Bagging : weak learners (overfitting)

Boosting : weak learners (underfitting)

提升树 — 基于残差的训练

优点：可并行计算，训练效率高

实际效果好

可控参数多

① 如何构造目标函数 → not continous (离散)

↓

② 目标函数如何优化 (如何近似) → Taylor Expansion

↓

③ 如何把树结构引入目标函数

↓

④ 仍然难优化

要不要用贪心算法 → NP-hard problem

优化

① 使用多棵树预测

训练了K棵树，对于第i个样本，最终预测值：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad f_k \in \mathcal{F}$$

目标函数 $Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$

Loss function

penalty — 控制复杂度.

(第K棵树复杂度)

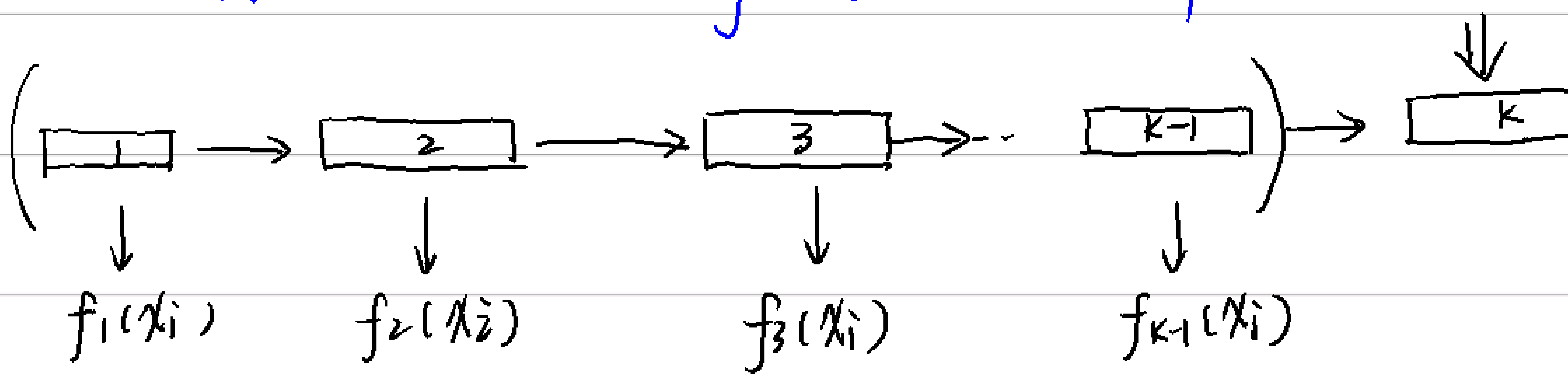
树的复杂度：叶节点

树的深度

叶节点的值

⋮

Additive Training（叠加式的训练）



给定：x_i

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)$$

$$\vdots$$

$$\hat{y}_i^{(K)} = f_1(x_i) + f_2(x_i) + \dots + f_K(x_i) = \underbrace{\hat{y}_i^{(K-1)} + f_K(x_i)}_{\hat{y}_i}$$

minimize.

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(K)}) + \sum_{k=1}^{K} \Omega(f_k)$$

$$= \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(K-1)} + f_K(x_i)) + \underbrace{\sum_{j=1}^{K-1} \Omega(f_j)}_{\text{常数项}} + \Omega(f_K)$$

⇒ 当训练第K棵树：

第K棵树预测值

$$\text{minimize} \quad \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(K-1)} + f_K(x_i)) + \Omega(f_K)$$

y_i：真实值

$\hat{y}_i^{(K-1)}$：前K-1棵预测（Constant）

$\Omega(f_K)$ ↳ 第K棵树复杂度

②

$$\text{minimize} \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(k-1)} + f_k(x_i)) + \Omega(f_k)$$

Talyor Expansion 近似目标函数

$$f(x+\Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2$$

$$f(x) = L(y_i, \hat{y}_i^{(k-1)}) \qquad f(x+\Delta x) = L(y_i, \underbrace{\hat{y}_i^{(k-1)}}_{x} + \underbrace{f_k(x_i)}_{\Delta x})$$

$$obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(k-1)} + f_k(x_i)) + \Omega(f_k)$$

$$= \sum_{i=1}^{n} \left[\boxed{L(y_i, \hat{y}_i^{(k-1)})} + \underbrace{\boxed{\frac{\partial L(y_i, \hat{y}_i^{(k-1)})}{\partial \hat{y}_i^{(k-1)}}}}_{g_i} \cdot f_k(x_i) + \underbrace{\boxed{\frac{1}{2} \frac{\partial^2 L(y_i, \hat{y}_i^{(k-1)})}{\partial^2 \hat{y}_i^{(k-1)}}}}_{h_i} \cdot f_k^2(x_i) \right] + \Omega(f_k)$$

带 k-1 项均已知

$$= \sum_{i=1}^{n} [\underbrace{\cancel{L(y_i, \hat{y}_i^{(k-1)})}}_{\text{constant}} + g_i f_k(x_i) + \frac{1}{2} h_i f_k^2(x_i)] + \Omega(f_k)$$

$$\Rightarrow \text{minimize} \sum_{i=1}^{n} [g_i f_k(x_i) + \frac{1}{2} h_i f_k^2(x_i)] + \Omega(f_k).$$

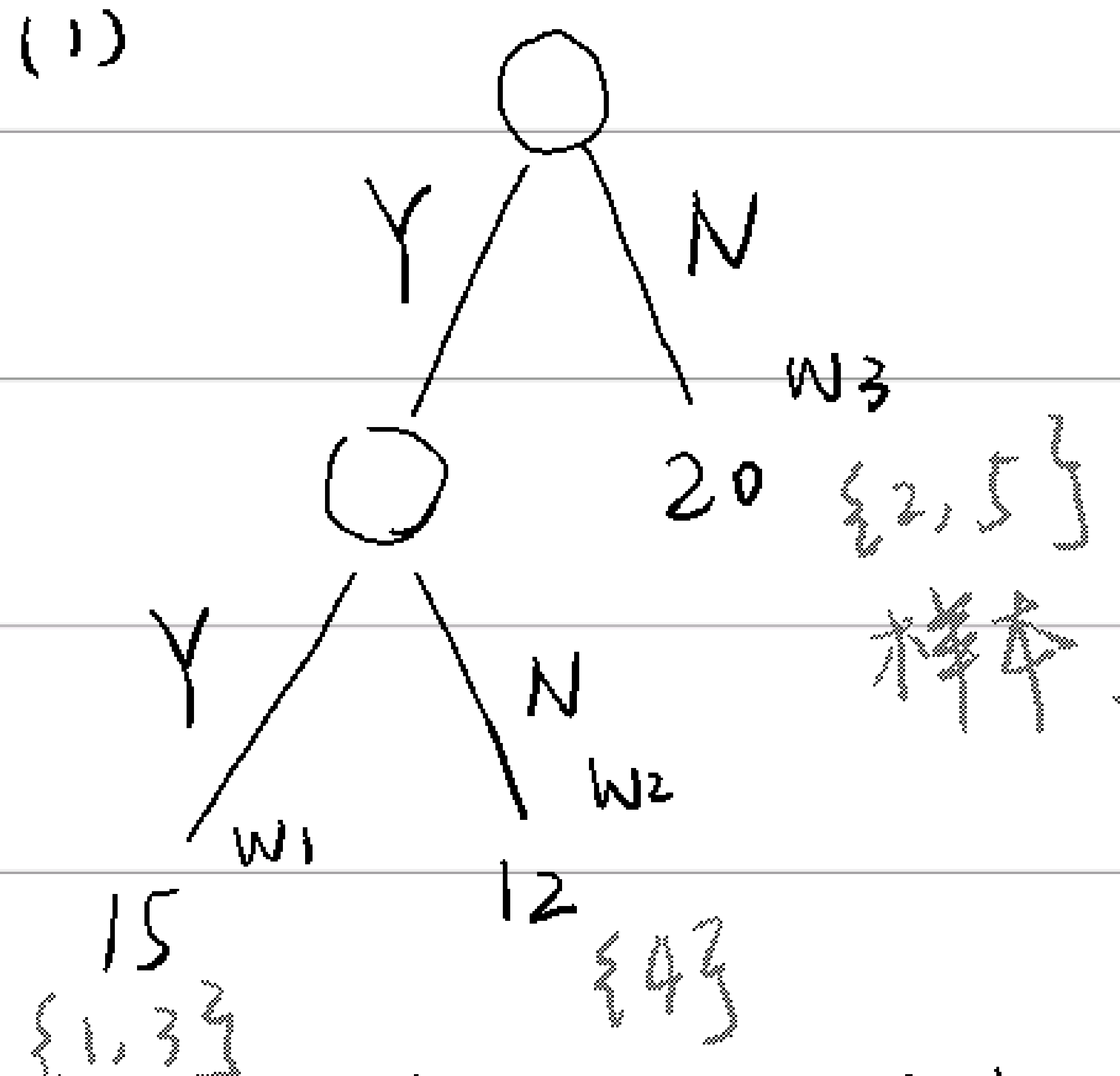
训练第 k 棵时 $\{h_i, g_i\}$ 是已知的.

(残差表现在 h_i, g_i 中)

③ minimize $\sum_{i=1}^{n}[g_i f_K(x_i) + \frac{1}{2} h_i f_K^2(x_i)] + \Omega(f_K)$.

(1) 参数化 (2) 参数化

(1)



$\Rightarrow$

$q(x)$：样本 x 的位置

$q(x_1)=1$

$q(x_2)=3$

$q(x_3)=1$

$q(x_4)=2$

$q(x_5)=3$.

$f_K(x_i) = W_{q(x_i)}$

参数化

$W = (w_1, w_2, w_3) = (15, 12, 20)$

$I_j = \{i \mid q(x_i) = j\} \Rightarrow$ $I_1 = \{1,3\}$, $I_2 = \{4\}$, $I_3 = \{2,5\}$

有哪些样本 x_i 在第 j 个位置上

(2) 树复杂度：叶节点个数 leaf value.

$\Omega(f_K) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$

叶节点个数

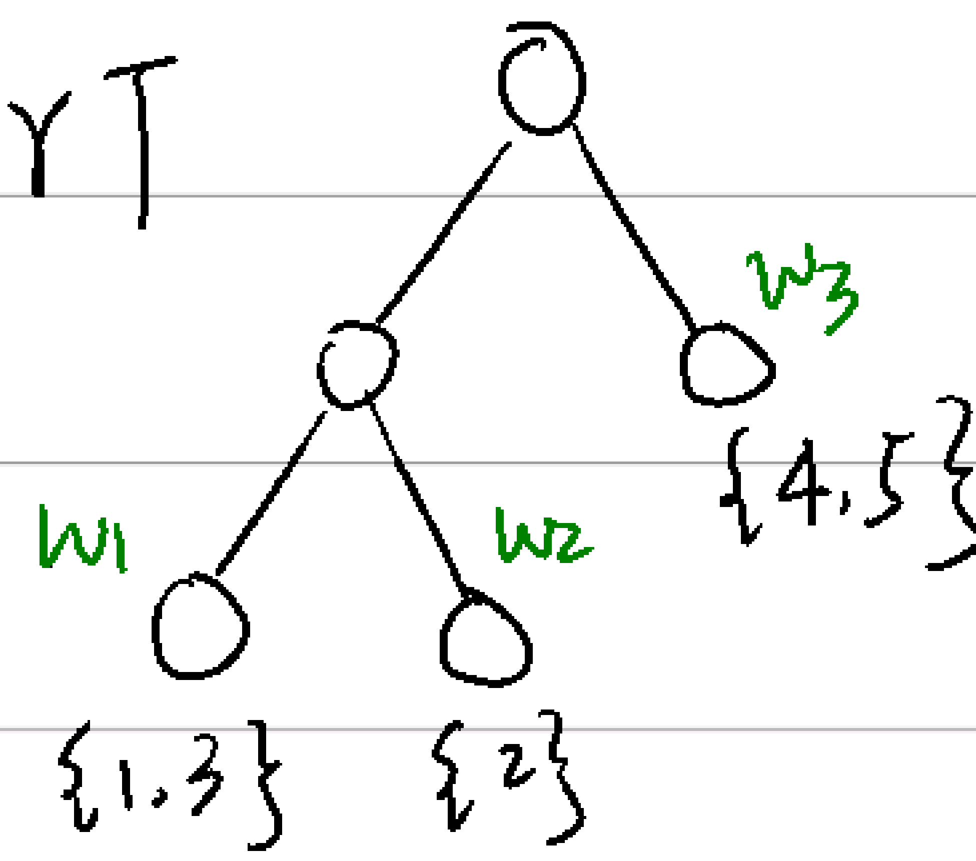
$$\text{minimize} \sum_{i=1}^{n} [g_i f_K(x_i) + \frac{1}{2} h_i f_K^2(x_i)] + \Omega(f_K).$$

$$= \sum_{i=1}^{n} [g_i \cdot w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

$a w_j + b w_j^2 \rightarrow$ 有最优解.

$$\Rightarrow \sum_{j=1}^{T} [(\underbrace{\sum_{i \in I_j} g_i}_{\text{constant}}) w_j + \frac{1}{2}(\underbrace{\sum_{i \in I_j} h_i + \lambda}_{\text{constant}}) w_j^2] + \gamma T$$

G_j $\quad H_j$



$$w_j^* = -\frac{b}{2a}$$

$$= -\frac{G_j}{H_j + \lambda}$$

$$Obj^* = -\frac{1}{2} \sum_{j=1}^{T} (\frac{G_j}{H_j + \lambda})^2 + \gamma T$$

样本组级:

$g_1 w_{q(x_1)} + g_2 w_{q(x_2)} + g_3 w_{q(x_3)}$

$+ g_4 w_{q(x_4)} + g_5 w_{q(x_5)}$

$= g_1 \cdot w_1 + g_3 w_1$ (按叶节点遍历)

$+ g_2 w_2$

$+ \underbrace{g_4 w_3 + g_5 w_3}$

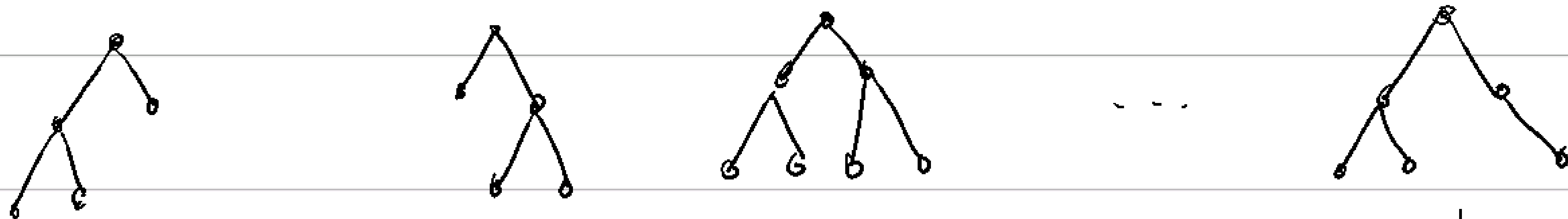
$$\sum_{i \in I_3} g_i w_3$$

☆: 树结构已知的情况下

(形状)

有以上结论

K棵树最小的目标函数 Obj_K^*



$Obj_K^{*(1)}$ $Obj_K^{*(2)}$ $Obj_K^{*(3)}$ . . . $Obj_K^{*(10000)}$

所有可能情况中寻找最小的那棵

$\min\{Obj_K^{*(i)}\}_{i=1}^{10000}$

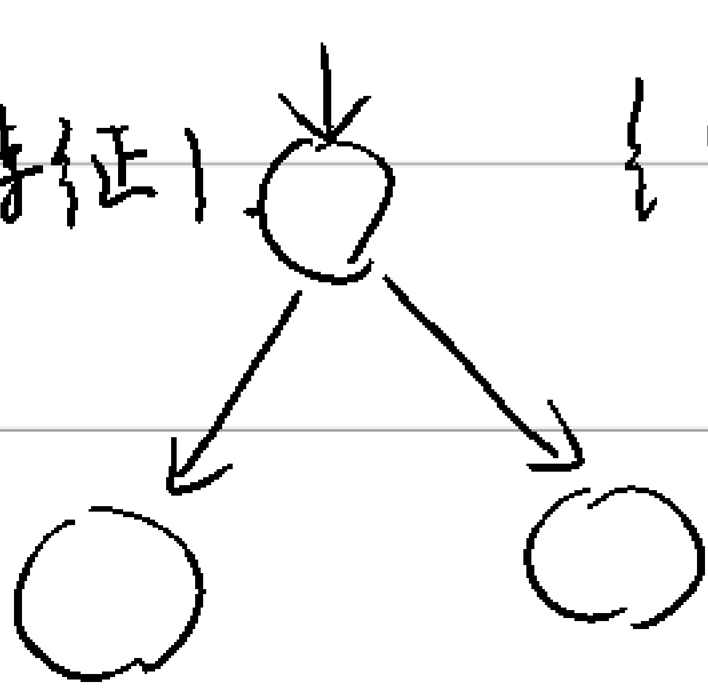
如何寻找树的形状 ？

Brute-Force Search (穷举) (X) 指数级

Greedy (√).

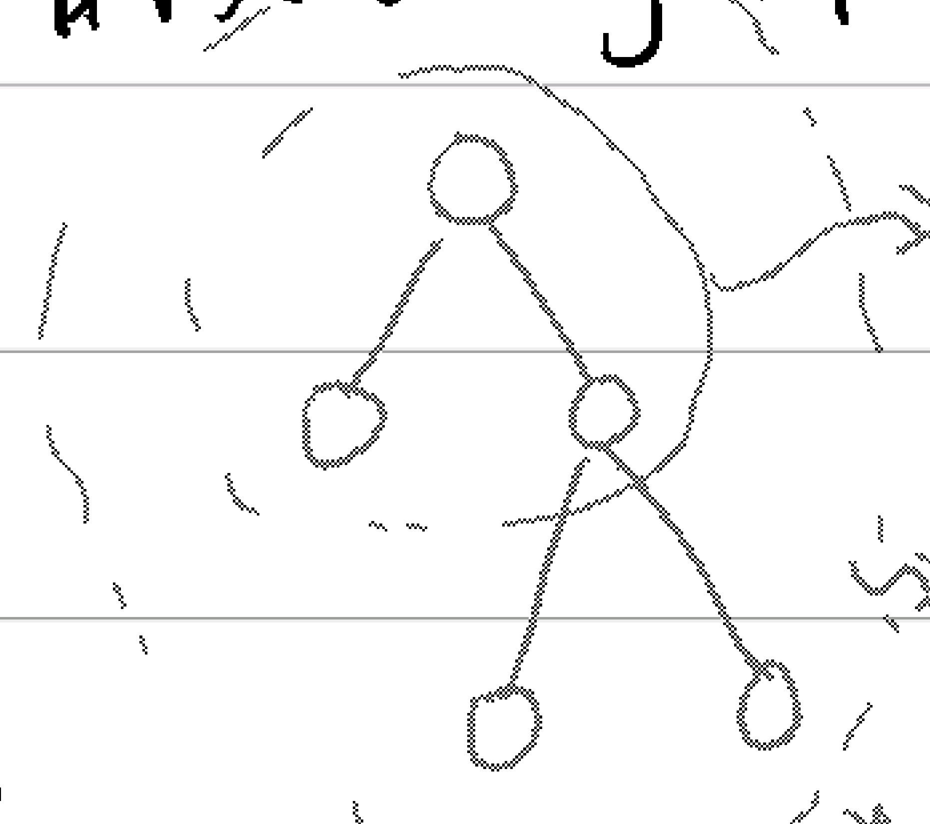
对于决策树：

特征1 $\{1,2,3,4,5,6,7,8\}$



$\{1,3,4\}$ $\{2,5,6,7,8\}$

这里把 Entropy 替换为 Obj 即可.



$Obj_K^{*(old)}$

$Obj_K^{*(now)}$

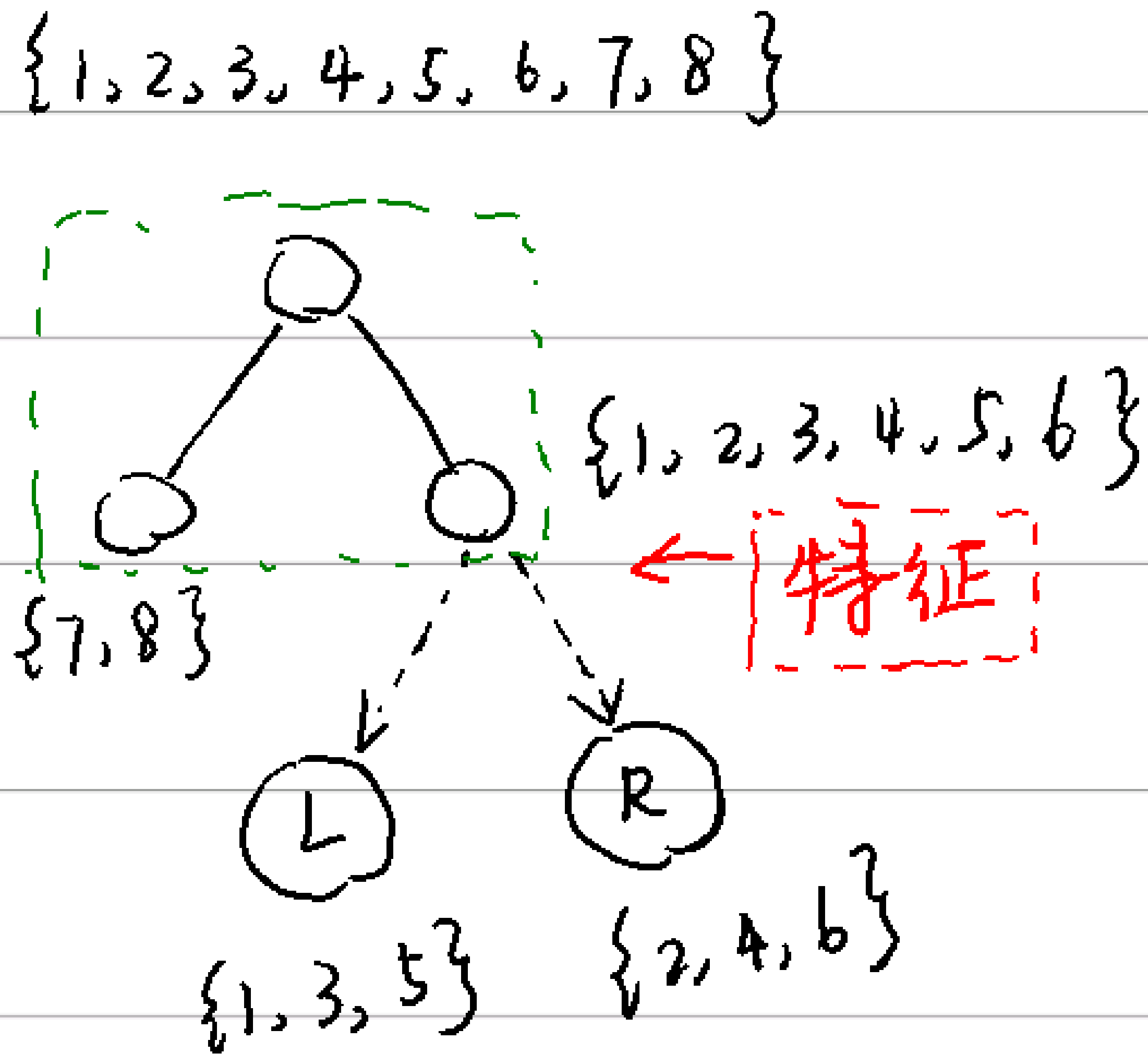
$Obj_K^{*(old)} - Obj_K^{*(now)}$

(Entropy) (Entropy)

特征1 score : 原不确定性 − 之后不确定性

maximize. Information Gain 信息增益.

寻找最好的split.



$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda} + \gamma T$$

$$\left\{\begin{aligned} \boxed{Obj^*_{old}} &= -\frac{1}{2}\left[\frac{(g_7+g_8)^2}{h_7+h_8+\lambda} + \frac{(g_1+\cdots+g_6)^2}{h_1+\cdots+h_6+\lambda}\right] + \gamma\cdot 2 \\ Obj^*_{new} &= -\frac{1}{2}\left[\frac{(g_7+g_8)^2}{h_7+h_8+\lambda} + \frac{(g_1+g_3+g_5)^2}{h_1+h_3+h_5+\lambda} + \frac{(g_2+g_4+g_6)^2}{h_2+h_4+h_6+\lambda}\right] + \gamma\cdot 3 \end{aligned}\right.$$

$$Obj^*_{old} - Obj^*_{new} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$

最大化

$$G_L = g_1+g_3+g_5 \qquad G_R = g_2+g_4+g_6$$

$$H_L = h_1+h_3+h_5 \qquad H_R = h_2+h_4+h_6$$