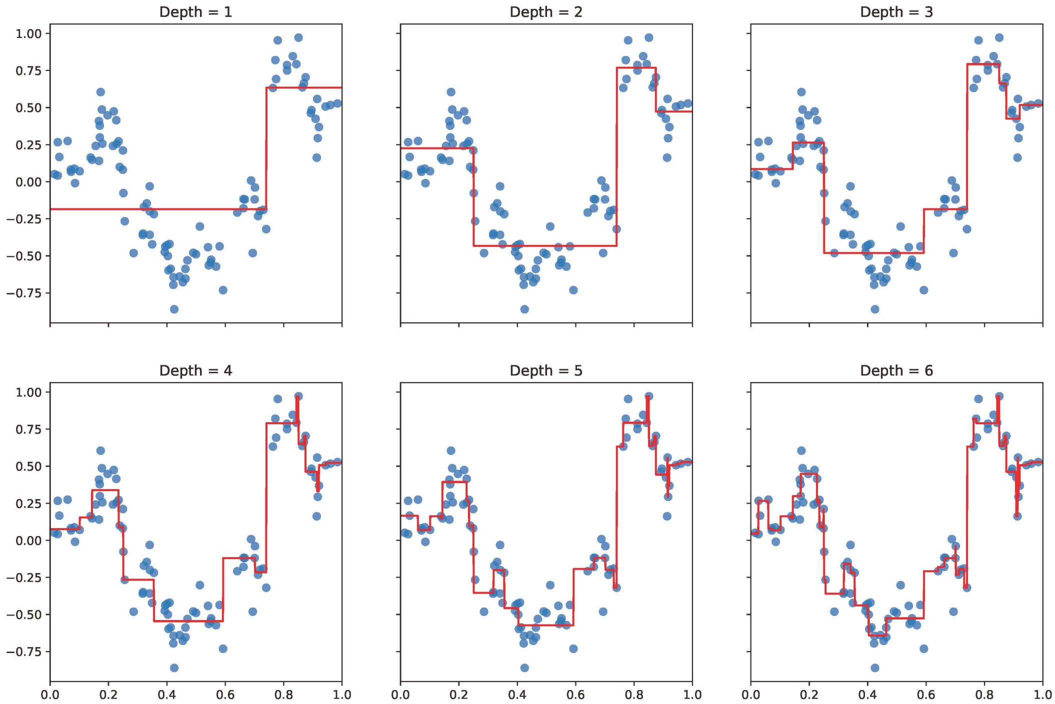
- 见 `tree.py`

2.3.2 实验结果

- 二分类问题



- 回归问题



- 实验现象
  - 可以看到，在以上二分类问题和回归问题中，决策树都表现出了较为显著的的拟合能力
  - 结果显示，随着超参数Depth增大，决策树的拟合能力不断增强，训练误差会下降；同时当Depth过大时也会出现过拟合现象，因此需要合理选取Depth的值

3 提升算法

3.1 噪声不敏感的AdaBoost 算法

3.1.1 函数G的凸性与可导性

- 凸性

即证明对于任意$x_1 < x_2 \in R$，均有

$$\forall\lambda \in [0,1], f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

（1）当$x_1 < x_2 \leq 0$或者$0 < x_1 < x_2$时，由e^x与$x+1$的凸性可知上式成立

（2）当$x_1 < 0 < x_2$时，

若$\lambda x_1 + (1-\lambda)x_2 \leq 0$，由$e^x$凸性：

$$f(\lambda x_1 + (1-\lambda)x_2) = f(\lambda' x_1 + (1-\lambda')\cdot 0) \leq \lambda' f(x_1) + (1-\lambda')f(0)$$

其中$\lambda' = \lambda + \frac{(1-\lambda)x_2}{x_1}$，于是

$$\begin{aligned}\lambda' f(x_1) + (1-\lambda')f(0) &= \lambda f(x_1) + \frac{(1-\lambda)x_2}{x_1} f(x_1) + (1-\lambda)(1-\frac{x_2}{x_1})f(x_2) \\ &\leq \lambda f(x_1) + \frac{(1-\lambda)x_2}{x_1} f(x_2) + (1-\lambda)(1-\frac{x_2}{x_1})f(x_2) \\ &= \lambda f(x_1) + +(1-\lambda)f(x_2)\end{aligned}$$

若$\lambda x_1+(1-\lambda)x_2>0$,则有：

$$
\begin{aligned}
f(x_1)=e^{x_1}\geq x_1+1 \\
\Rightarrow \lambda f(x_1)+(1-\lambda)f(x_2)\geq \lambda(x_1+1)+(1-\lambda)(x_2+1)=\lambda x_1+(1-\lambda)x_2+1 \\
=f(\lambda x_1+(1-\lambda)x_2)
\end{aligned}
$$

故G的凸性得证。

- 可导性

只需验证G在$x=0$处可导，而

$$
\lim_{\Delta x\to 0-}\frac{G(x+\Delta x)-G(x)}{\Delta x}=(e^x)'|_{x=0}=1
$$
$$
\lim_{\Delta x\to 0+}\frac{G(x+\Delta x)-G(x)}{\Delta x}=(x+1)'|_{x=0}=1
$$

故G在$x=0$处可导，且

$$
G'(0)=1
$$

故G处处可导。

3.1.2 损失函数表达式

类比一般的Adaboost，有

$$
\epsilon_t=P_{i\sim D_t}[h_t(x_i)\neq y_i]
$$

其中

$$
D_1(i)=\frac{1}{m}
$$

递推关系为

$$
\overline{D}_{t+1}(i)=\frac{G'(-y_i\sum_{j=1}^{N}\overline{\alpha}_{t,j}h_j(x_i))}{\sum_{i=1}^{N}G'(-y_i\sum_{j=1}^{N}\overline{\alpha}_{t,j}h_j(x_i))}
$$
$$
\overline{\alpha}_t=\overline{\alpha}_{t-1}+\eta e_k
$$

而e_k为具有最小ϵ_t的基分类器h_t所对应维度

其合理性由下式给出：

$$
\frac{\partial F(\overline{\alpha}_{t-1}+\eta e_k)}{\partial\eta}|_{\eta=0}=-\frac{1}{m}\sum_{i=1}^{m}y_ih_k(x_i)G'(\overline{\alpha}_{t-1})
$$
$$
\propto\sum_{i=1}^{m}(-y_ih_k(x_i))\overline{D}_t(i)=(2\epsilon_t-1)
$$

即选取具有最小ϵ_t的基分类器；

同时令

$$
\frac{\partial F(\overline{\alpha}_{t-1}+\eta e_k)}{\partial\eta}=0
$$

可以得到

$$
\frac{1}{m}\sum_{i=1}^{m}y_ih_k(x_i)G'(-y_i\sum_{j=1}^{N}\overline{\alpha}_{t-1,j}h_j(x_i)-\eta y_ih_k(x_i))=0
$$

由此可解出η的值

3.2 探究 AdaBoost 算法是否能使用完全相同的弱分类器

$$\sum_i D_{t+1}(i) \cdot 1_{[y_i \neq h_t(x_i)]} = \sum_i \frac{D_t(i)exp(-\alpha_t y_i h_t(x_i))}{\sum_i D_t(i)exp(-\alpha_t y_i h_t(x_i))} \cdot 1_{[y_i \neq h_t(x_i)]}$$

代入

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

有

$$\begin{aligned}\sum_i D_{t+1}(i) \cdot 1_{[y_i \neq h_t(x_i)]} &= \frac{\sum_i (\frac{1-\epsilon_t}{\epsilon_t})^{-y_i h_t(x_i)/2} D_t(i) \cdot 1_{[y_i \neq h_t(x_i)]}}{\sum_i (\frac{1-\epsilon_t}{\epsilon_t})^{-y_i h_t(x_i)/2}} \\ &= \frac{\epsilon_t (\frac{1-\epsilon_t}{\epsilon_t})^{1/2}}{\epsilon_t (\frac{1-\epsilon_t}{\epsilon_t})^{1/2} + (1-\epsilon_t)(\frac{1-\epsilon_t}{\epsilon_t})^{-1/2}} \\ &= \frac{1-\epsilon_t}{2(1-\epsilon_t)} = \frac{1}{2}\end{aligned}$$

而在t+1步时，选取的h_{t+1}必然使得ϵ_{t+1}达到最小；而对于二分类问题，有

$$\min \epsilon_{t+1} = \min \sum_i D_{t+1}(i) \cdot 1_{[y_i \neq h_{t+1}(x_i)]} < 1/2$$

故h_{t+1}与h_t不会相同

3.3 AdaBoost 的训练误差

当$T > \frac{\log m}{2\gamma^2}$时，

由课件上给出的经验误差界，

$$\hat{\epsilon}(h) \le exp[-2\sum_{t=1}^{T}(\frac{1}{2} - \epsilon_t)^2] < exp[-2 \cdot \frac{\log m}{2\gamma^2} \cdot \gamma^2] = \frac{1}{m}$$

而在m条数据集上，经验误差只可能为$\frac{k}{m}$，$k = 0, 1 \ldots m$

故得到$k = 0$，即$\hat{\epsilon}(h) = 0$，训练误差达到0

3.4 简化版本的 AdaBoost

首先有

$$\hat{\epsilon}(h) \le \prod_{t=1}^{T} Z_t$$

而

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

回代得到

$$\hat{\epsilon}(h) \le \prod_{t=1}^{T}(e^{-\alpha} + \epsilon_t(e^{\alpha} - e^{-\alpha}))$$

$$\le \prod_{t=1}^{T}(e^{-\alpha} + (1/2-\gamma)(e^{\alpha} - e^{-\alpha})) = (e^{-\alpha} + (1/2-\gamma)(e^{\alpha} - e^{-\alpha}))^T$$

$$= ((1/2-\gamma)e^{\alpha} + (1/2+\gamma)e^{-\alpha})^T$$

取

$$e^{\alpha} = \sqrt{\frac{1/2+\gamma}{1/2-\gamma}}$$
$$\Rightarrow \alpha = 1/2 \log \frac{1/2+\gamma}{1/2-\gamma}$$

则有

$$\hat{\epsilon}(h) \leq (2\sqrt{(1/2+\gamma)(1/2-\gamma)})^{T} = (1-4\gamma^2)^{T/2}$$

3.5 Gradient Boosting Machines

3.5.1 总结算法流程

(b)

$$h_t = arg\min_{h \in F} \ell(h, -g_t)$$

(c)

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

3.5.2 回归问题

$$g_t = (f_{t-1}(x_i) - y_i)_{i=1}^n$$
$$h_t = arg\min_{h \in F} \frac{1}{2}(h(x) + (f_{t-1}(x_i) - y_i)_{i=1}^n)^2 = arg\min_{h \in F} \sum_{i=1}^{n}(h(x_i) + f_{t-1}(x_i) - y_i)^2$$

3.5.3 二分类问题

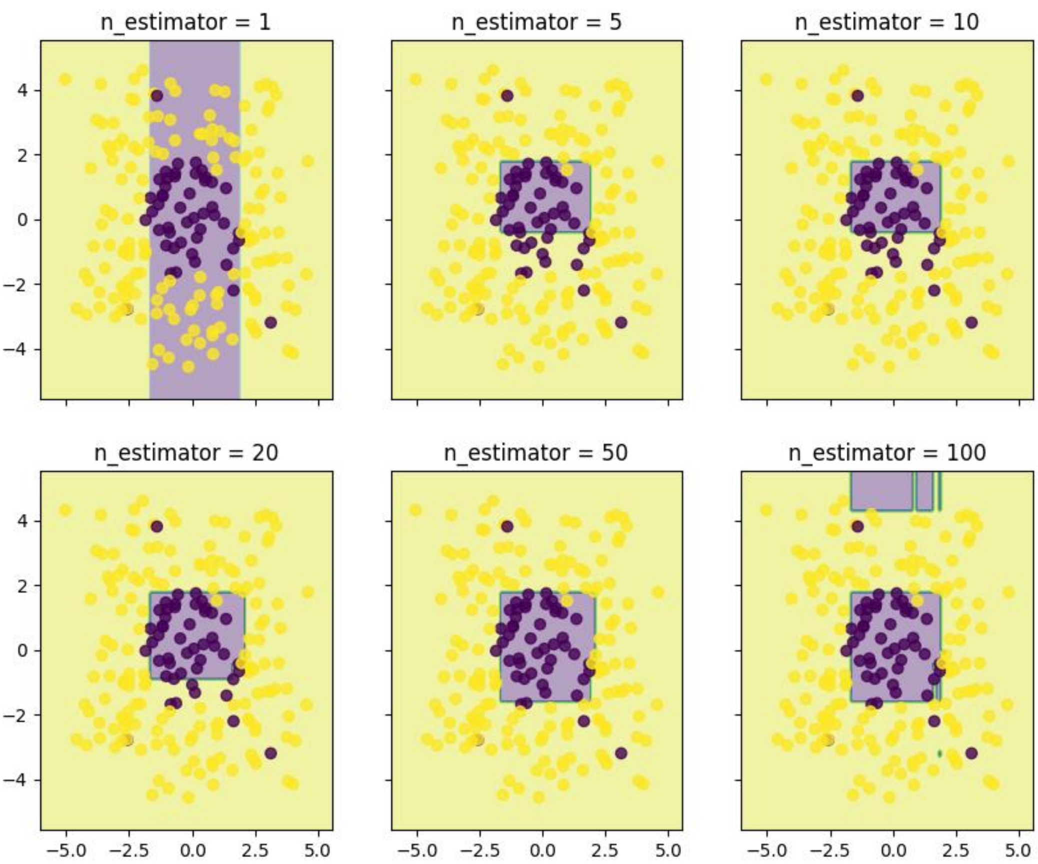
$$g_t = -(\frac{y_i e^{-y_i f_{t-1}(x_i)}}{1+e^{-y_i f_{t-1}(x_i)}})_{i=1}^n = (-\frac{y_i}{1+e^{y_i f_{t-1}(x_i)}})_{i=1}^n$$
$$h_t = arg\min_{h \in F} \sum_{i=1}^{n}(h(x_i) - \frac{y_i}{1+e^{y_i f_{t-1}(x_i)}})^2$$

3.5.4 代码补全

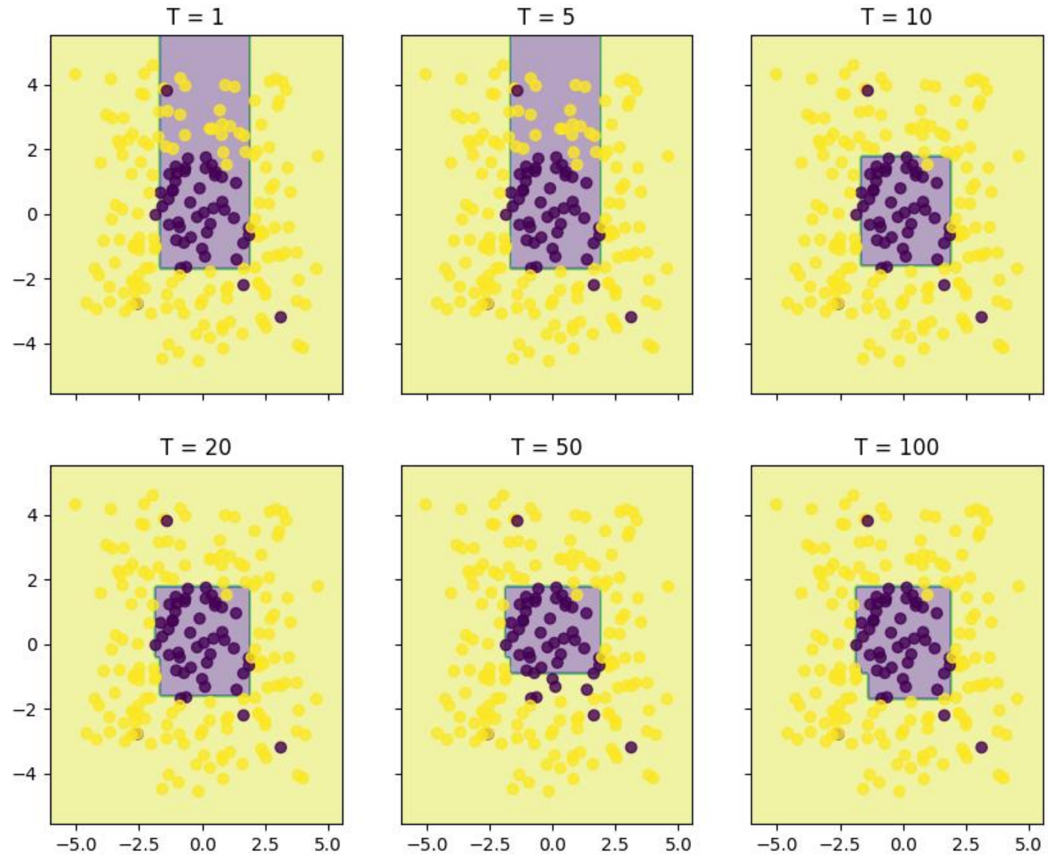
- 见 `boosting.py`

3.5.5 实验结果

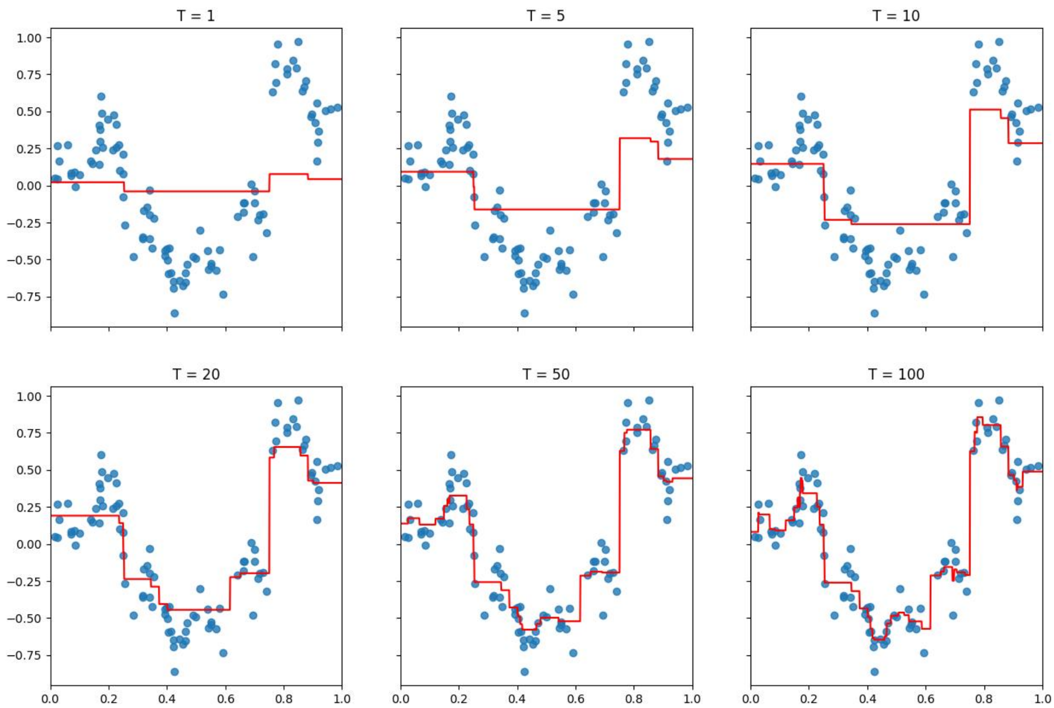
- L2 Loss 在二分类问题上的结果



- logistic loss 在二分类问题上的结果



- L2 loss 在回归问题上的结果



3.5.6 实验现象

- 可以看到，超参数T决定了GBM的拟合能力，T越大时GBM拟合能力越强，训练误差越低；同时也更容易过拟合
- 同时注意到，在GBM的二分类问题上，使用Logistic Loss相比较于L2 Loss能够更好地避免过拟合；这是因为Logistic Loss对噪音/异常数据更加鲁棒