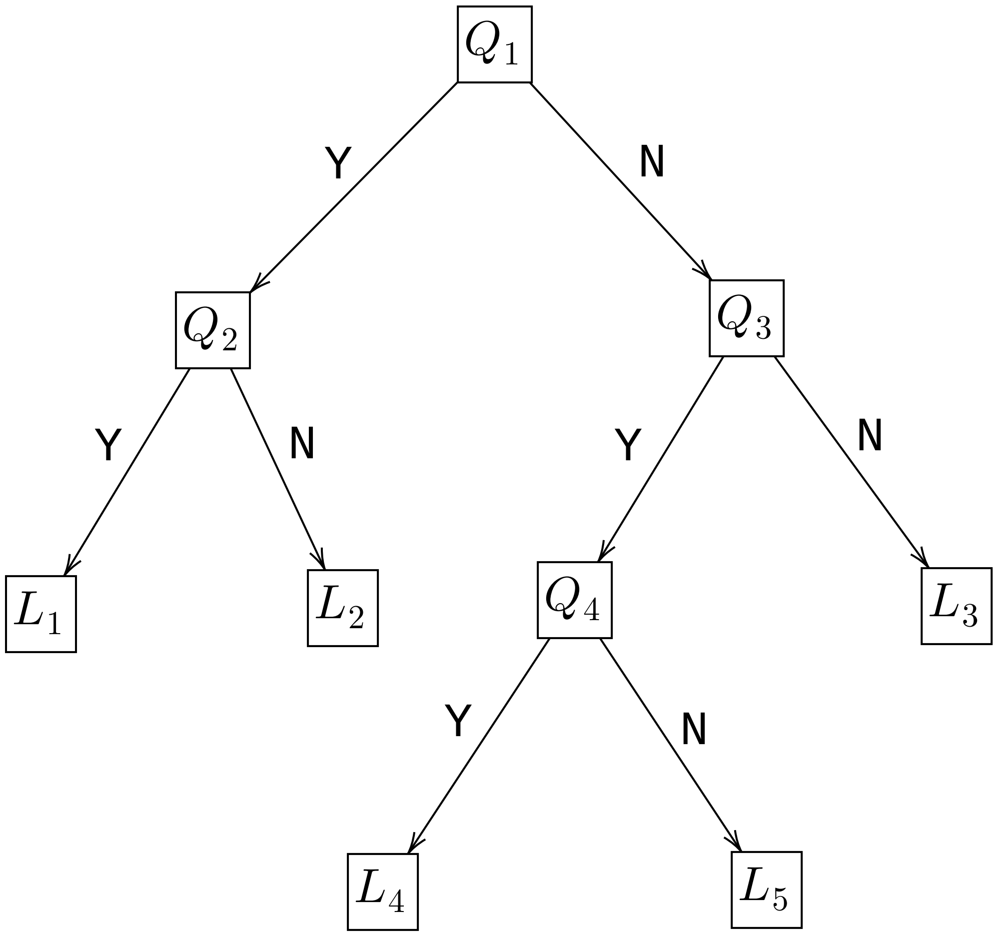
Decision Trees-1

Machine Learning Techniques

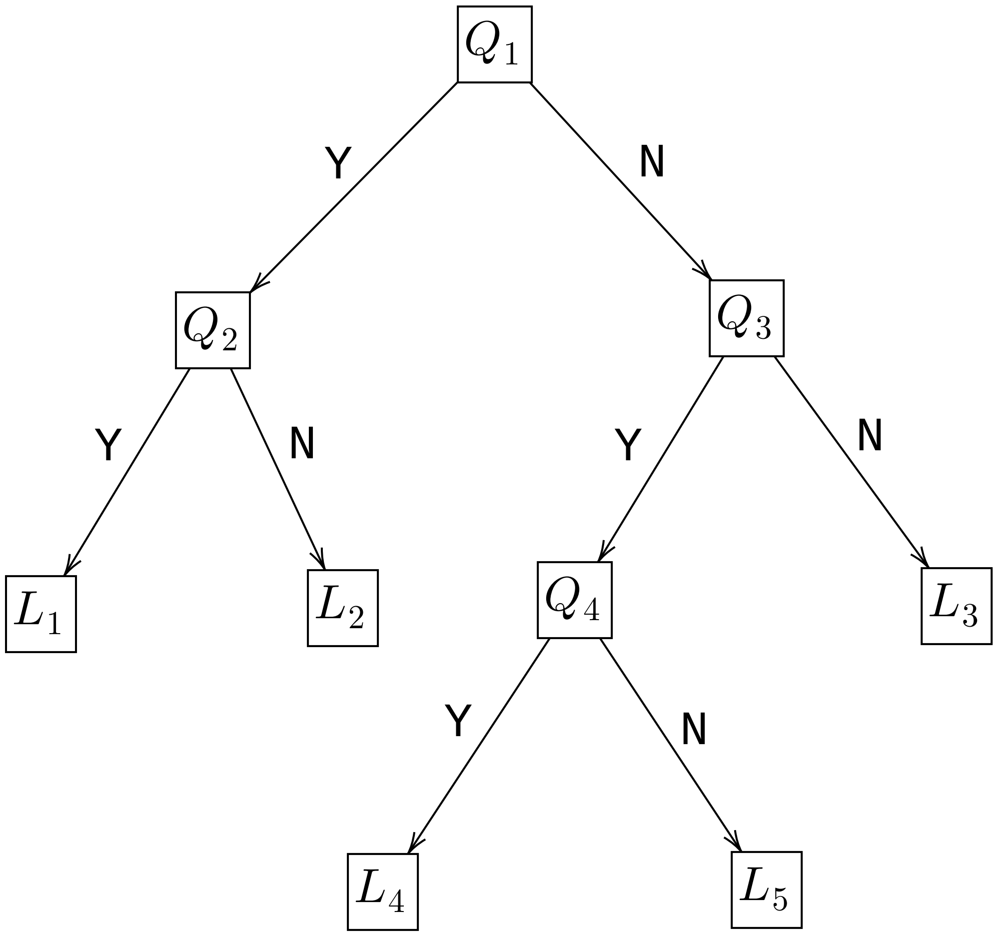
Karthik Thiagarajan

Decision Tree



Data Structure

Decision Tree

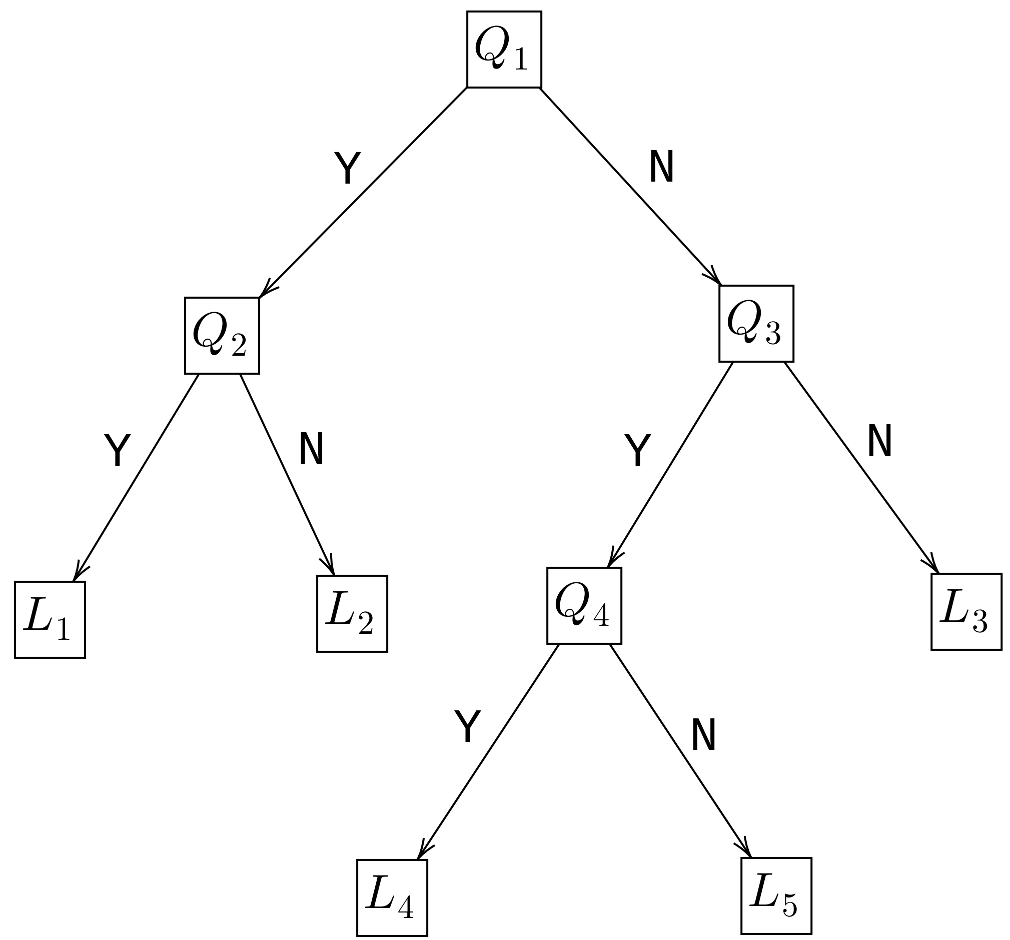


Data Structure

Tree

- Binary tree
- $Q_i :$ feature $<$ value
- L_i: label
- Depth $= 3$

Decision Tree



Data Structure

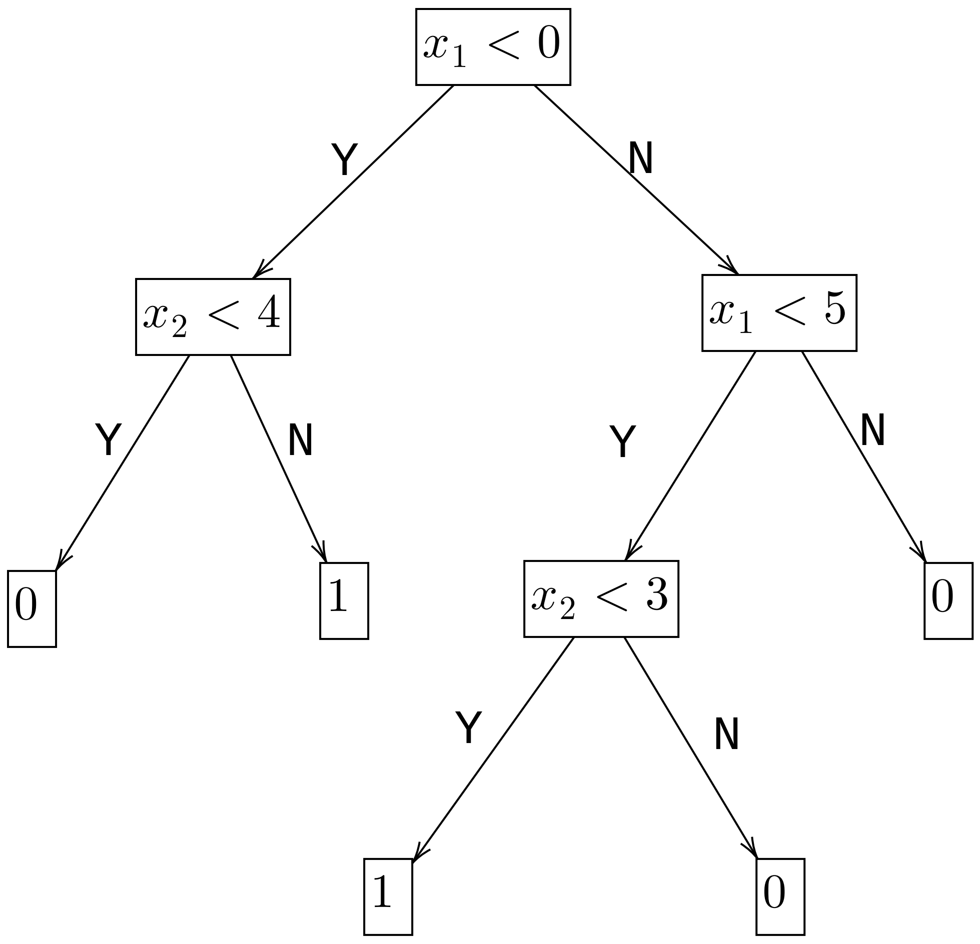
<u>Tree</u>

- Binary tree
- Q_i : feature $<$ value
- L_i: label
- Depth $= 3$

<u>Nodes</u>

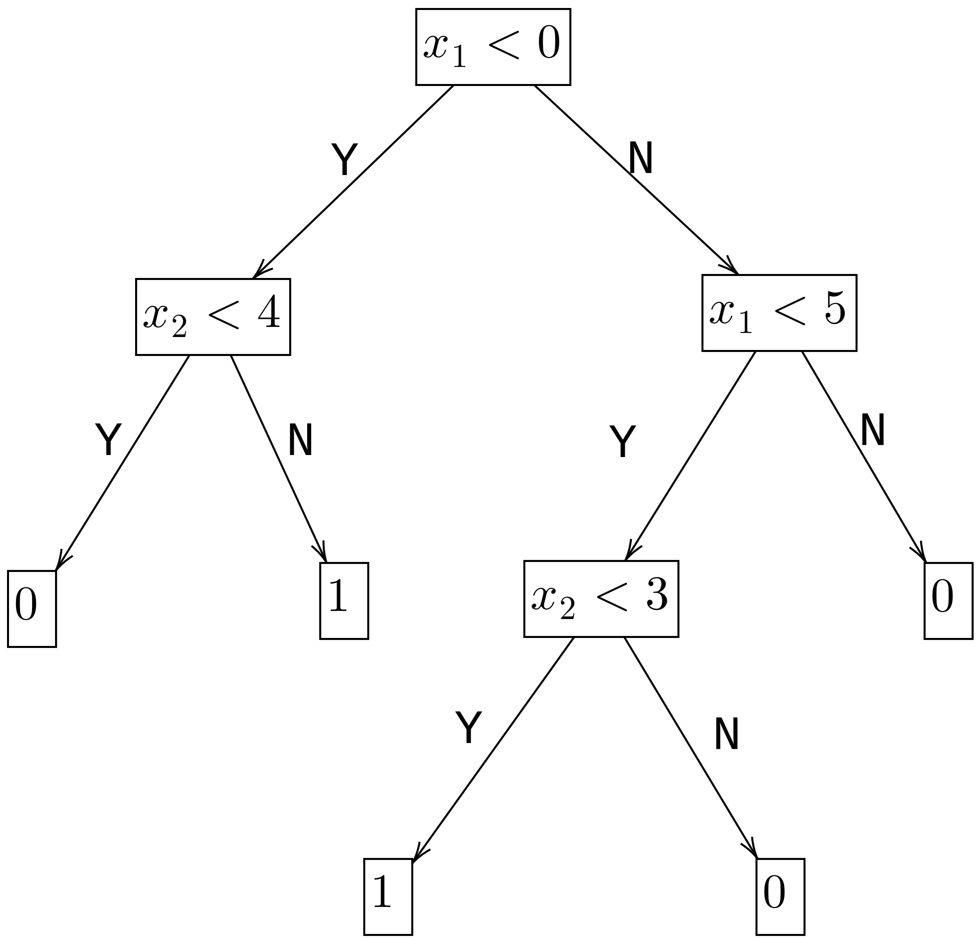
- Root node: Q_1
- Internal nodes: Q_1, Q_2, Q_3, Q_4
  - left $\rightarrow$ yes
  - right $\rightarrow$ no
  - questions
- Leaves: L_1, L_2, L_3, L_4, L_5
  - predictions

Decision Tree



Function

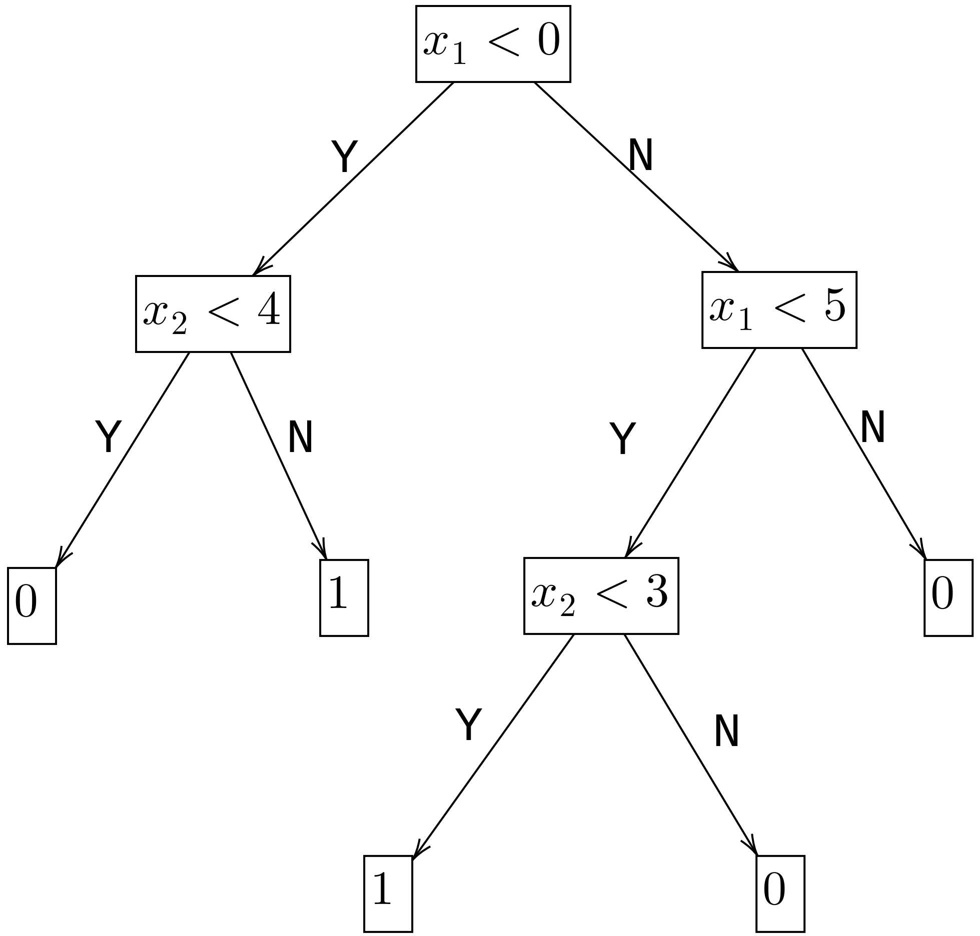
Decision Tree



Function

$$h : \mathbb{R}^d \to \{0, 1\}$$
$$h(\mathbf{x}) = y$$

Decision Tree

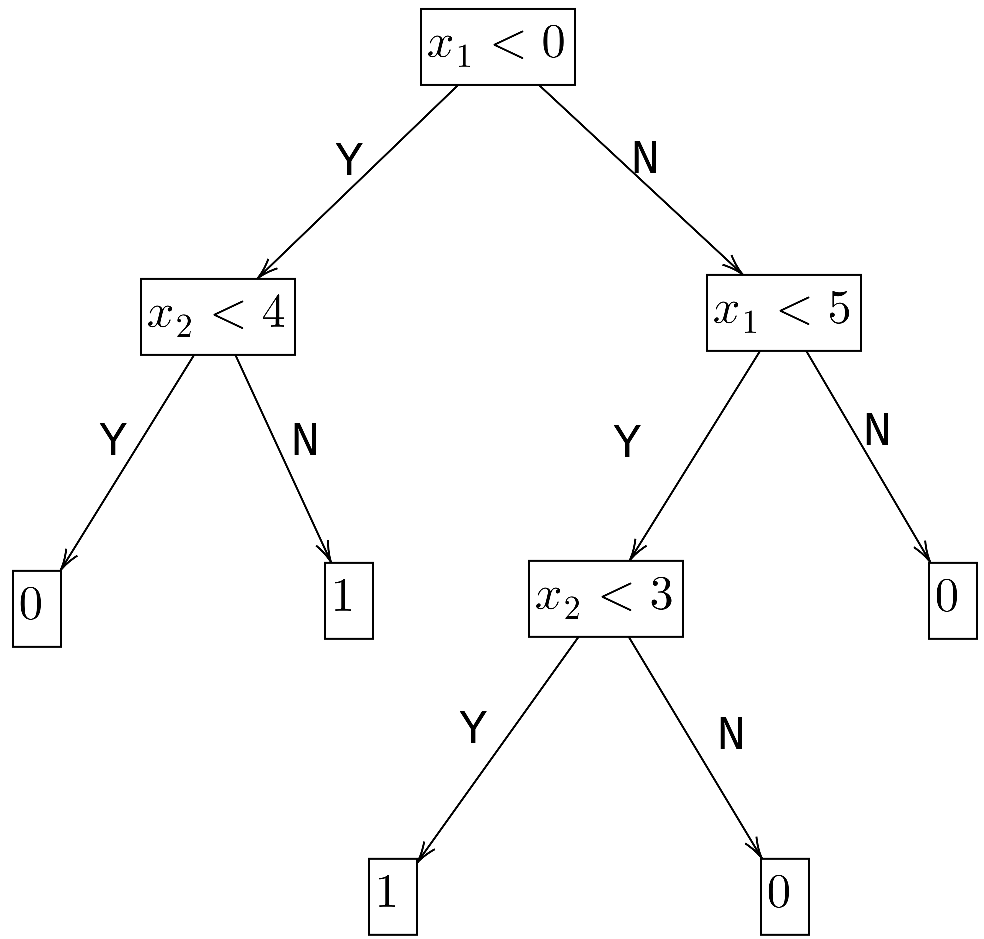


Function

$$h : \mathbb{R}^d \to \{0, 1\}$$
$$h(\mathbf{x}) = y$$

x_1	x_2	y
3	4	
-2	5	
10	4	

Decision Tree

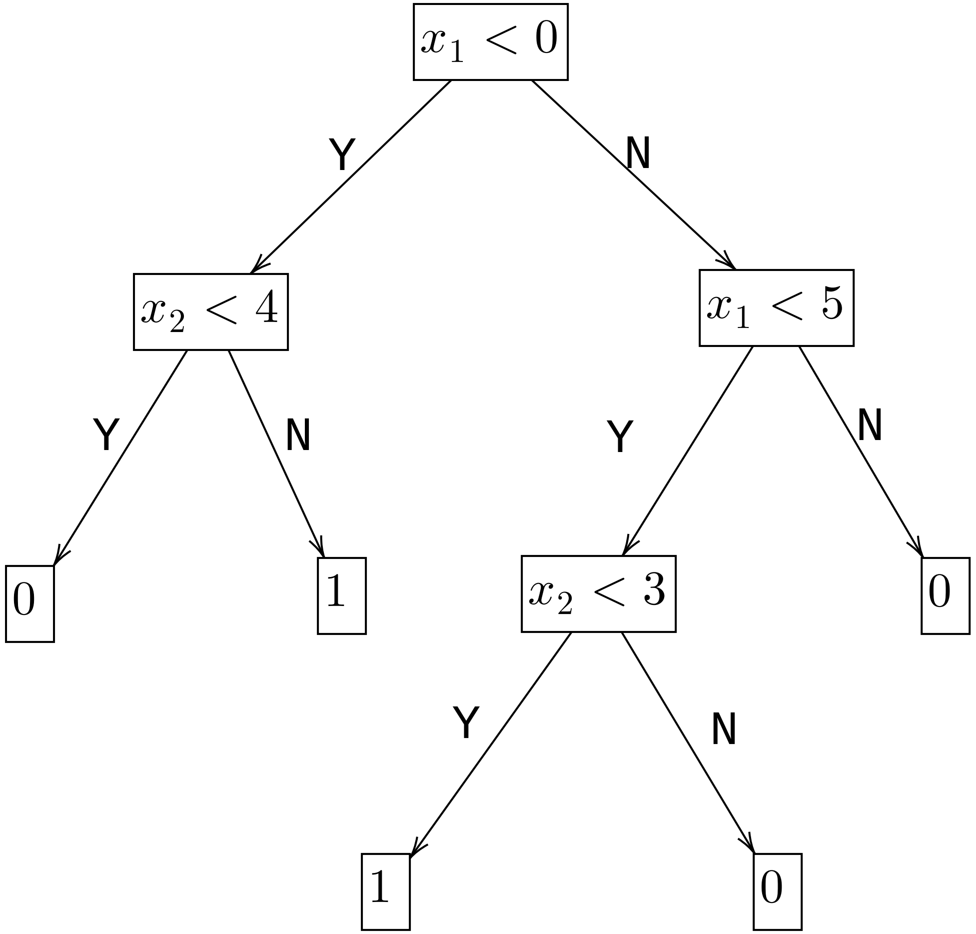


Function

$$h : \mathbb{R}^d \to \{0, 1\}$$
$$h(\mathbf{x}) = y$$

x_1	x_2	y
3	4	0
-2	5	
10	4	

Decision Tree

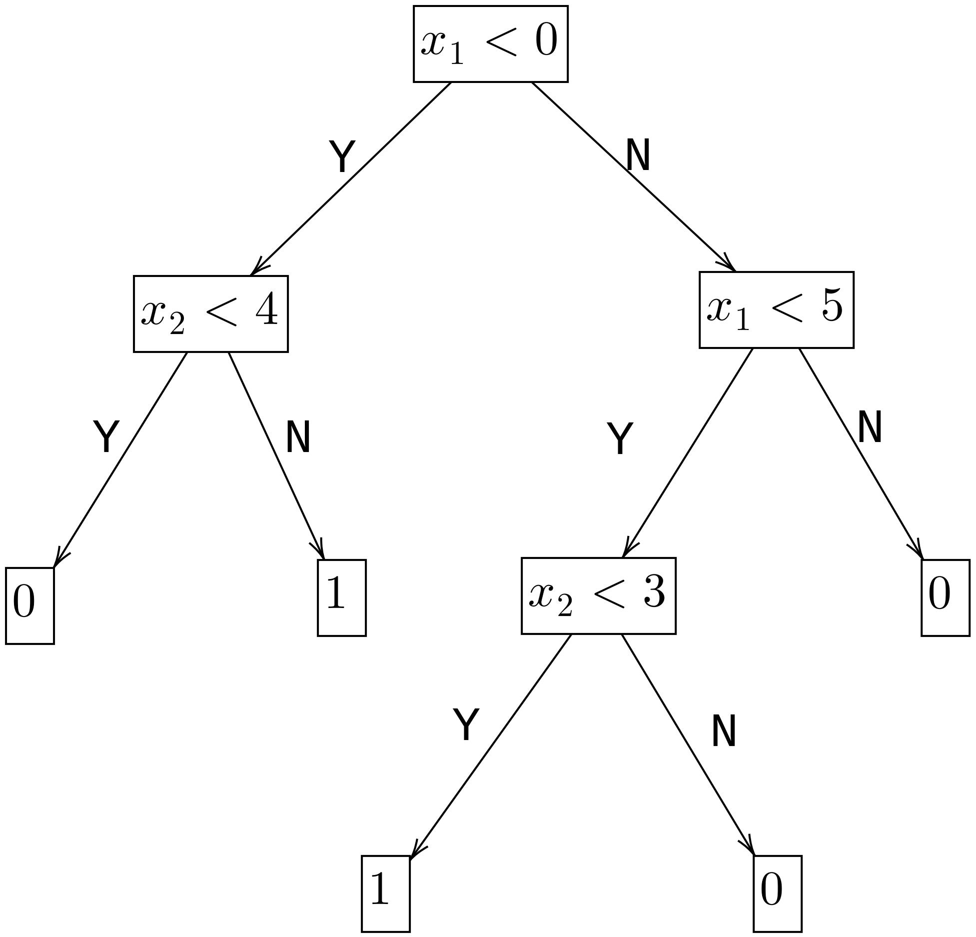


Function

$$h : \mathbb{R}^d \to \{0, 1\}$$
$$h(\mathbf{x}) = y$$

x_1	x_2	y
3	4	0
-2	5	1
10	4	

Decision Tree

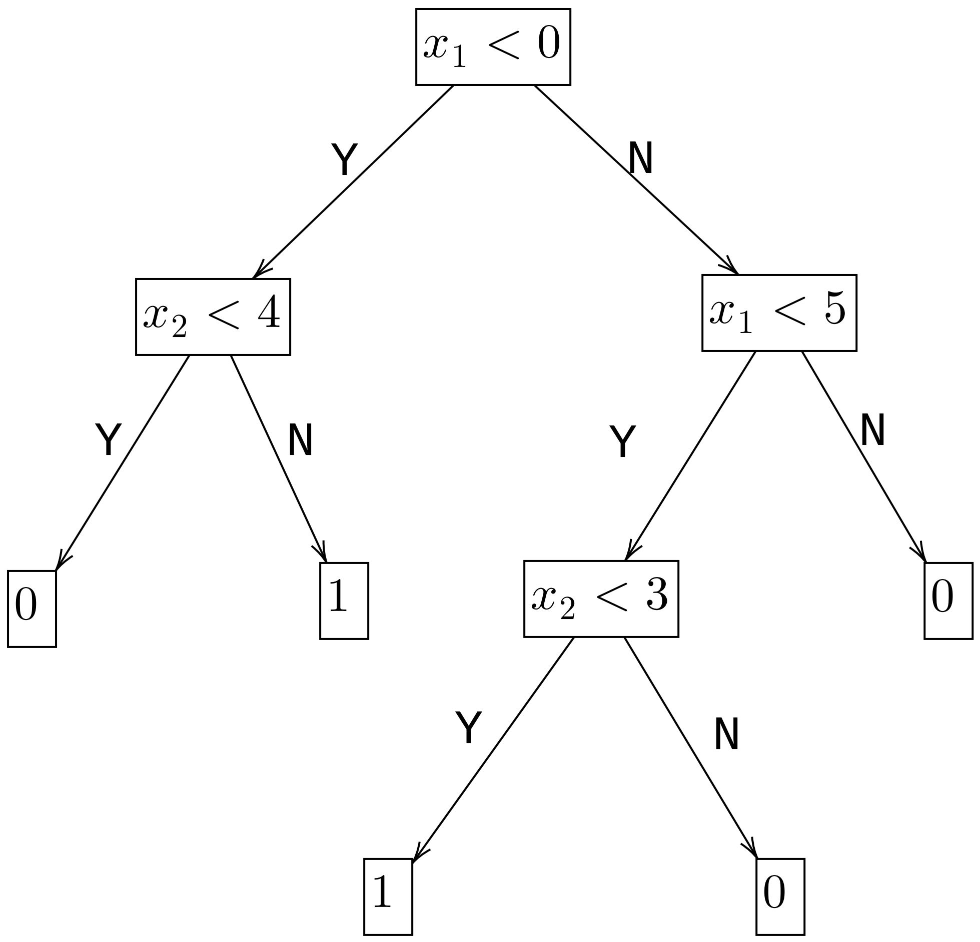


Function

$$h : \mathbb{R}^d \to \{0, 1\}$$
$$h(\mathbf{x}) = y$$

x_1	x_2	y
3	4	0
-2	5	1
10	4	0

Decision Tree



Function

$$h : \mathbb{R}^d \to \{0, 1\}$$
$$h(\mathbf{x}) = y$$

x_1	x_2	y
3	4	0
-2	5	1
10	4	0

Prediction

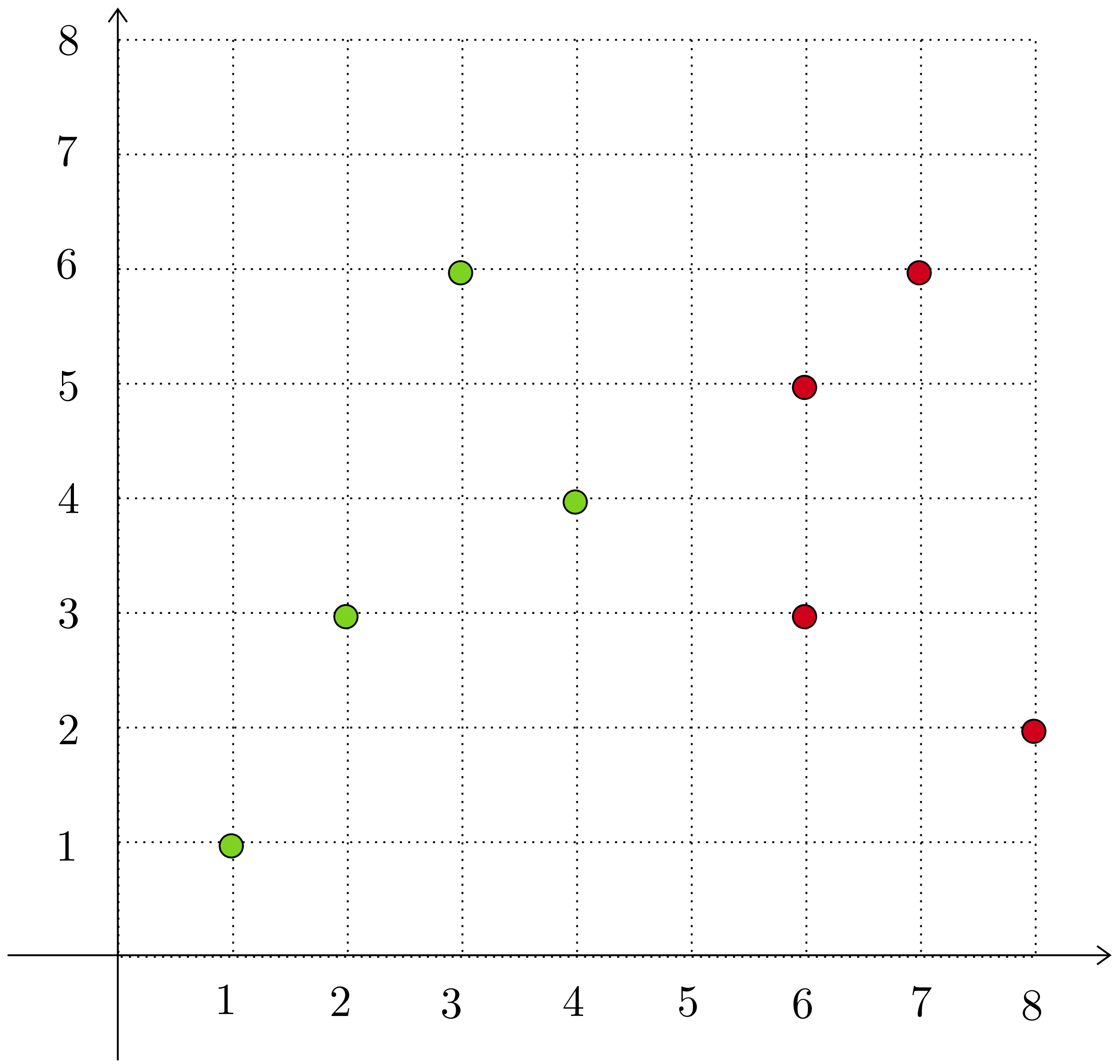
- Prediction $\to$ Traversal from root to leaf
- Longer paths $\to$ more complex decisions
- Interpretable

Growing a Tree

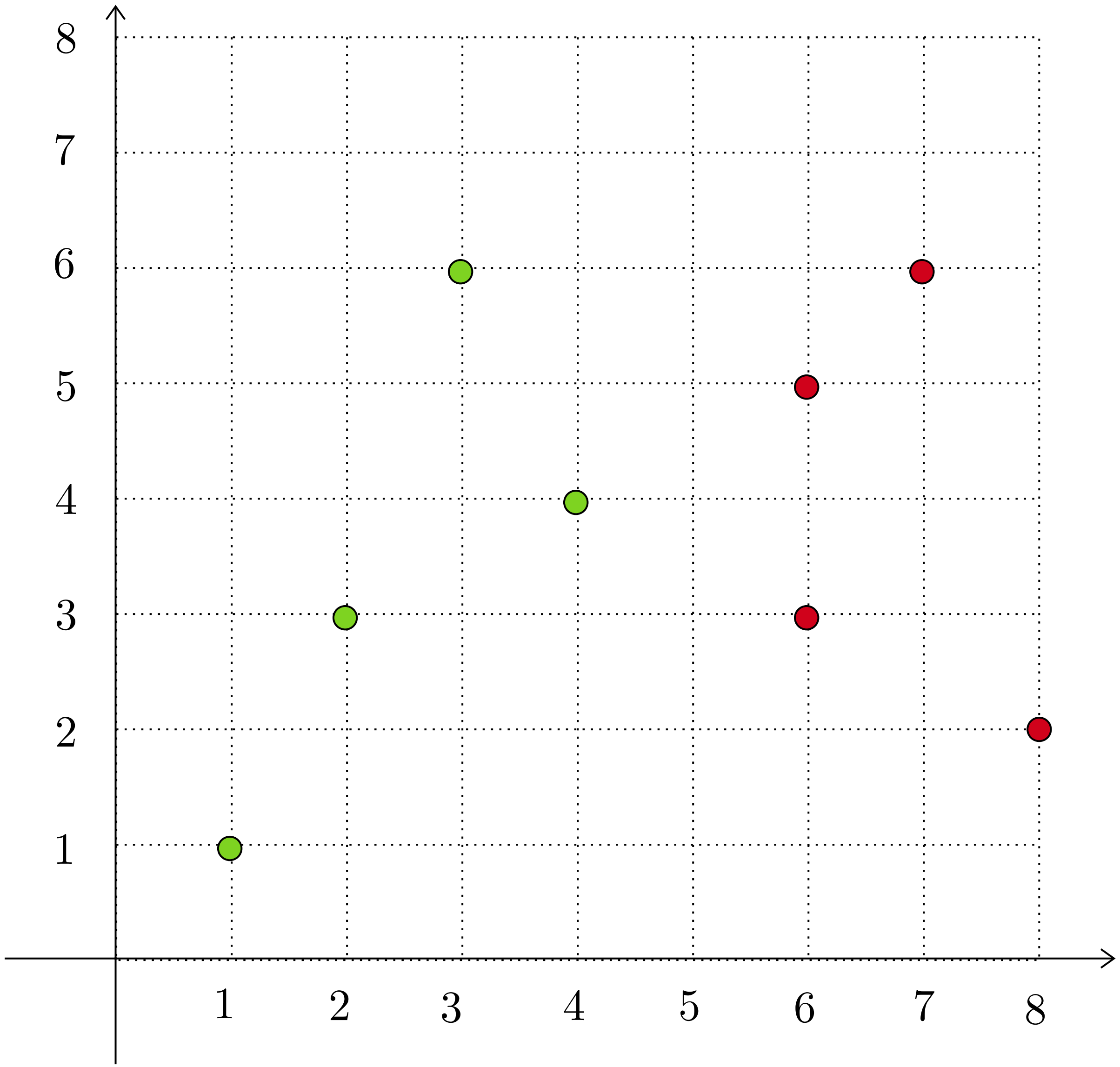
$$D_{\texttt{train}} = \{(\mathbf{x}_1, y_1), \;\cdots, (\mathbf{x}_n, y_n)\}$$

$$\mathbf{x}_i \in \mathbb{R}^d, \;\; y_i \in \{0, 1\}$$

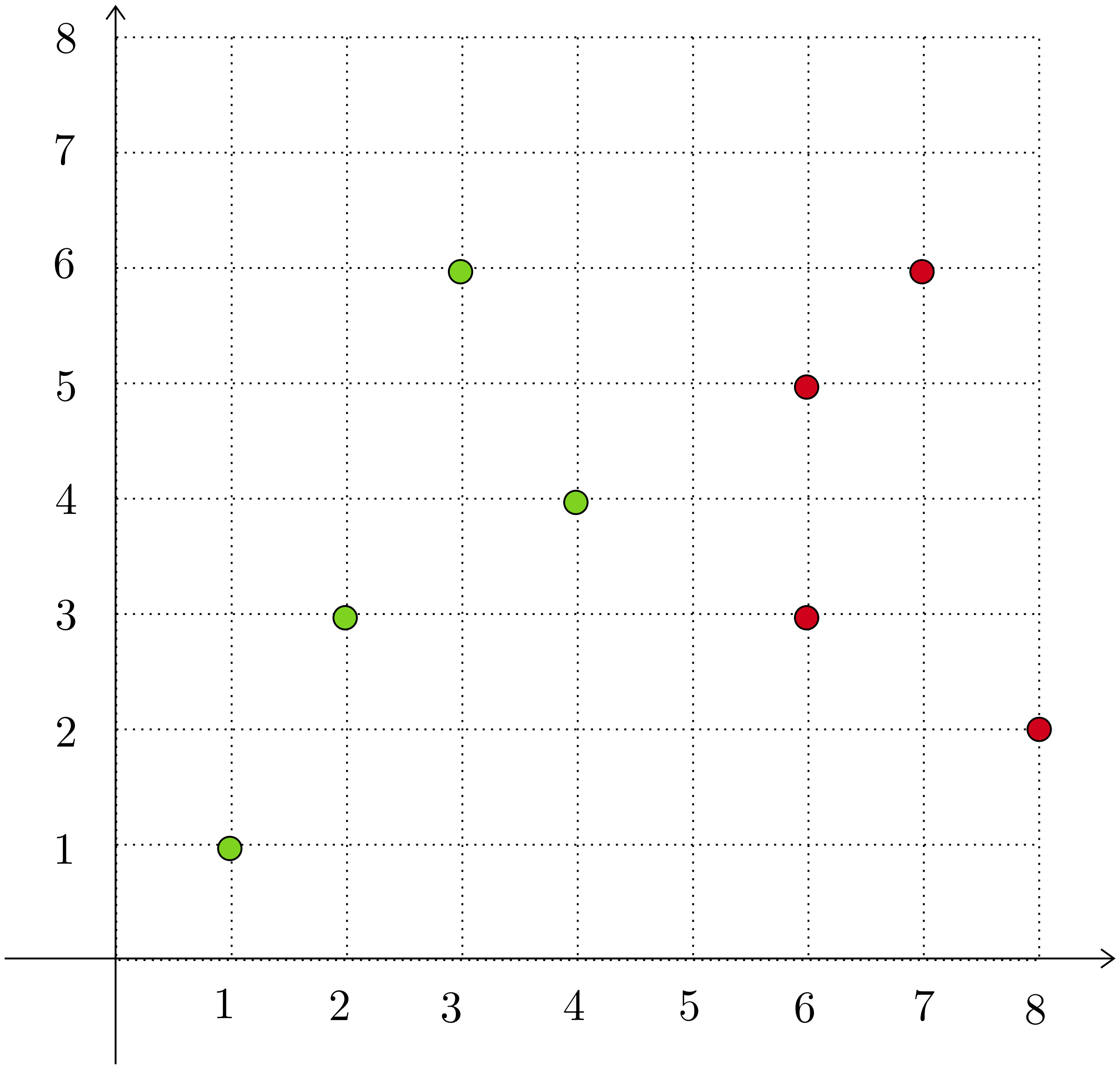
x_1	x_2	y
1	1	1
2	3	1
3	6	1
4	4	1
6	3	0
6	5	0
7	6	0
8	2	0



Growing a Tree

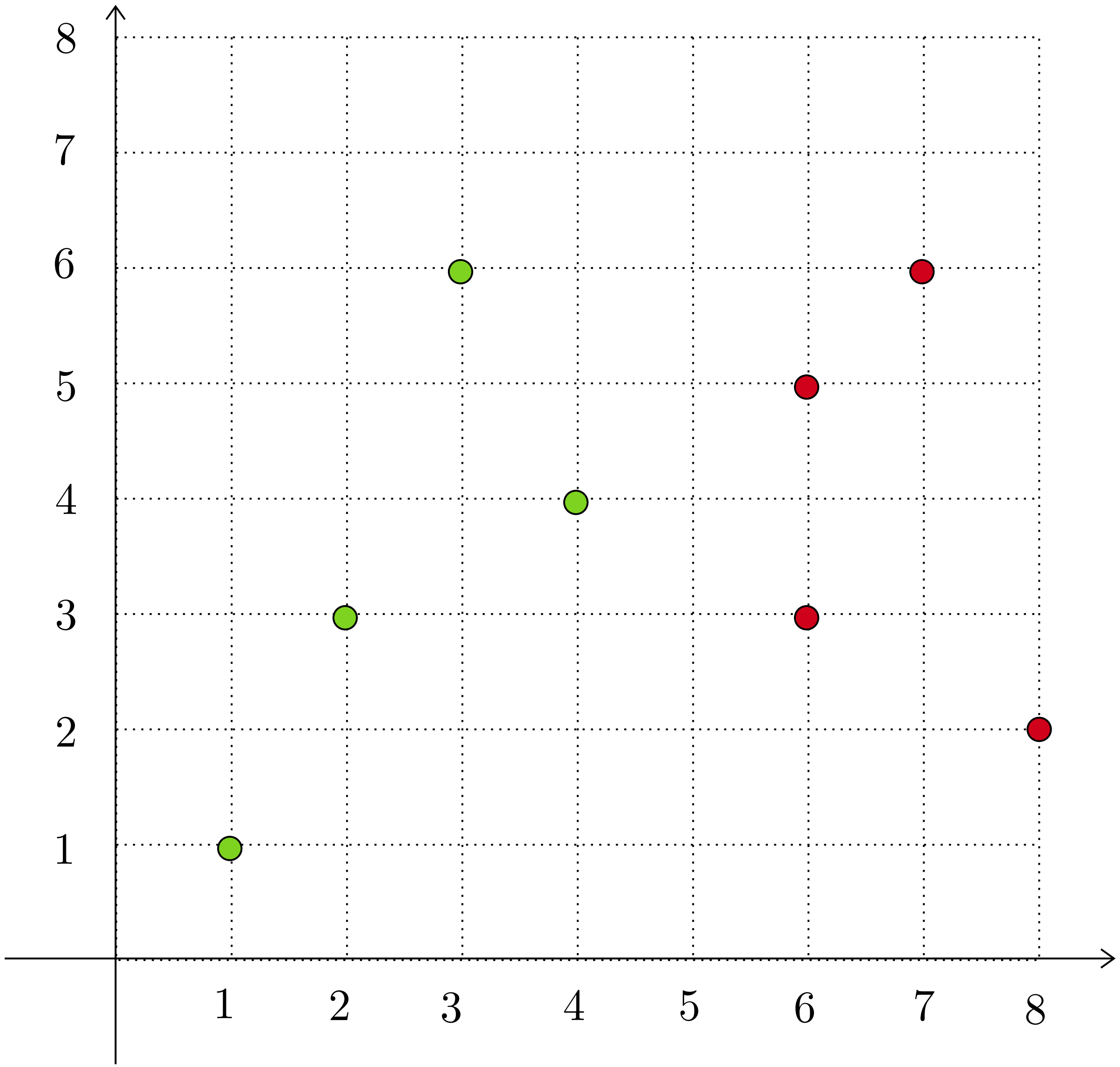


Growing a Tree



Good Q_1:

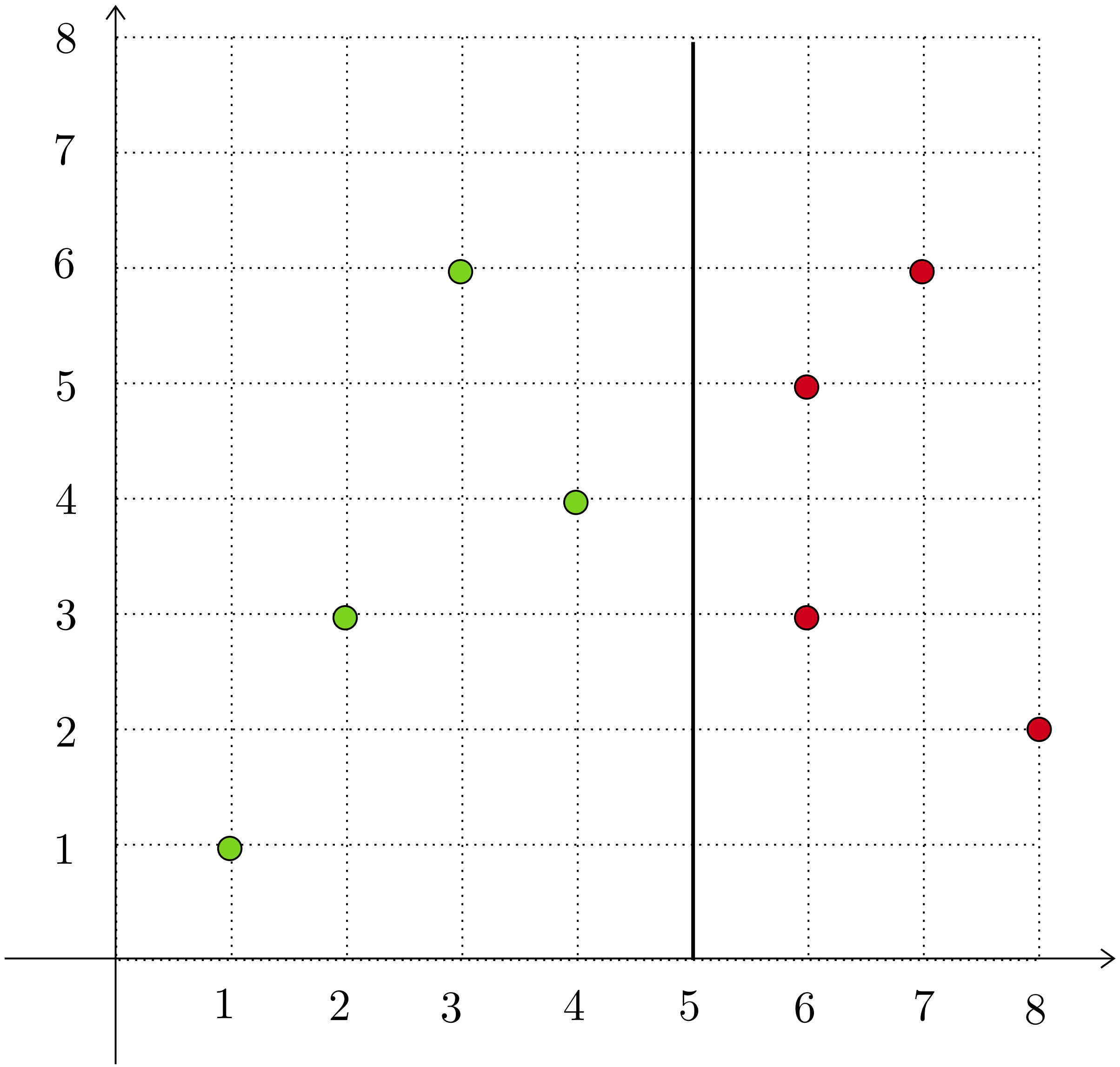
Growing a Tree



Good Q_1:

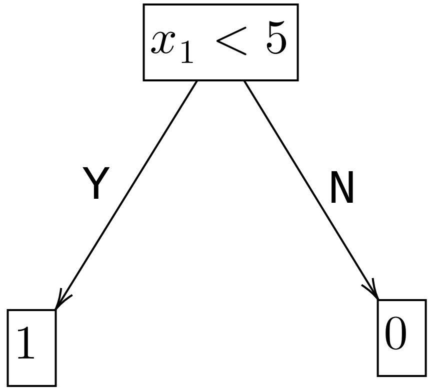
- $x_1 < 5$
- $x_1 \leq 4$
- $x_1 < 6$
- $x_1 < a$, $a \in (4, 6)$

Growing a Tree



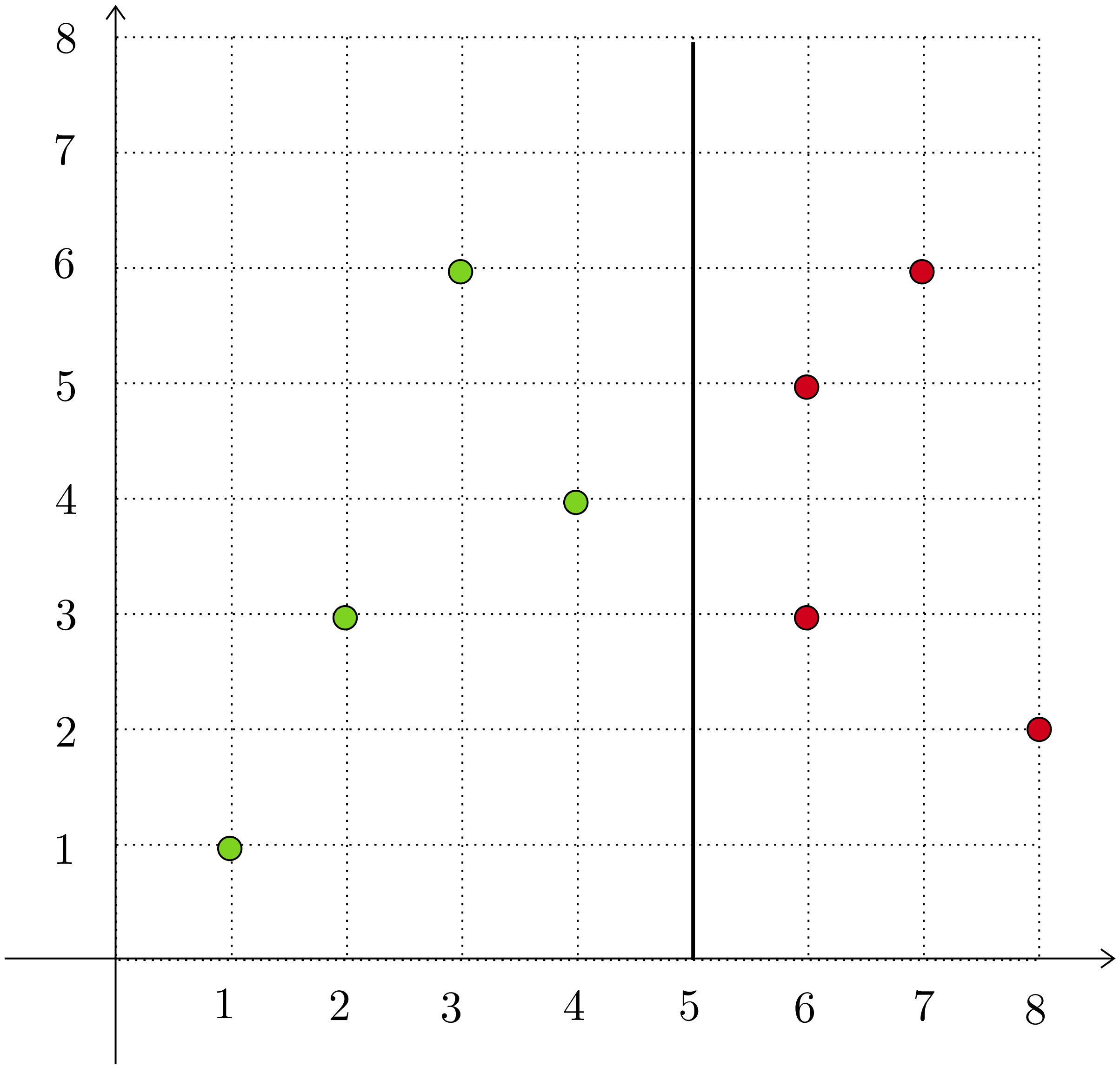
Good Q_1:

- $x_1 < 5$
- $x_1 \leq 4$
- $x_1 < 6$
- $x_1 < a, \quad a \in (4, 6)$



Decision Stump

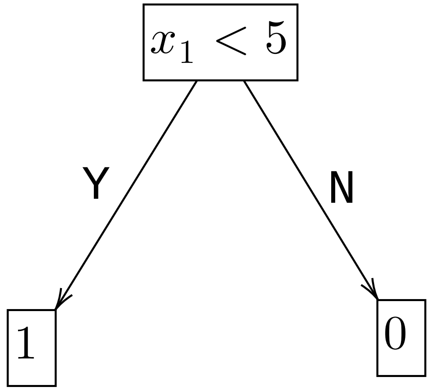
Growing a Tree



Node Purity

Good Q_1:

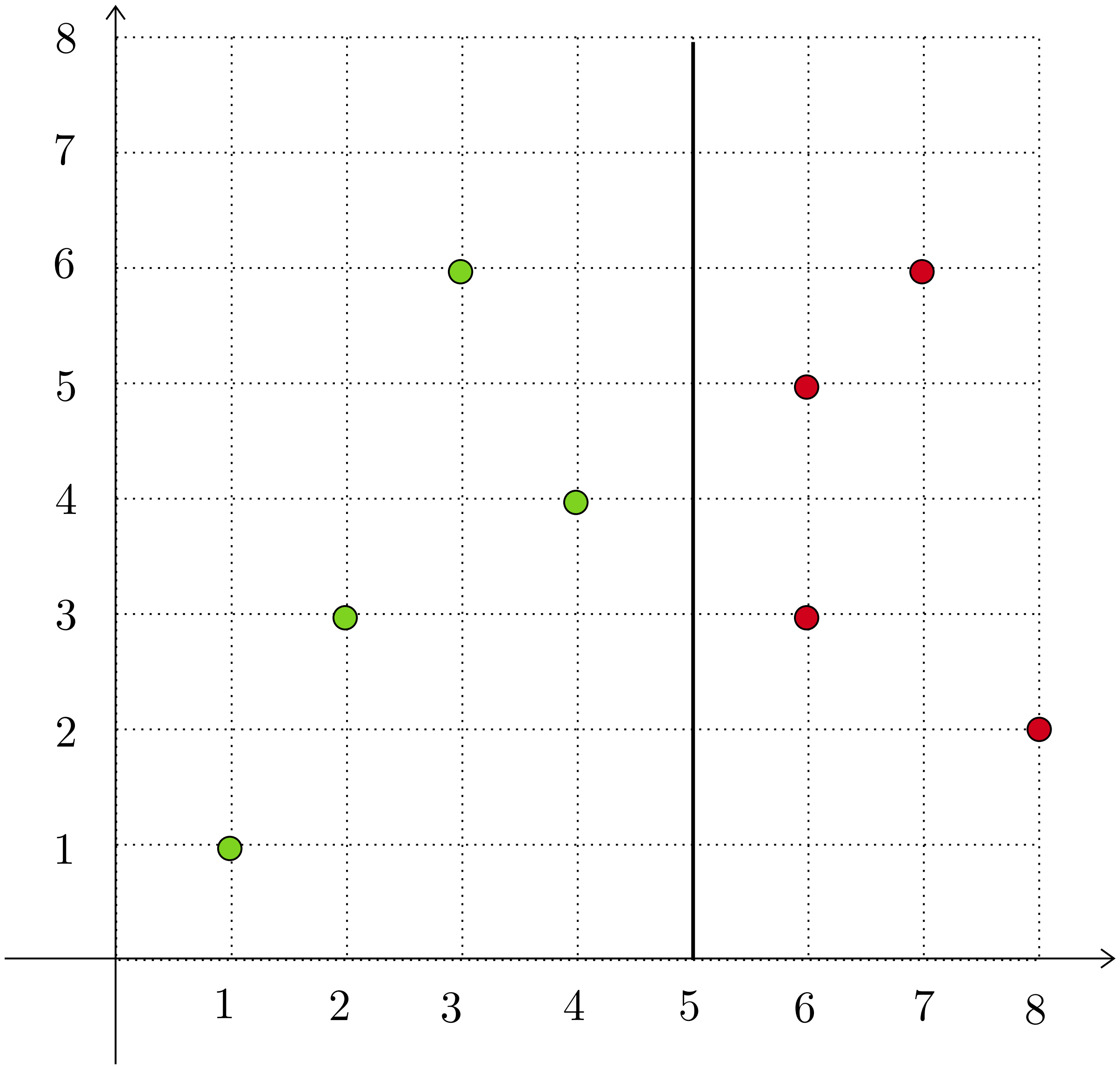
- $x_1 < 5$
- $x_1 \leq 4$
- $x_1 < 6$
- $x_1 < a, \quad a \in (4, 6)$



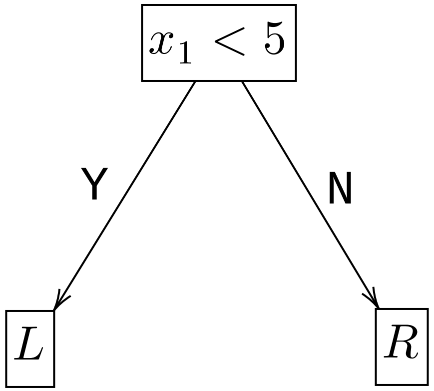
Decision Stump

- Q_1 partitions the dataset
- Decision stump
- Clean or "pure" partitions

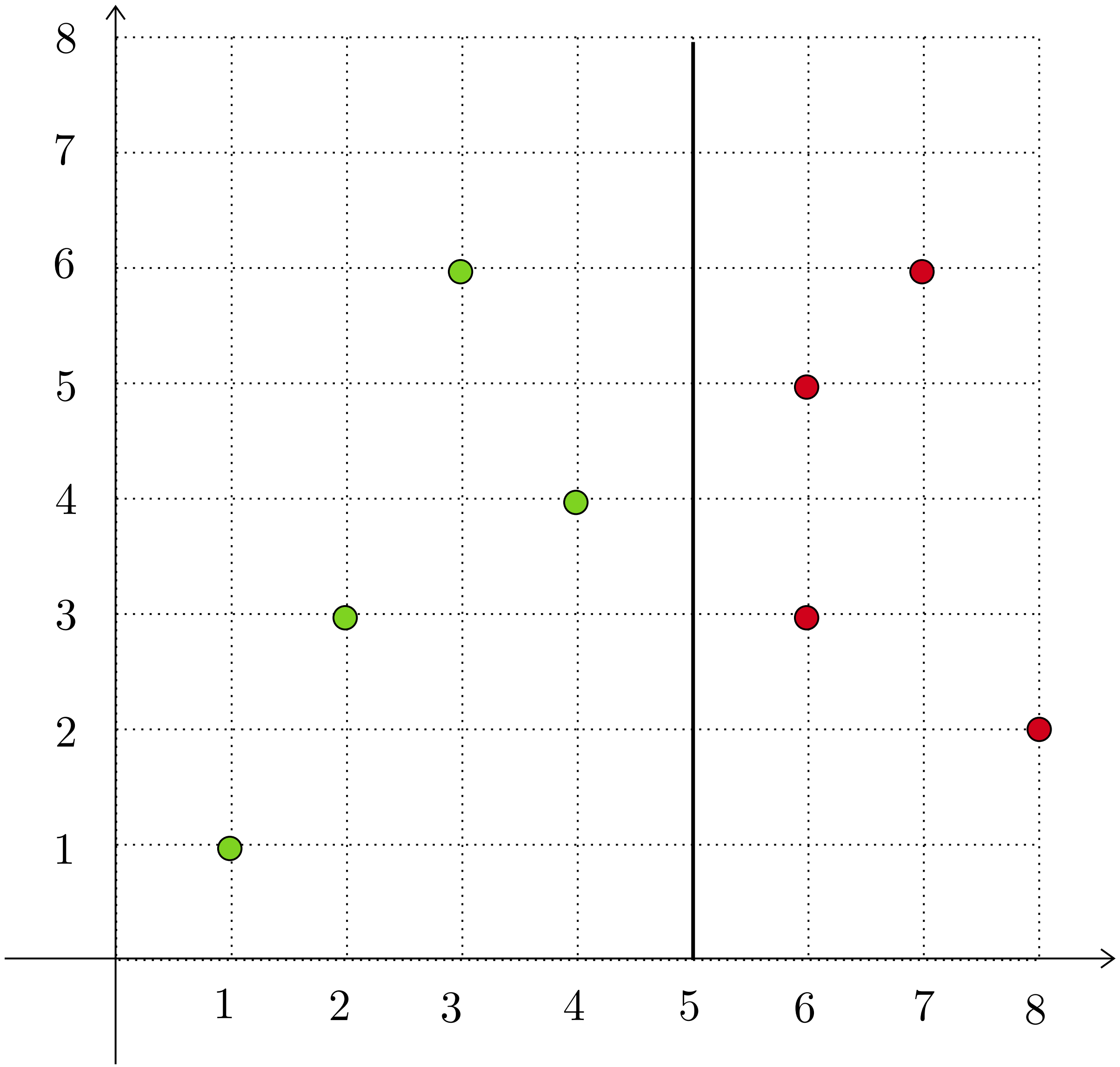
Growing a Tree



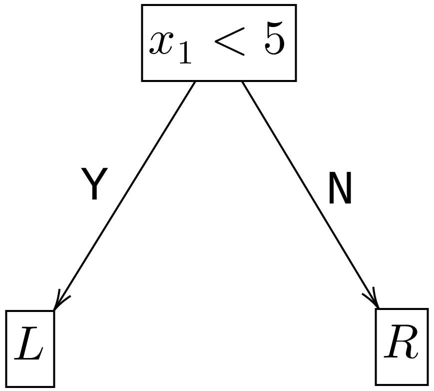
Node Impurity



Growing a Tree



Node Impurity

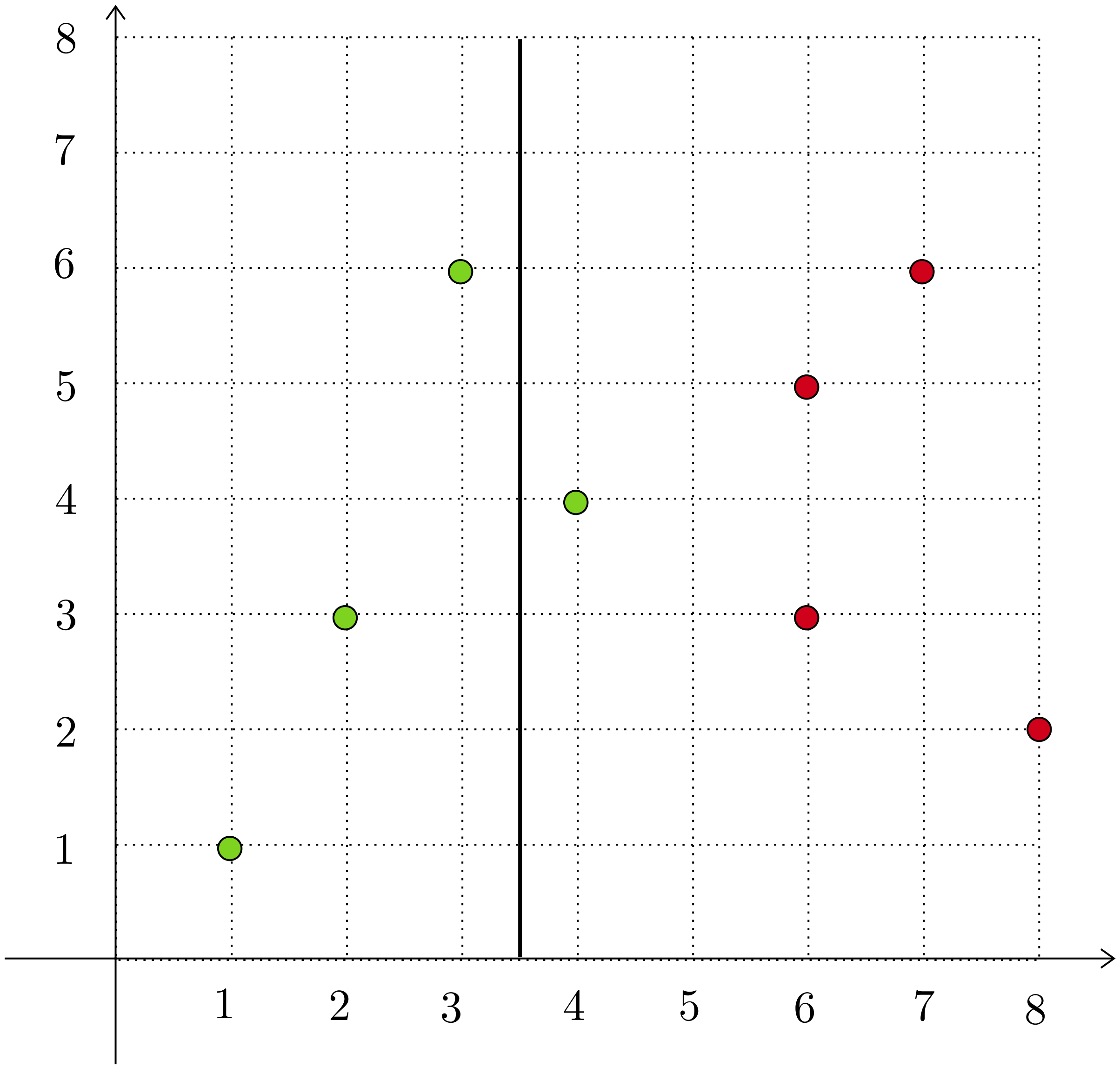


$$L = \{1, 1, 1, 1\}, \;\; R = \{0, 0, 0, 0\}$$

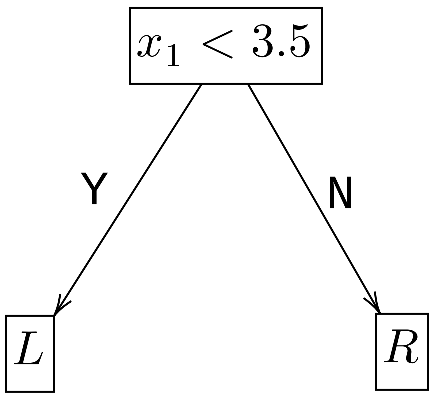
$$\begin{aligned} p_{_L} &= \frac{\sum_{y_i \in L} \mathbf{1}[y_i = 1]}{|L|} \\ &= \frac{4}{4} \\ &= 1 \end{aligned}$$

$$\begin{aligned} p_{_R} &= \frac{\sum_{y_i \in R} \mathbf{1}[y_i = 1]}{|R|} \\ &= \frac{0}{4} \\ &= 0 \end{aligned}$$

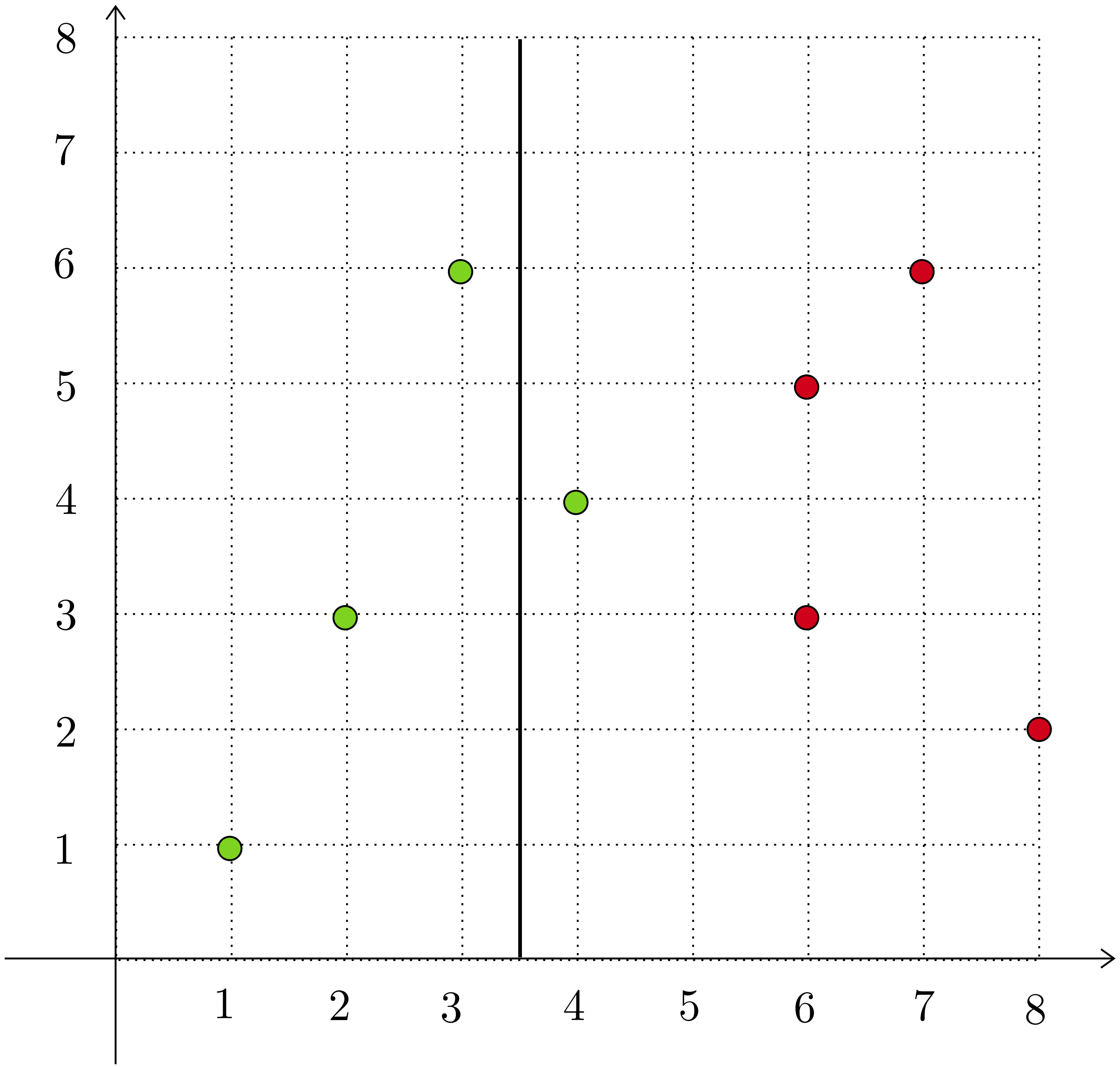
Growing a Tree



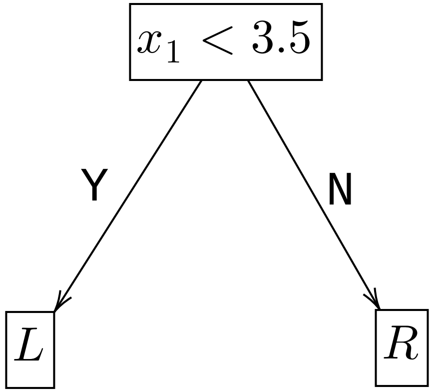
Node Impurity



Growing a Tree



Node Impurity

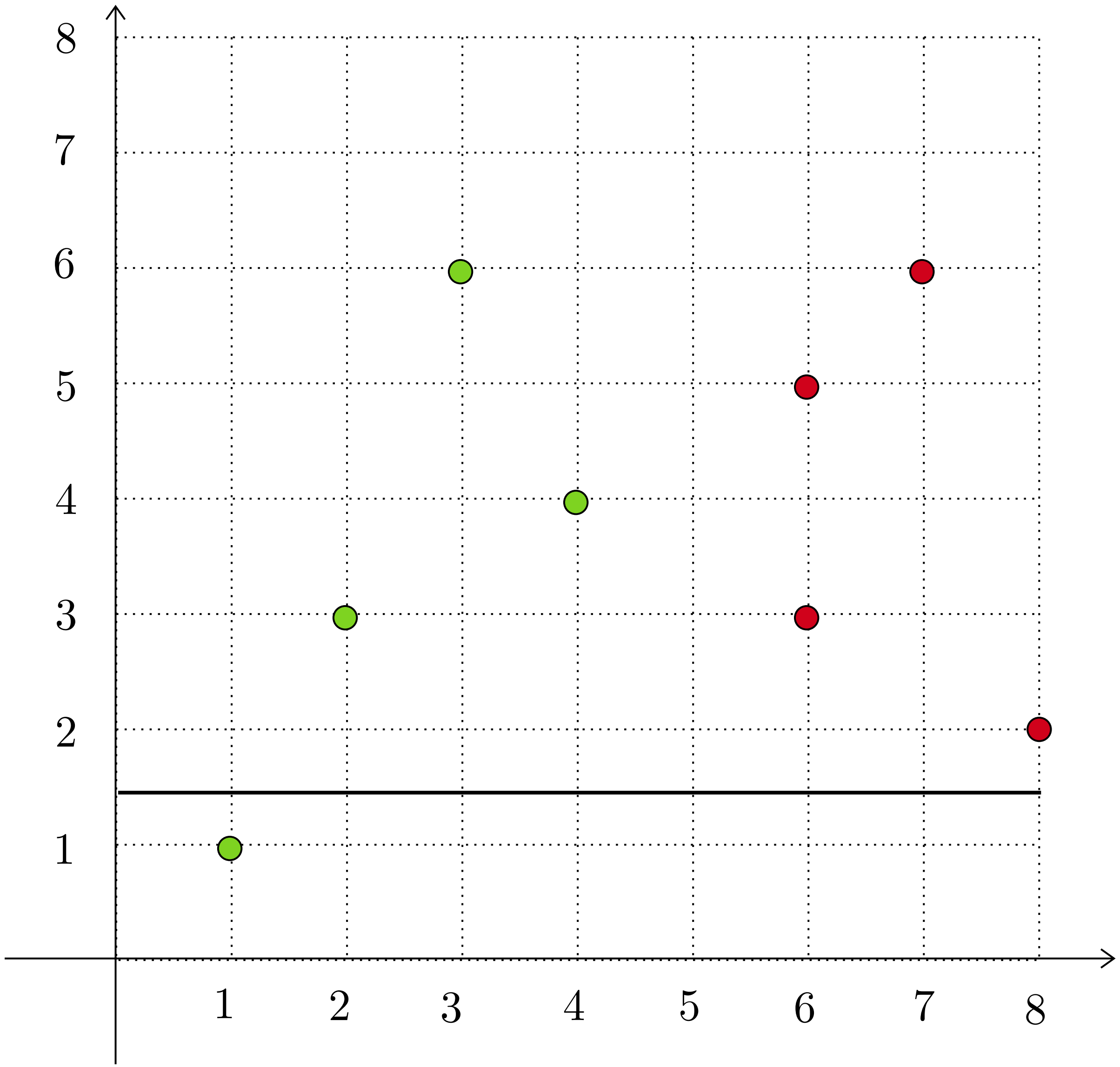


$$L = \{1, 1, 1\},\ R = \{1, 0, 0, 0, 0\}$$

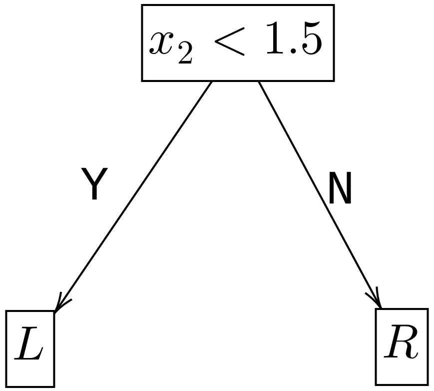
$$p_L = \frac{\sum_{y_i \in L} \mathbf{1}[y_i = 1]}{|L|} = \frac{3}{3} = 1$$

$$p_R = \frac{\sum_{y_i \in R} \mathbf{1}[y_i = 1]}{|R|} = \frac{1}{5} = 0.2$$

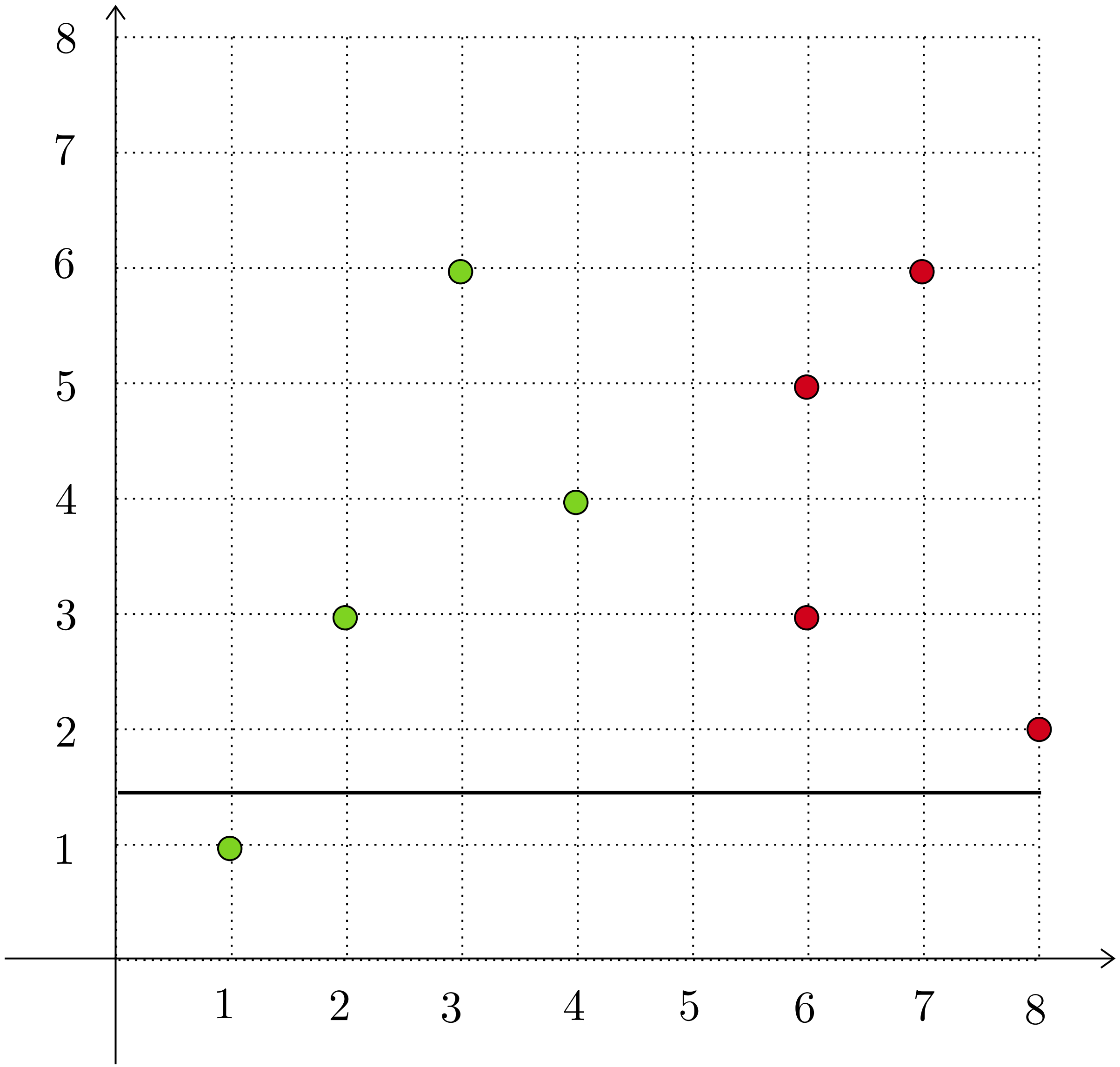
Growing a Tree



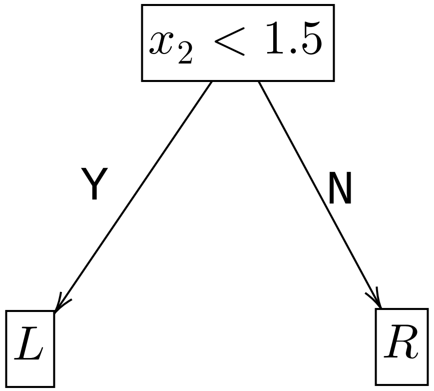
Node Impurity



Growing a Tree



Node Impurity

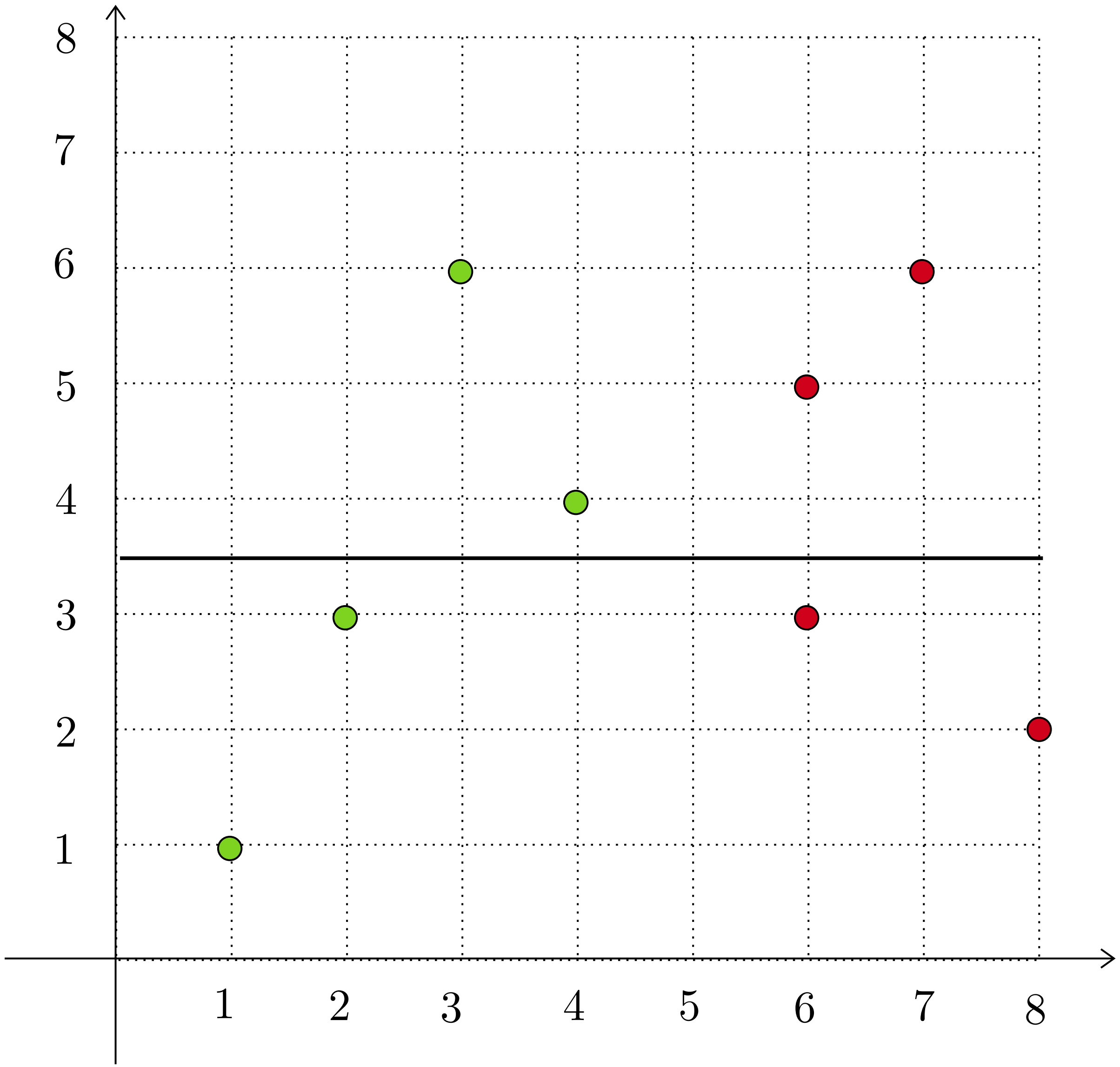


$$L = \{1, \}, \;\; R = \{1, 1, 1, 0, 0, 0, 0\}$$

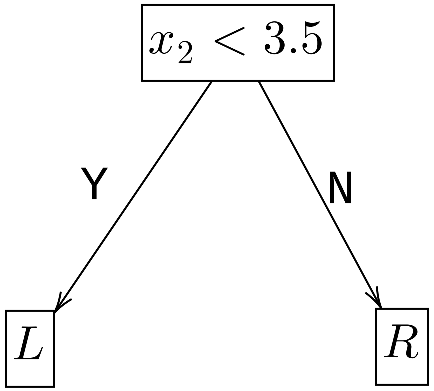
$$p_L = \frac{\sum_{y_i \in L} \mathbf{1}[y_i = 1]}{|L|} = \frac{1}{1} = 1$$

$$p_R = \frac{\sum_{y_i \in R} \mathbf{1}[y_i = 1]}{|R|} = \frac{3}{7} = 0.43$$

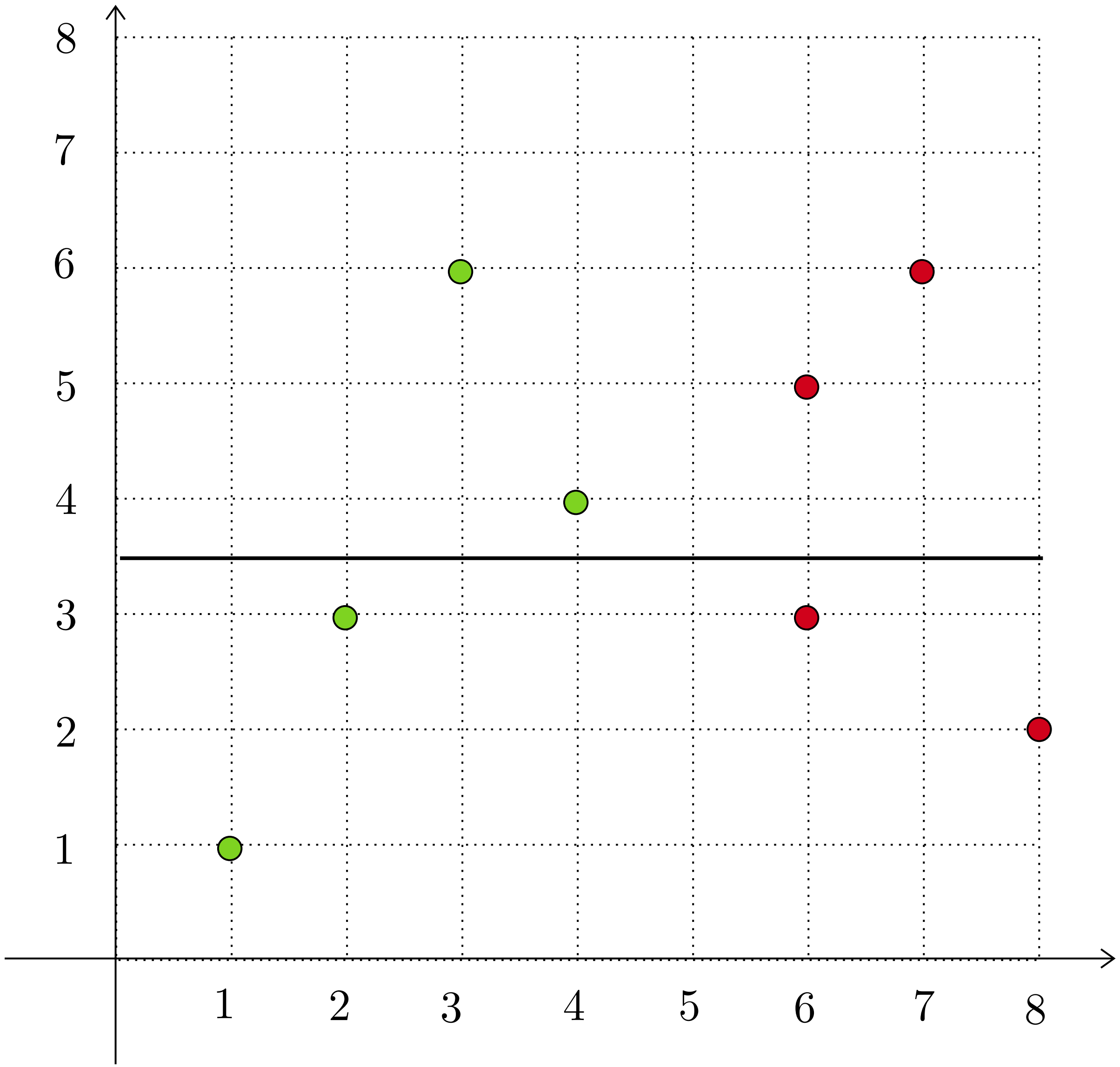
Growing a Tree



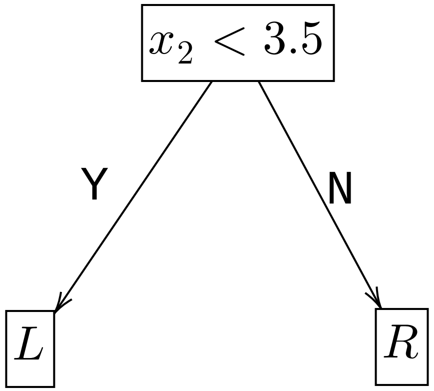
Node Impurity



Growing a Tree



Node Impurity



$$L = \{1, 1, 0, 0\},\ \ R = \{1, 1, 0, 0\}$$

$$p_{_L} = \frac{\sum_{y_i \in L} \mathbf{1}[y_i = 1]}{|L|} = \frac{2}{4} = 0.5$$

$$p_{_R} = \frac{\sum_{y_i \in R} \mathbf{1}[y_i = 1]}{|R|} = \frac{2}{4} = 0.5$$

Node Impurity

Impurity measure

p	Impurity
Low	
Medium	
High	

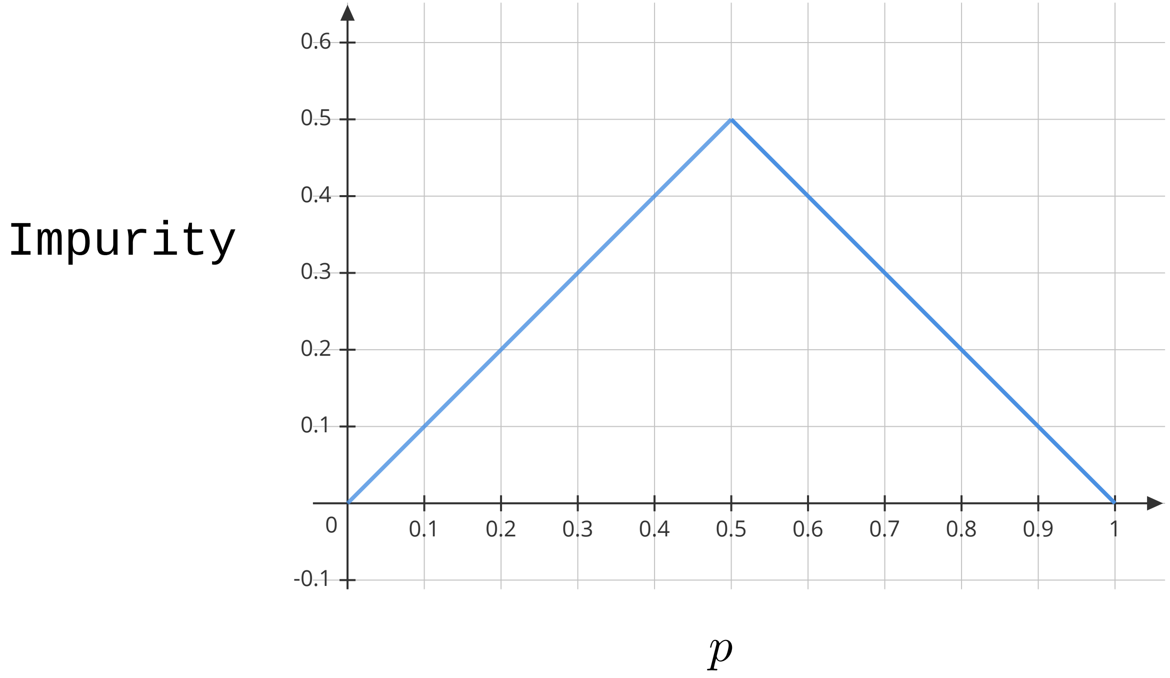
Node Impurity

Impurity measure

p	Impurity
Low	Low
Medium	High
High	Low

Node Impurity

Impurity measure



p	Impurity
Low	Low
Medium	High
High	Low

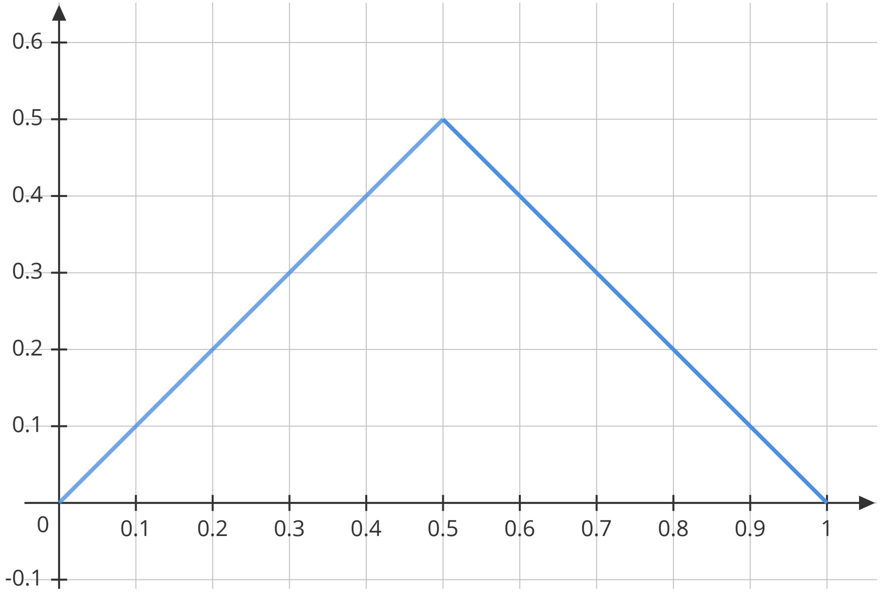
Impurity measure

Misclassification Error

N is a node with label l

$$\text{misclassification error (0-1 loss)} = \frac{\sum_{y_i \in N} \mathbf{1}[y_i \neq l]}{|N|}$$

Impurity

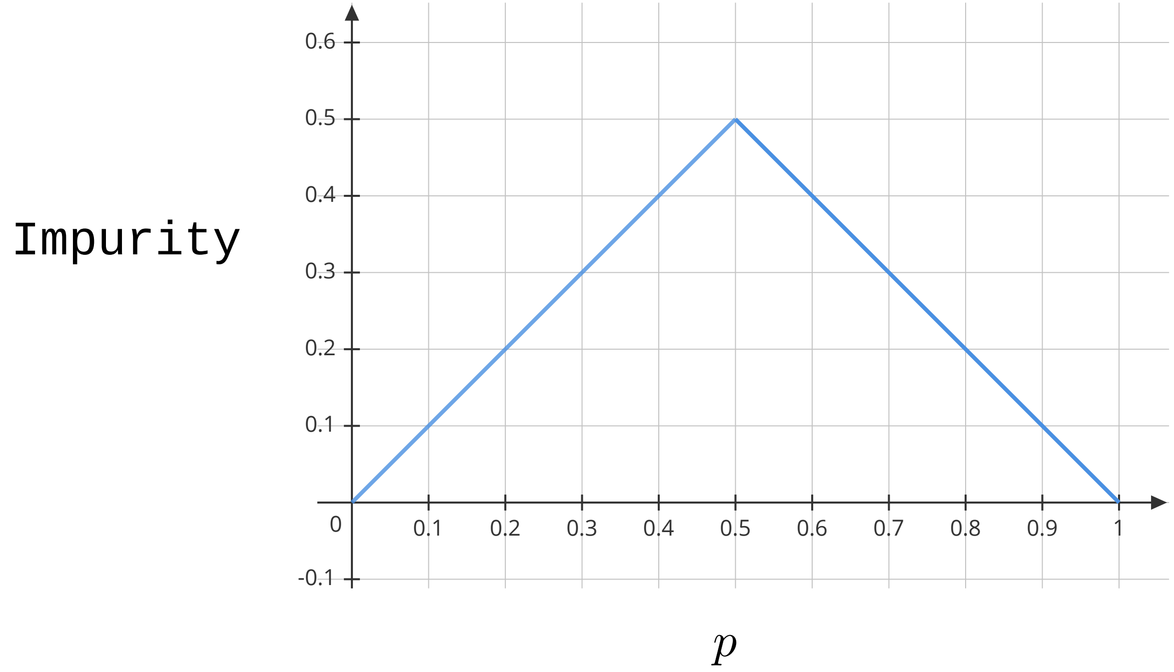


p

Impurity measure

Misclassification Error

N is a node
with label l



$$\text{misclassification error (0-1 loss)} = \frac{\sum_{y_i \in N} \mathbf{1}[y_i \neq l]}{|N|}$$

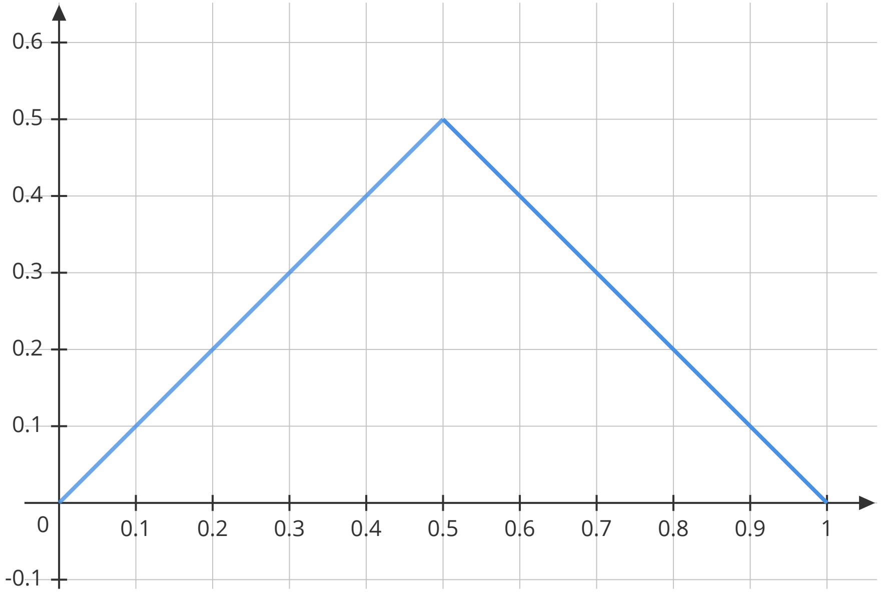
$$= \begin{cases} p, & 0 \leqslant p \leqslant 0.5 \\ 1 - p, & 0.5 < p \leqslant 1 \end{cases}$$

Impurity measure

Misclassification Error

N is a node with label l

Impurity



p

misclassification error (0-1 loss) $= \frac{\sum_{y_i \in N} \mathbf{1}[y_i \neq l]}{|N|}$

$$= \begin{cases} p, & 0 \leqslant p \leqslant 0.5 \\ 1 - p, & 0.5 < p \leqslant 1 \end{cases}$$

$$= 1 - \max(p, 1 - p)$$

Impurity measure

Misclassification Error

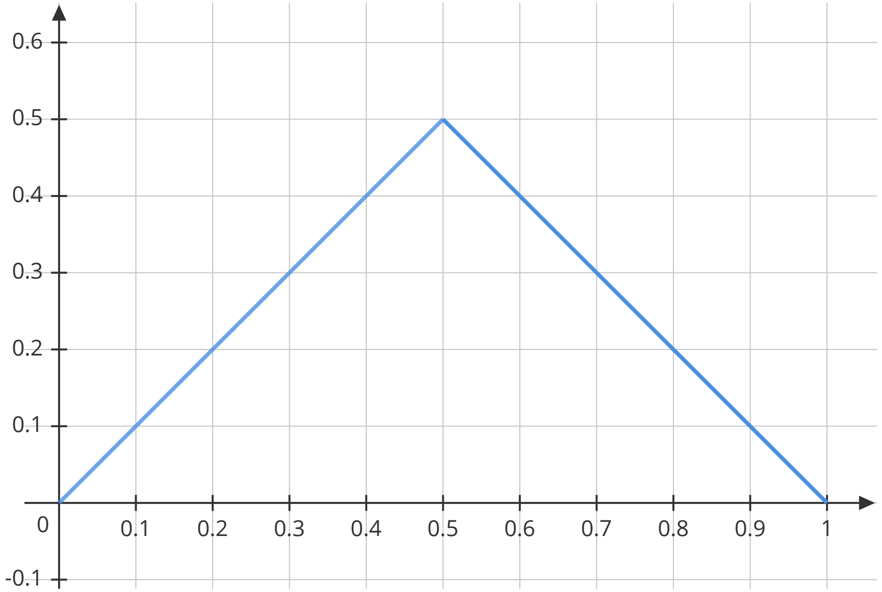
N is a node with label l

misclassification error (0-1 loss) $= \dfrac{\sum_{y_i \in N} \mathbf{1}[y_i \neq l]}{|N|}$

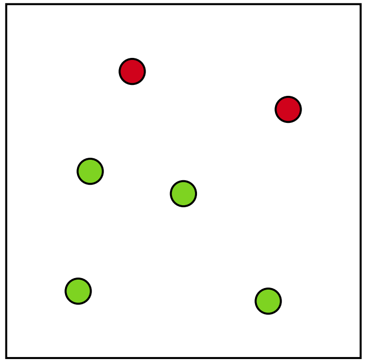
$$= \begin{cases} p, & 0 \leqslant p \leqslant 0.5 \\ 1 - p, & 0.5 < p \leqslant 1 \end{cases}$$

$$= 1 - \max(p, 1 - p)$$

Impurity



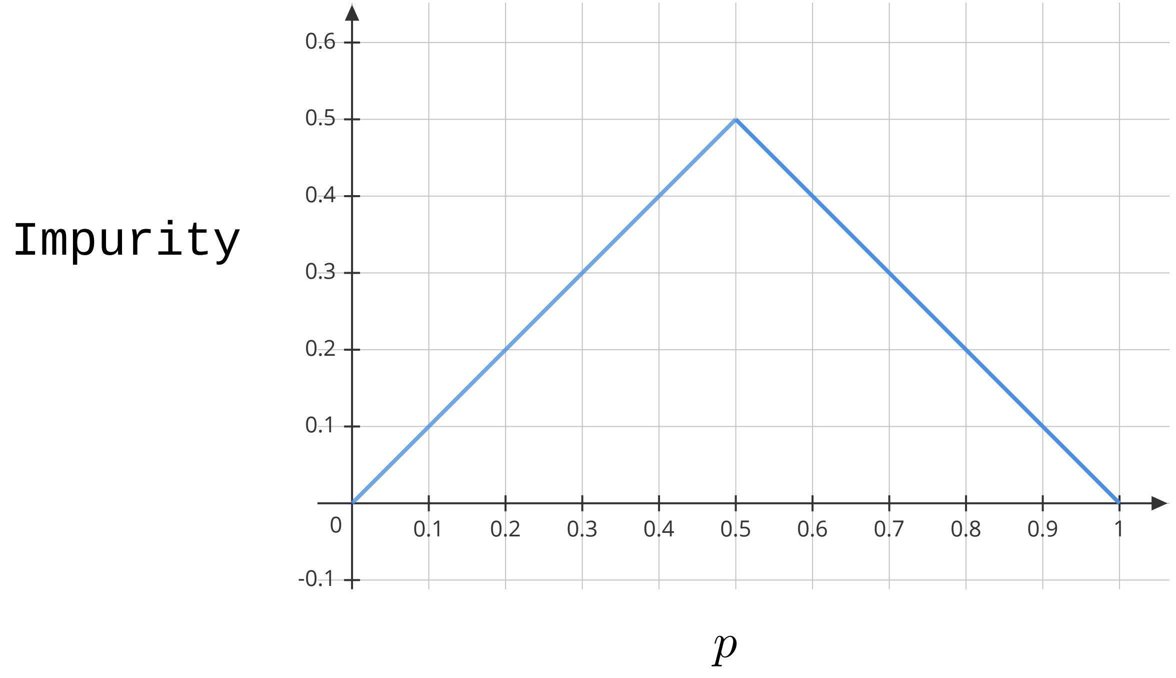
p



Impurity measure

Misclassification Error

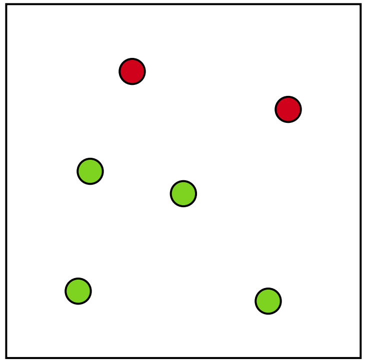
N is a node with label l



misclassification error (0-1 loss) $= \frac{\sum_{y_i \in N} \mathbf{1}[y_i \neq l]}{|N|}$

$$= \begin{cases} p, & 0 \leqslant p \leqslant 0.5 \\ 1-p, & 0.5 < p \leqslant 1 \end{cases}$$

$$= 1 - \max(p, 1-p)$$



$|N| = 6$

$p = \frac{2}{3}$

$l = 1$

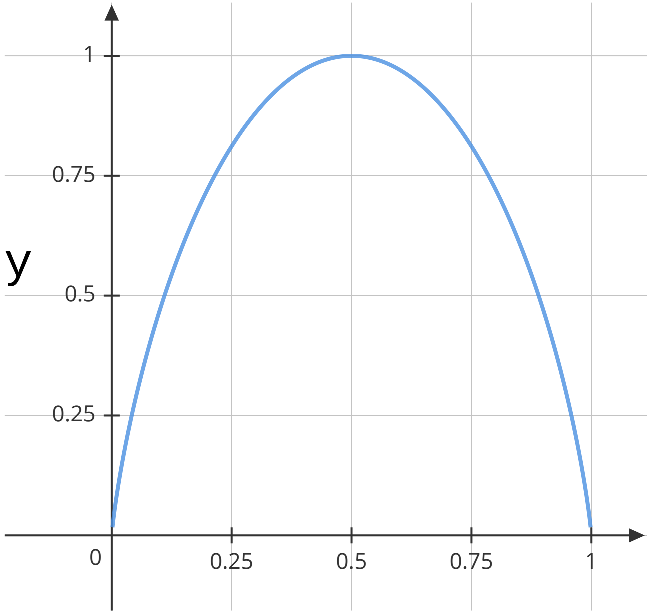
$\texttt{Imp}(N) = \frac{1}{3}$

Impurity measure

Entropy

$$\text{Entropy} = -p\log_2 p - (1-p)\log_2(1-p)$$

Impurity

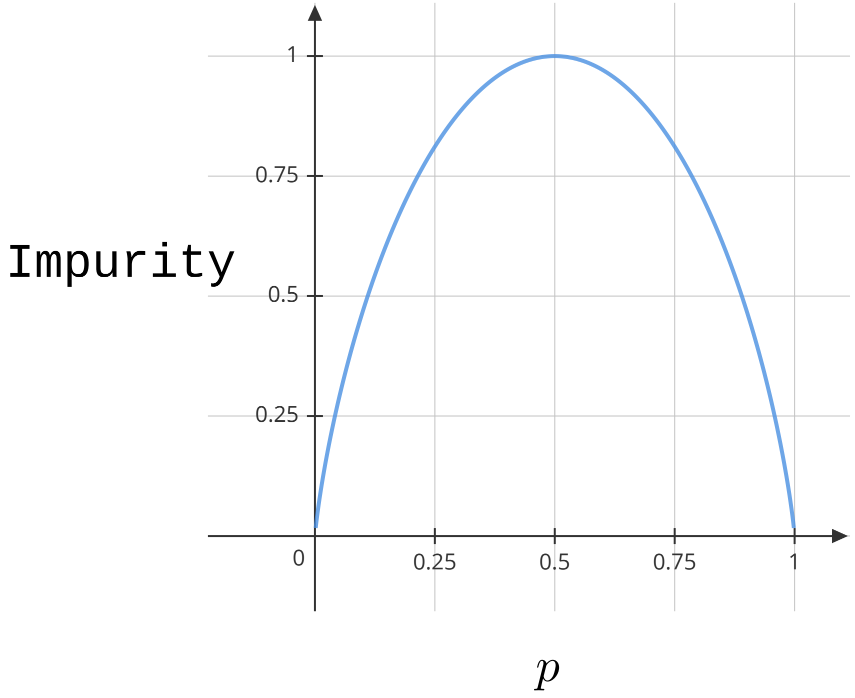


p

Impurity measure

Entropy

Entropy $= -p\log_2 p - (1-p)\log_2(1-p)$

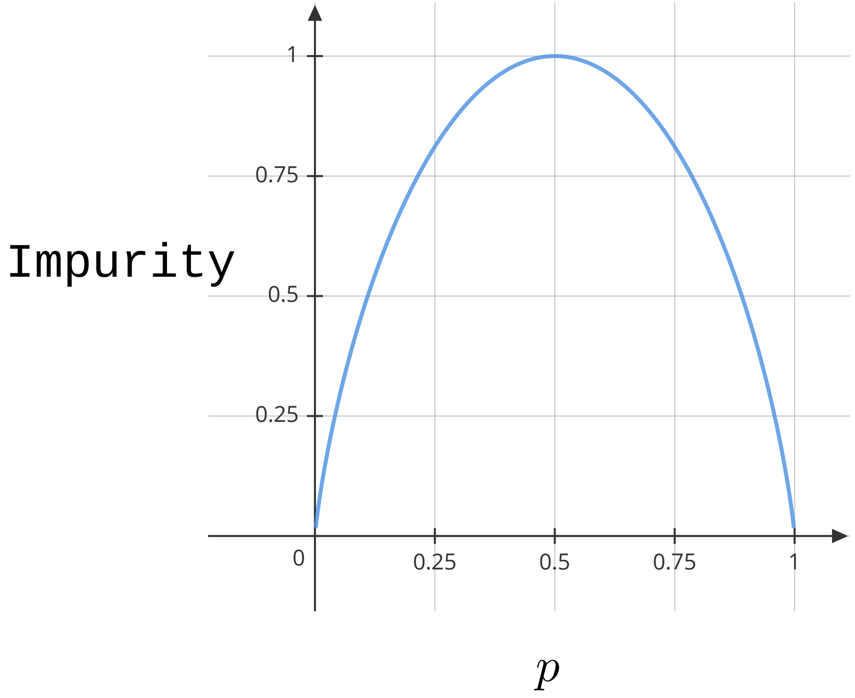


- Base is 2.
- We take $0\log_2 0 = 0$.
- Entropy is between 0 and 1.

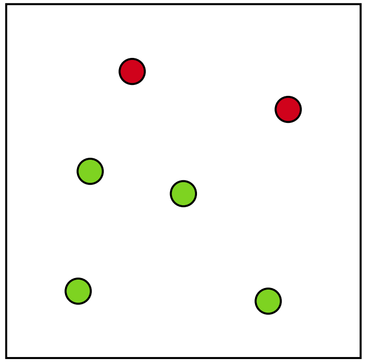
Impurity measure

Entropy

$$\text{Entropy} = -p\log_2 p - (1-p)\log_2(1-p)$$



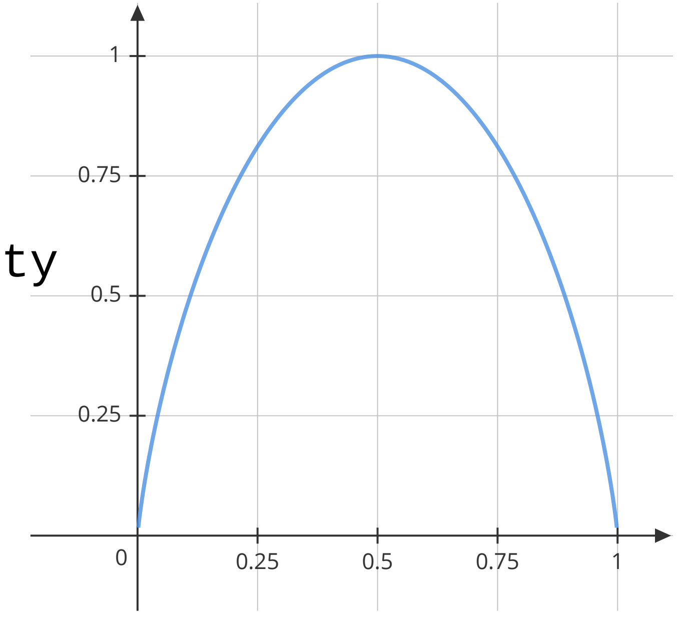
- Base is 2.
- We take $0\log_2 0 = 0$.
- Entropy is between 0 and 1.



Impurity measure

Entropy

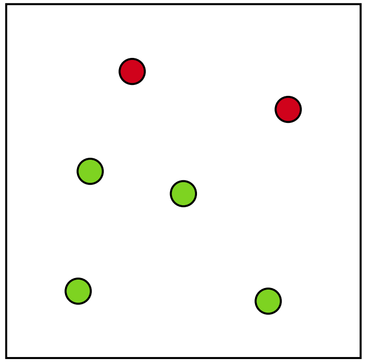
Impurity



p

Entropy $= -p\log_2 p - (1-p)\log_2(1-p)$

- Base is 2.
- We take $0\log_2 0 = 0$.
- Entropy is between 0 and 1.



$|N| = 6$

$p = \frac{2}{3}$

$l = 1$

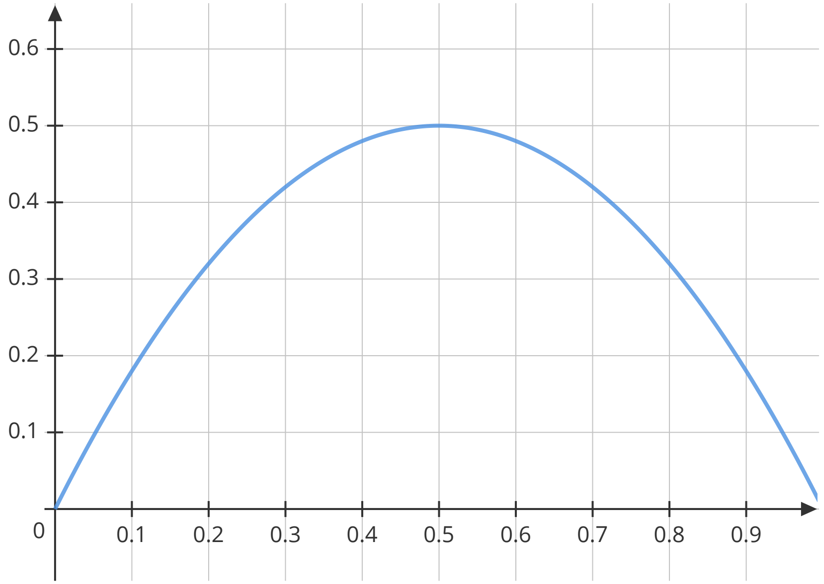
Imp$(N) = 0.92$

Impurity measure

Gini Index

Gini index $= 2p(1-p)$

Impurity

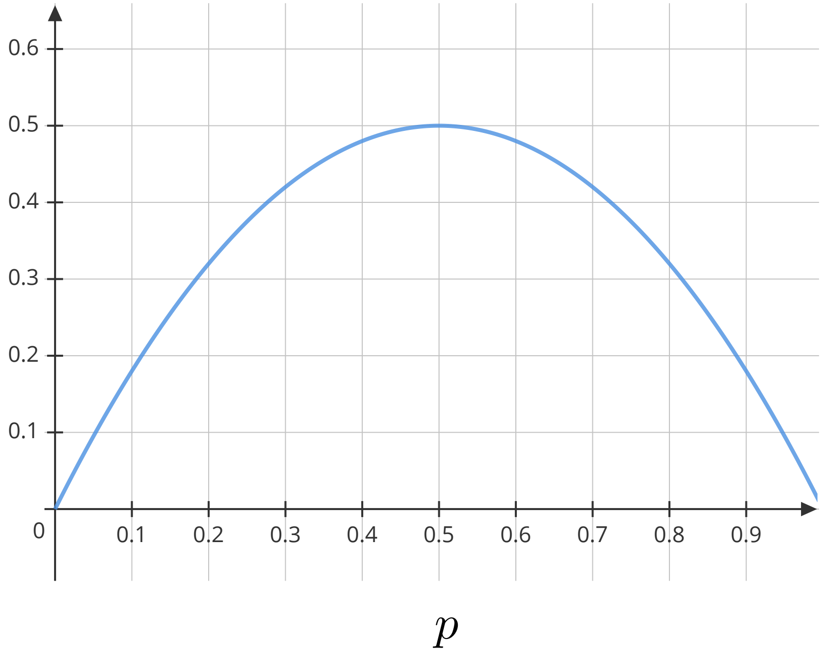


p

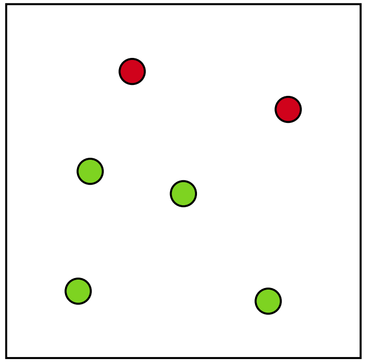
Impurity measure

Gini Index

Impurity



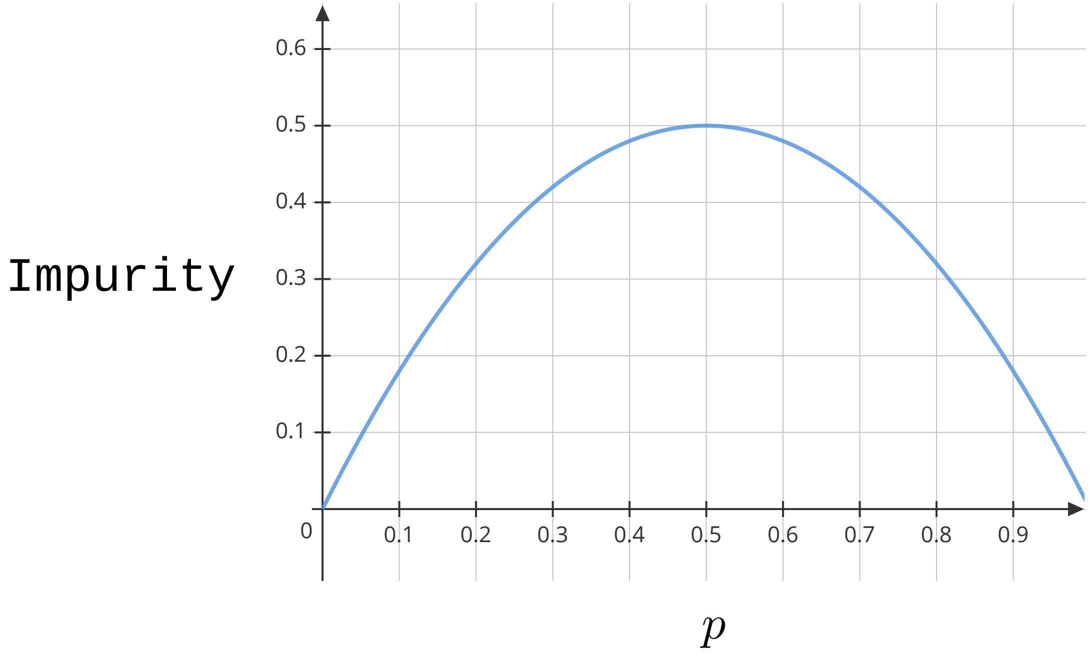
Gini index $= 2p(1-p)$

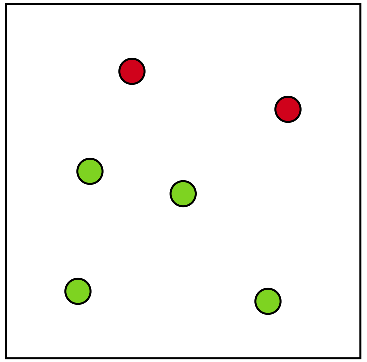


Impurity measure

Gini Index

$$\texttt{Gini index} = 2p(1-p)$$

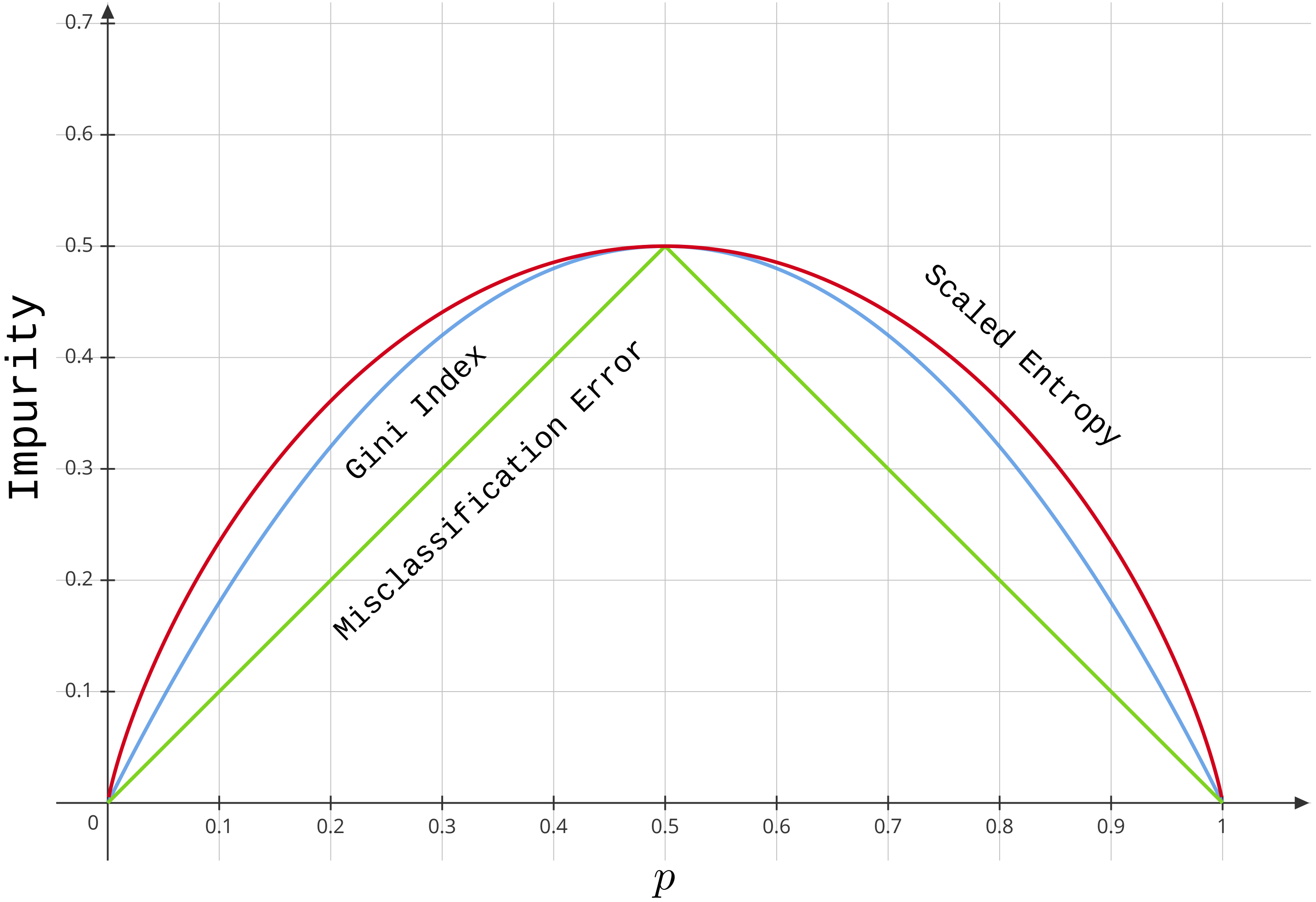




$$|N| = 6$$
$$p = \frac{2}{3}$$
$$l = 1$$
$$\texttt{Imp}(N) = 0.44$$

Impurity measures

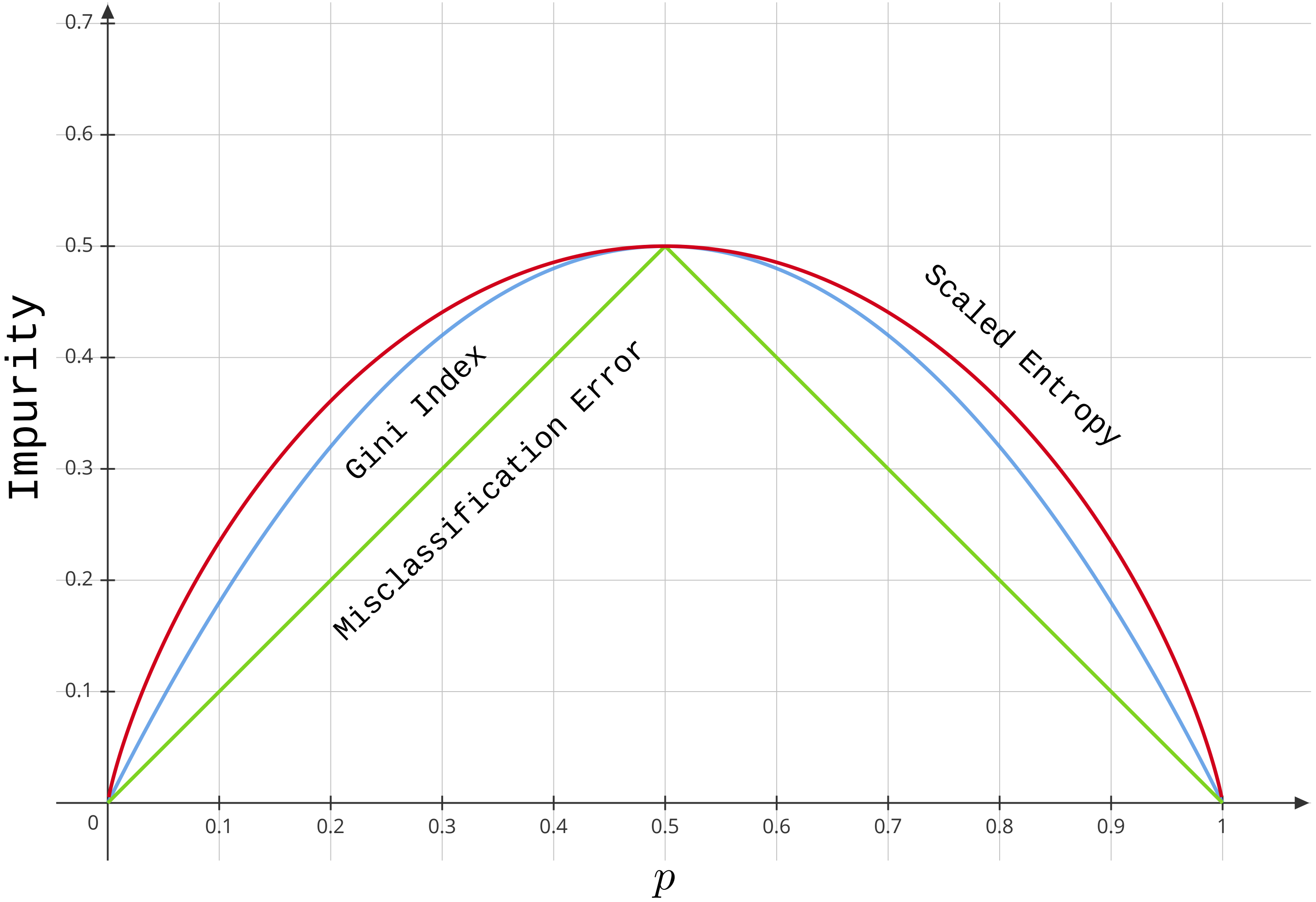
Comparison



Source: ESL (309)

Impurity measures

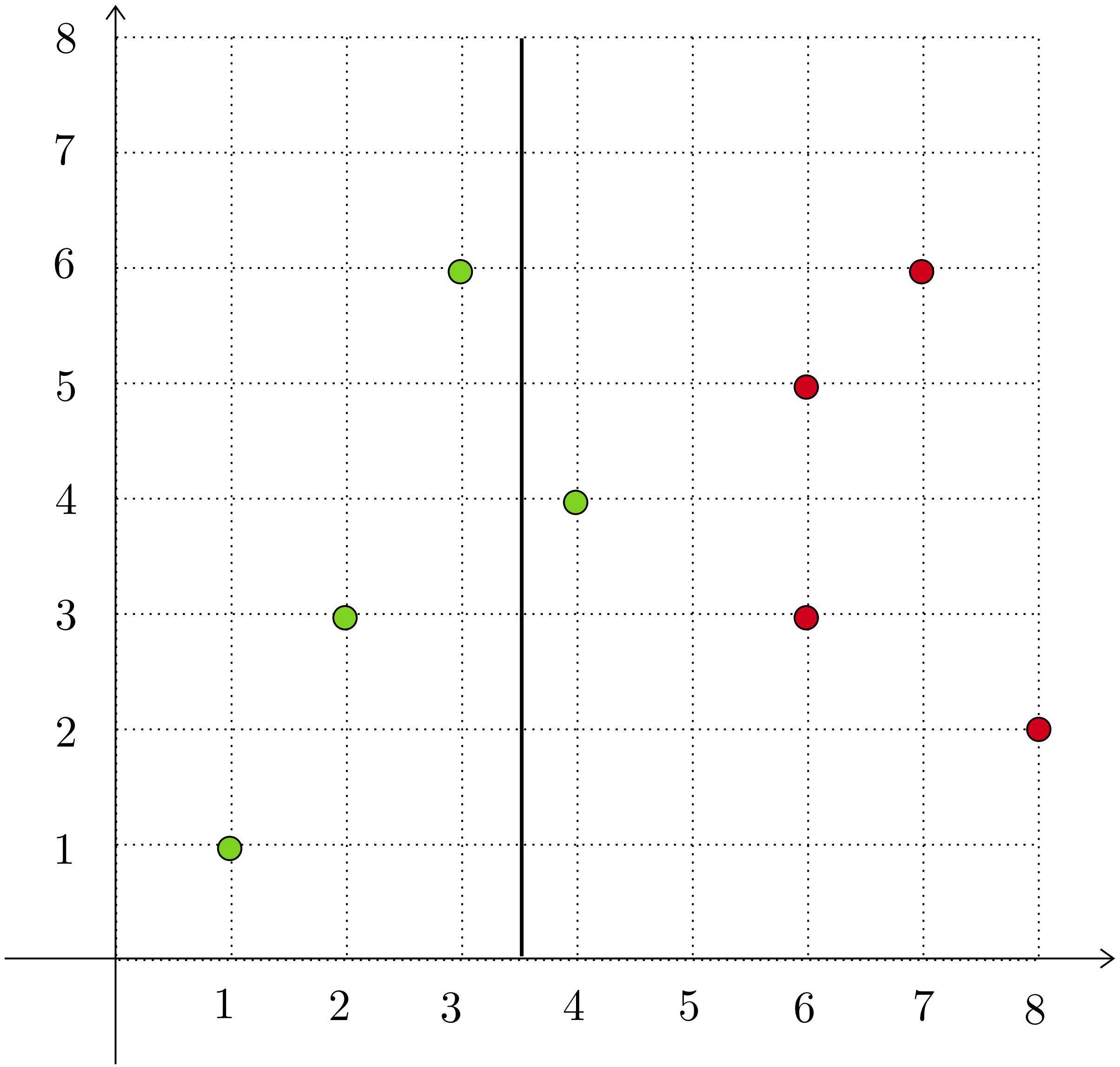
Comparison



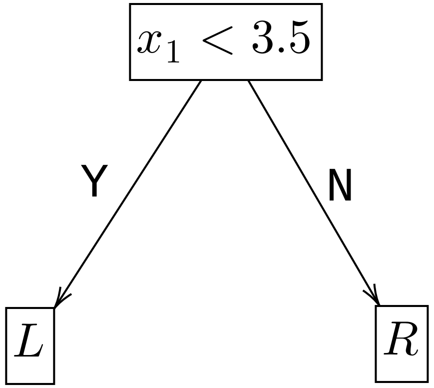
- Gini index and entropy are differentiable.
- Better suited to numerical optimization.
- Gini index and entropy are more sensitive to changes in node probabilities.
- Use Gini index or **entropy** for growing a tree.

Source: ESL (309)

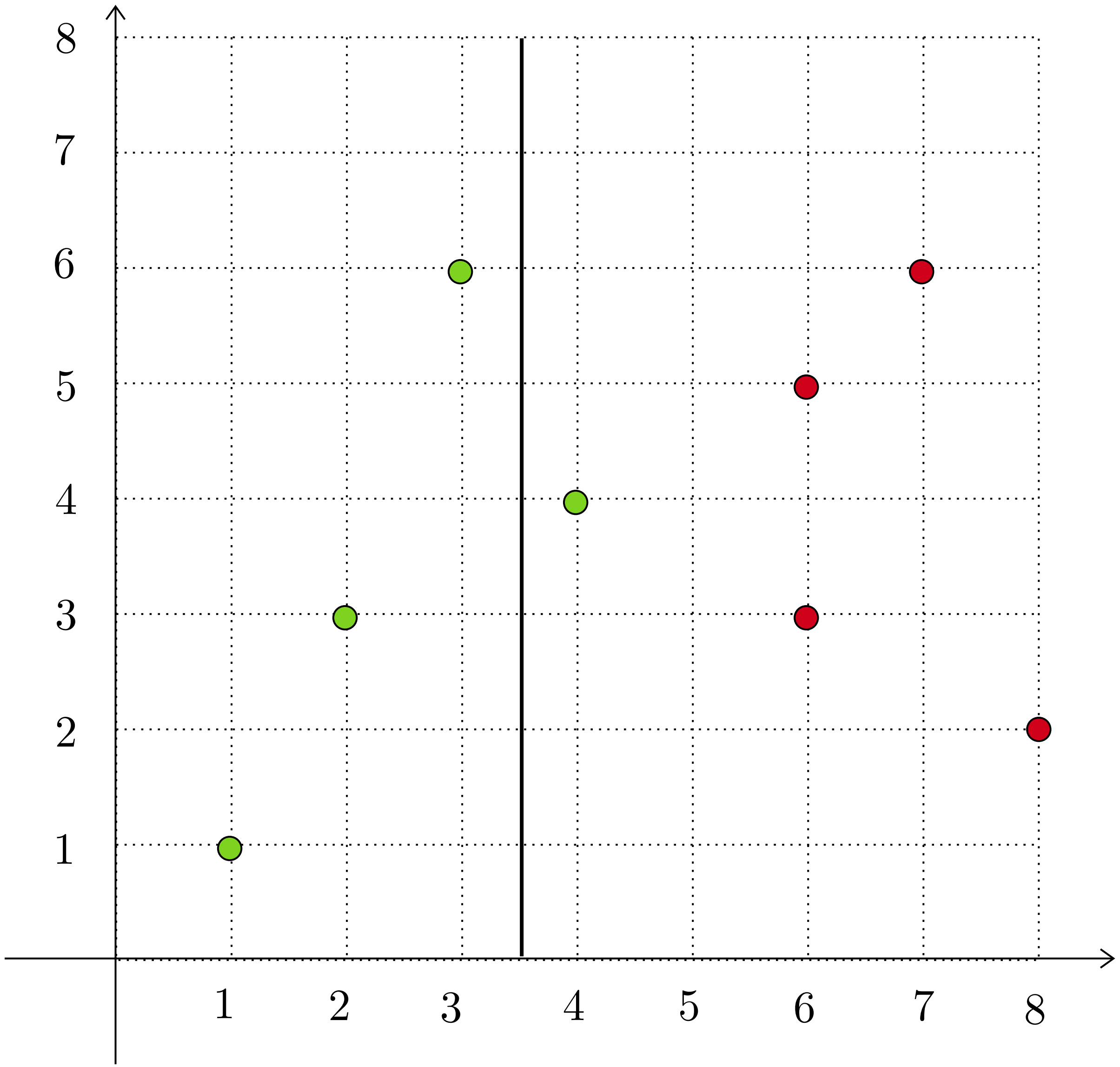
Growing a Tree

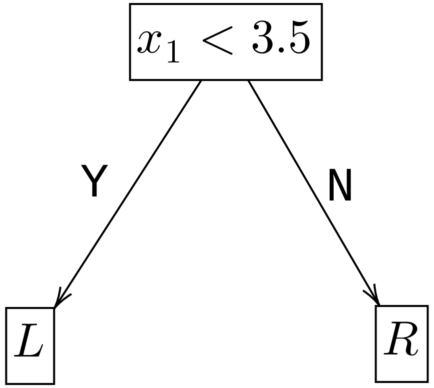


Entropy of children



Growing a Tree

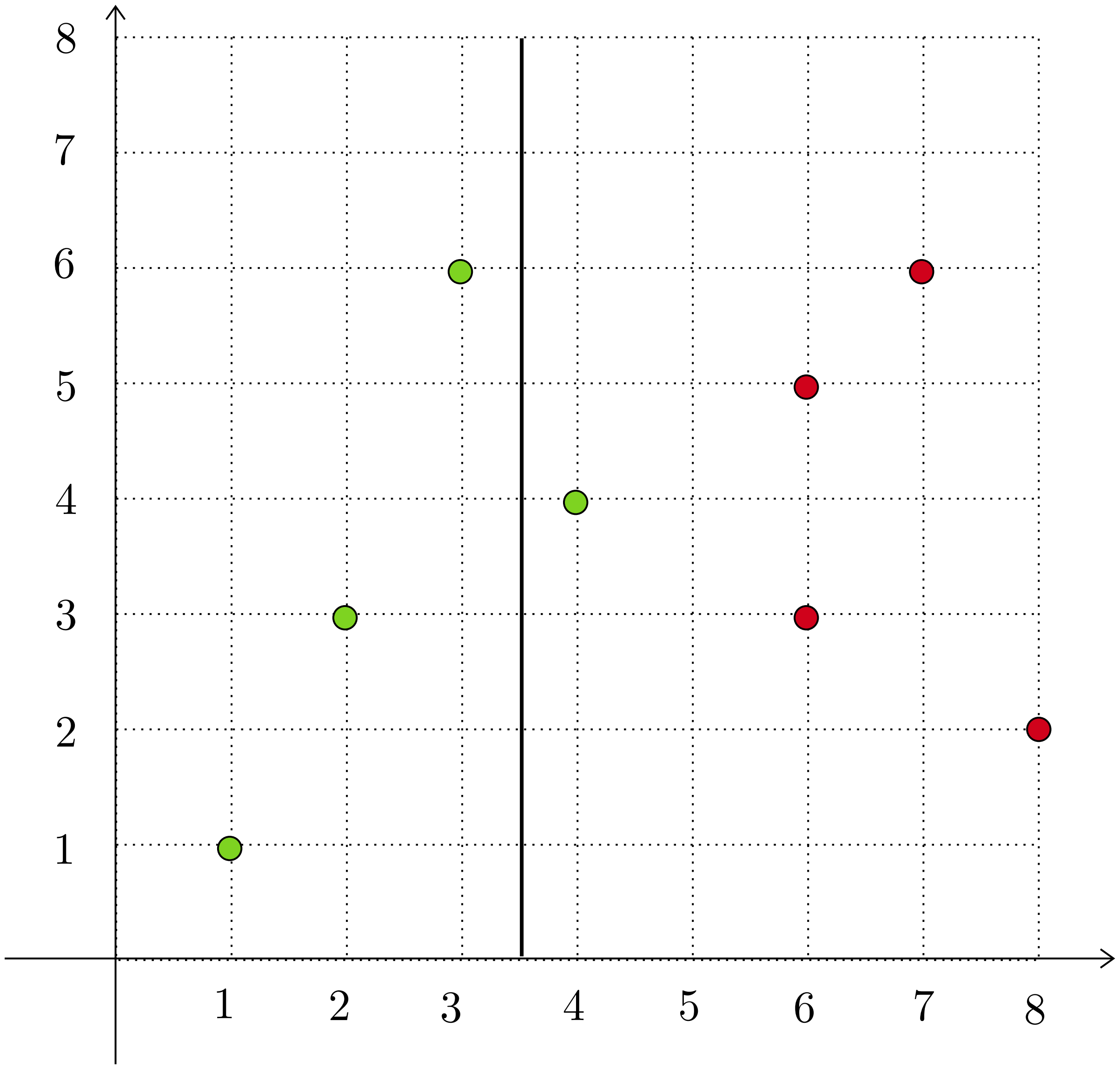


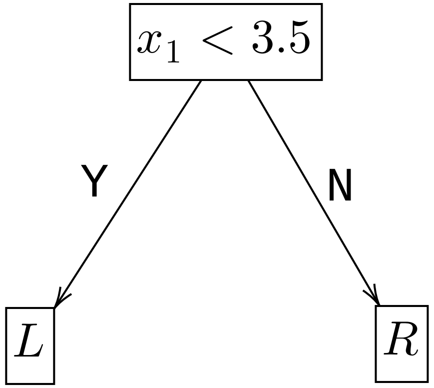


Entropy of children

$L = \{1, 1, 1\}$
$R = \{1, 0, 0, 0, 0\}$

Growing a Tree



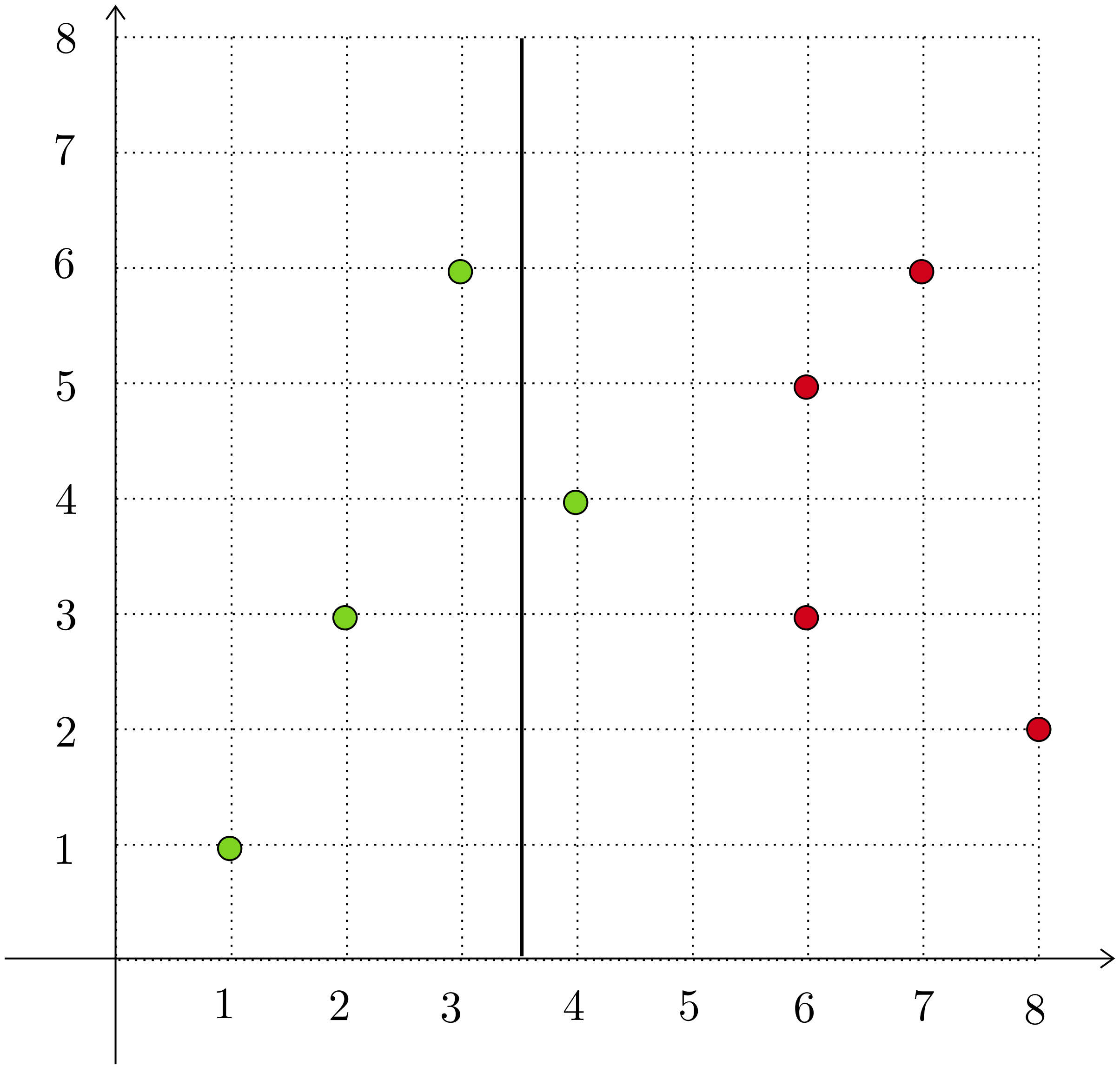


Entropy of children

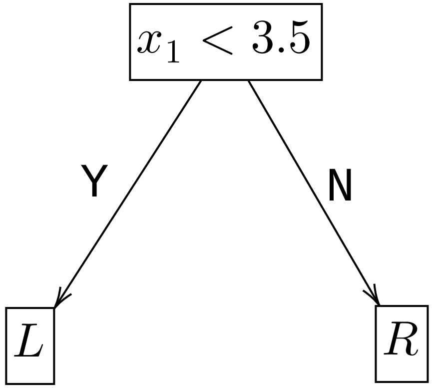
$L = \{1, 1, 1\}$
$R = \{1, 0, 0, 0, 0\}$

$p_L = 1,\ p_R = {}^1/_5$

Growing a Tree



Entropy of children

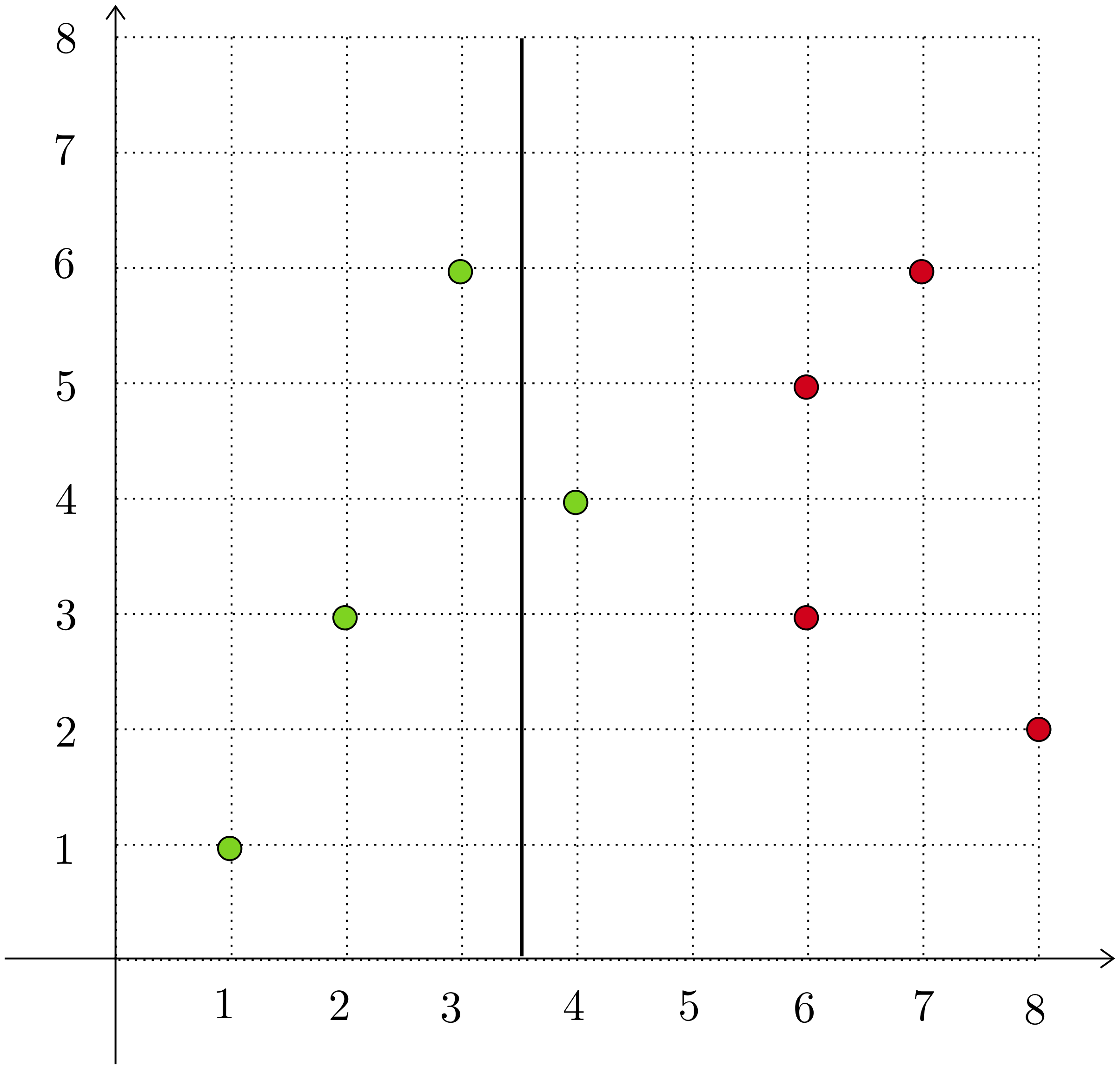


$L = \{1, 1, 1\}$
$R = \{1, 0, 0, 0, 0\}$

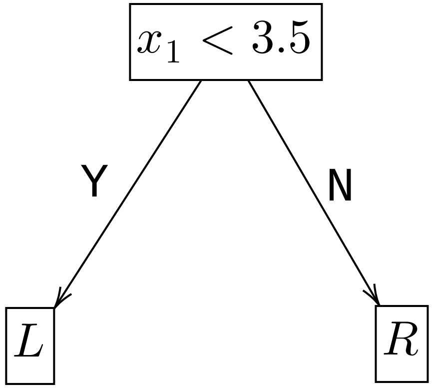
$p_L = 1, \; p_R = {}^{1}/_{5}$

$$E_{_L} = -p_{_L} \log p_{_L} - (1 - p_{_L}) \log (1 - p_{_L})$$
$$= 0$$

Growing a Tree



Entropy of children



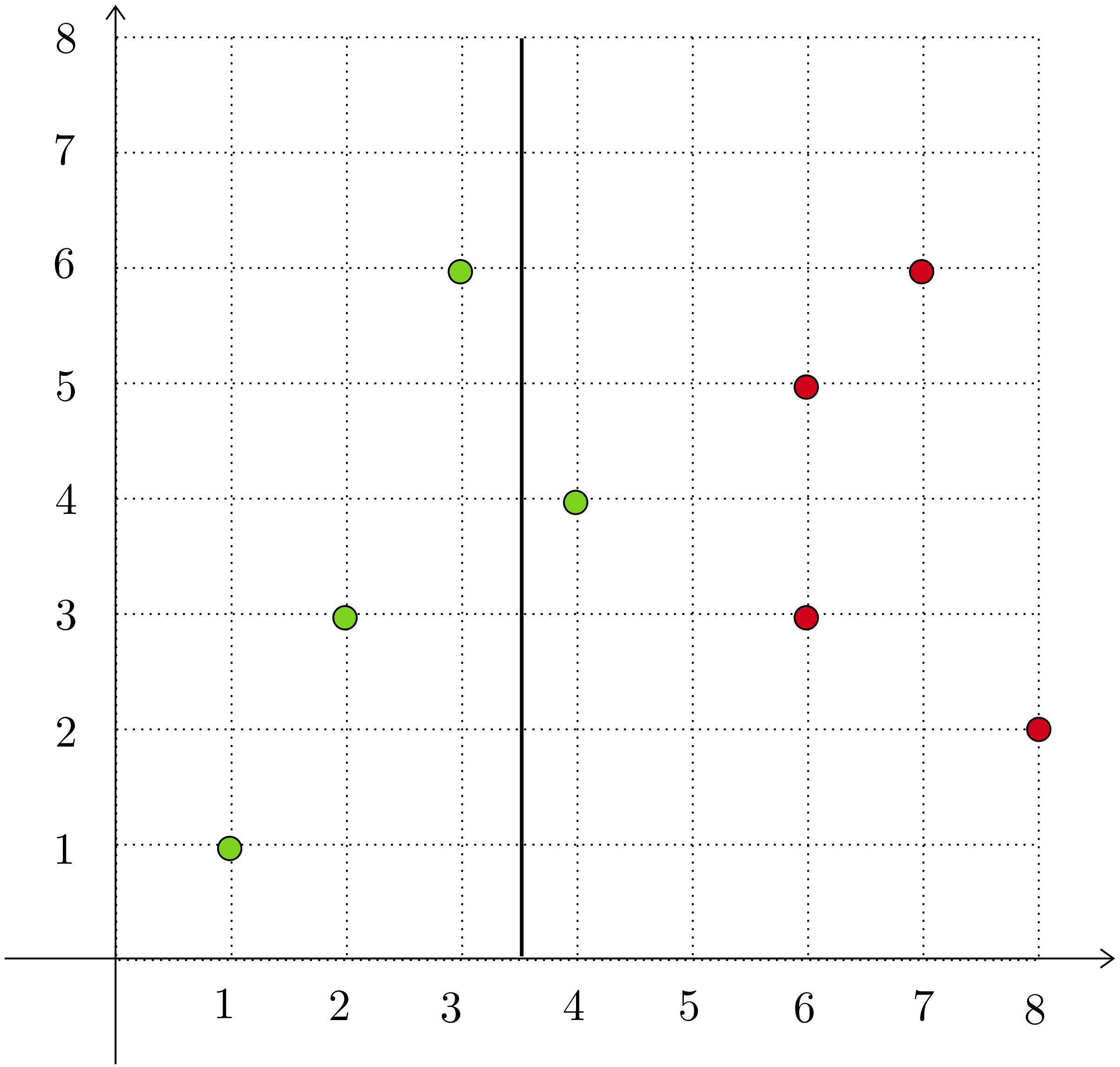
$$L = \{1, 1, 1\}$$
$$R = \{1, 0, 0, 0, 0\}$$

$$p_L = 1, \; p_R = {}^1\!/_5$$

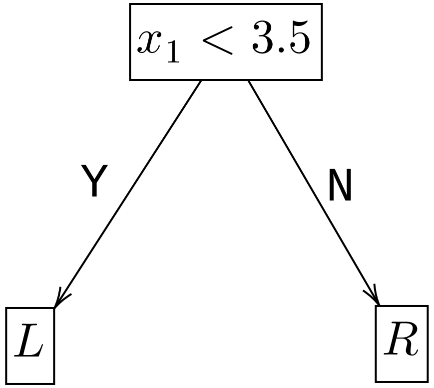
$$\begin{aligned} E_L &= -p_L \log p_L - (1 - p_L)\log(1 - p_L) \\ &= 0 \end{aligned}$$

$$\begin{aligned} E_R &= -p_R \log p_R - (1 - p_R)\log(1 - p_R) \\ &= 0.722 \end{aligned}$$

Growing a Tree



Entropy of children



$L = \{1, 1, 1\}$
$R = \{1, 0, 0, 0, 0\}$

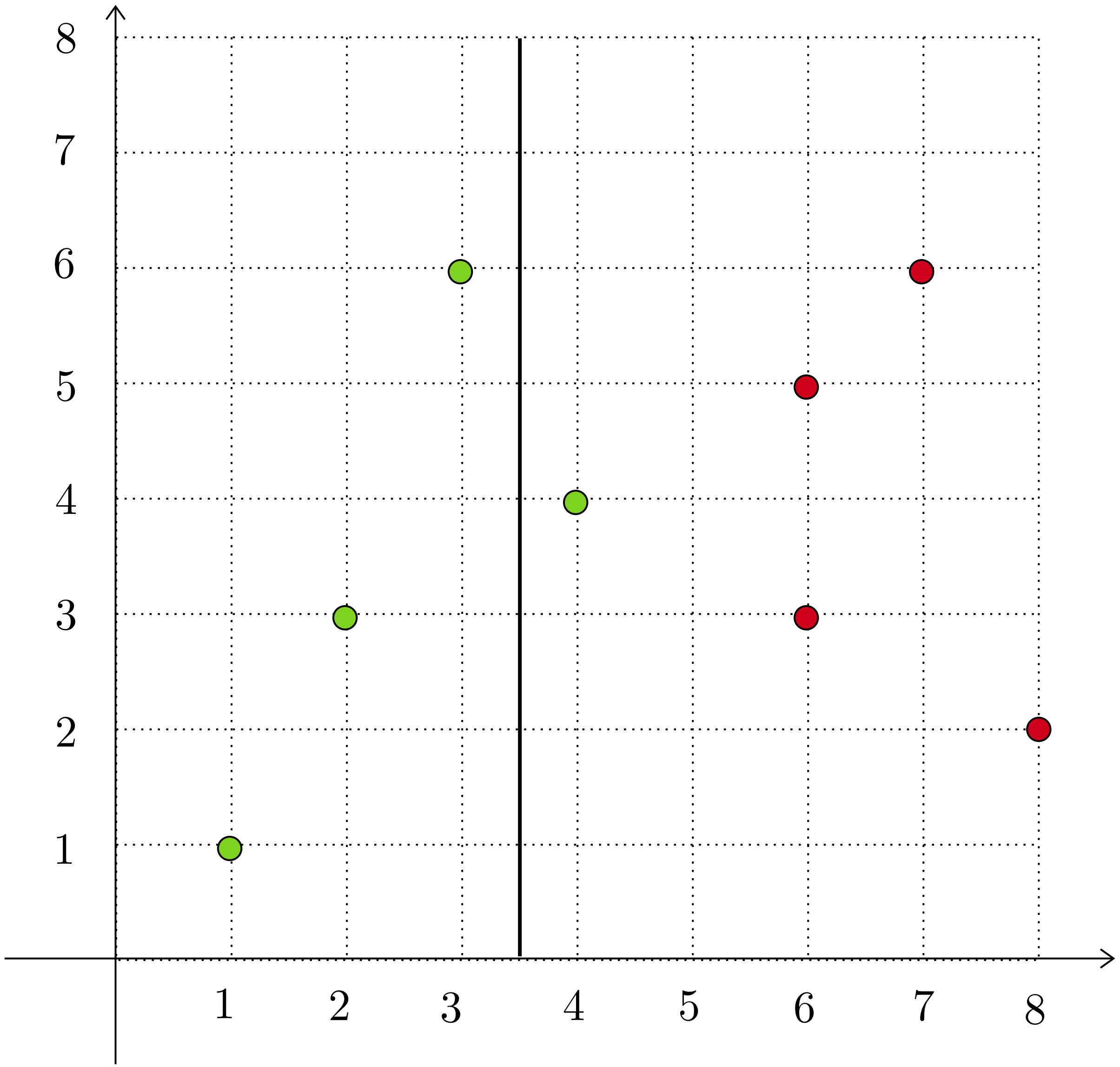
$p_L = 1, \; p_R = {}^1/_5$

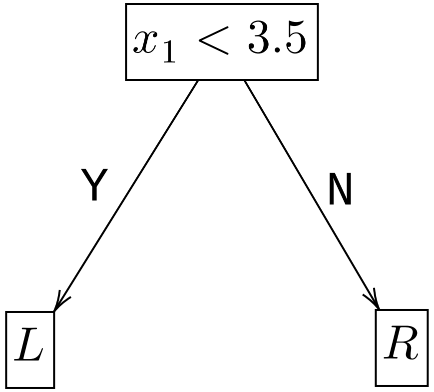
$$E_{L} = -p_{L} \log p_{L} - (1 - p_{L}) \log(1 - p_{L})$$
$$= 0$$

$$E_{R} = -p_{R} \log p_{R} - (1 - p_{R}) \log(1 - p_{R})$$
$$= 0.722$$

$$E_{LR} = \frac{3}{8} E_{L} + \frac{5}{8} E_{R}$$
$$= 0.451$$

Growing a Tree



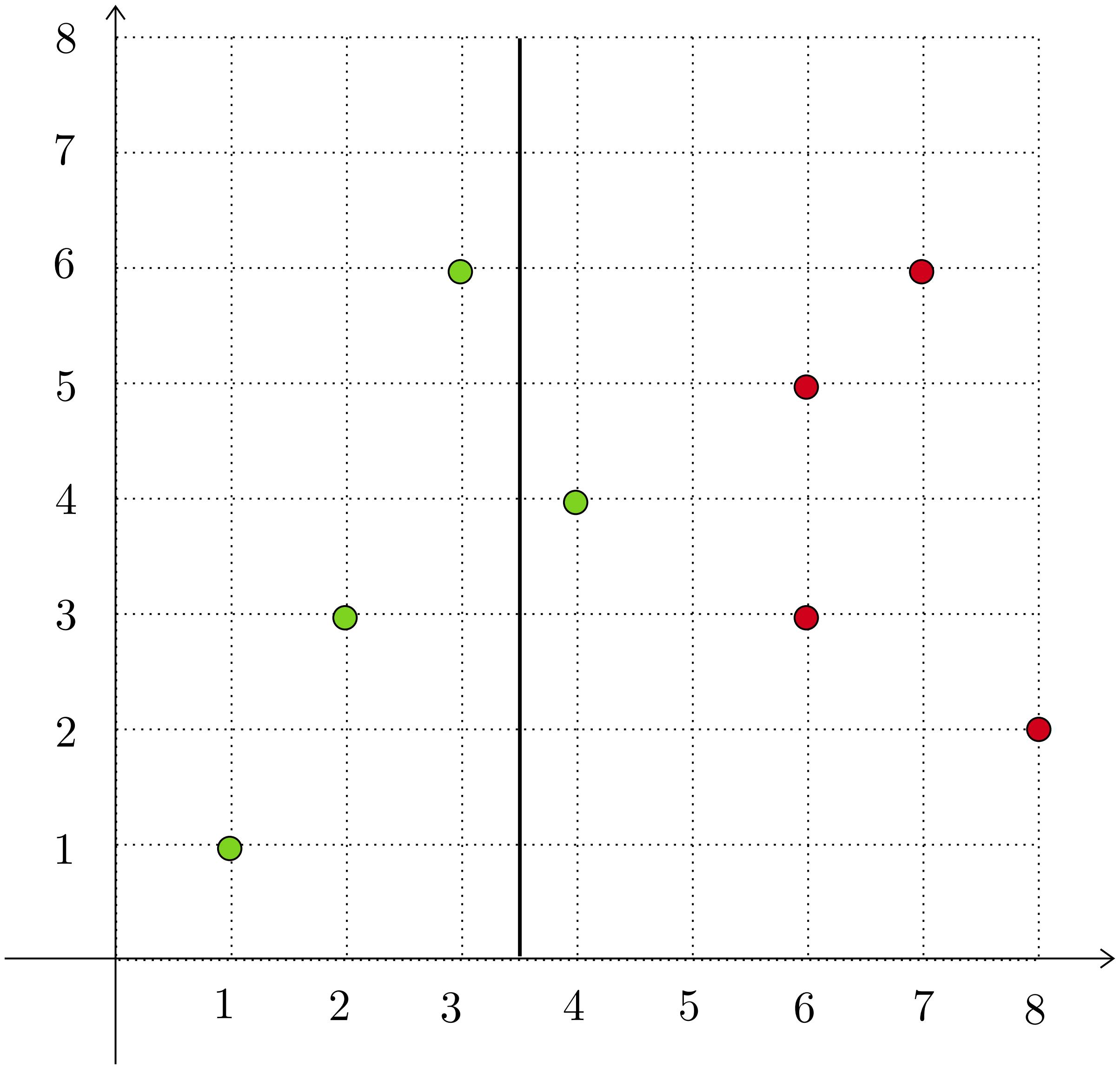


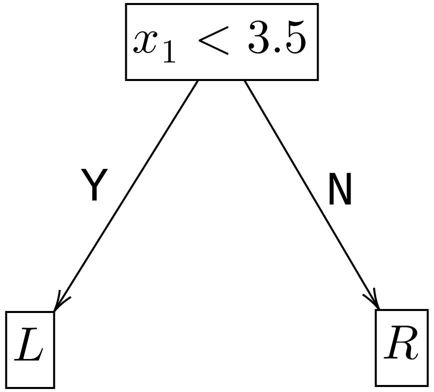
Information Gain

$$P = \{1,1,1,1,0,0,0,0\}$$
$$L = \{1,1,1\}$$
$$R = \{1,0,0,0,0\}$$

$$p_{_P} = {}^1\!/_2, p_{_L} = 1,\ p_{_R} = {}^1\!/_5$$

Growing a Tree





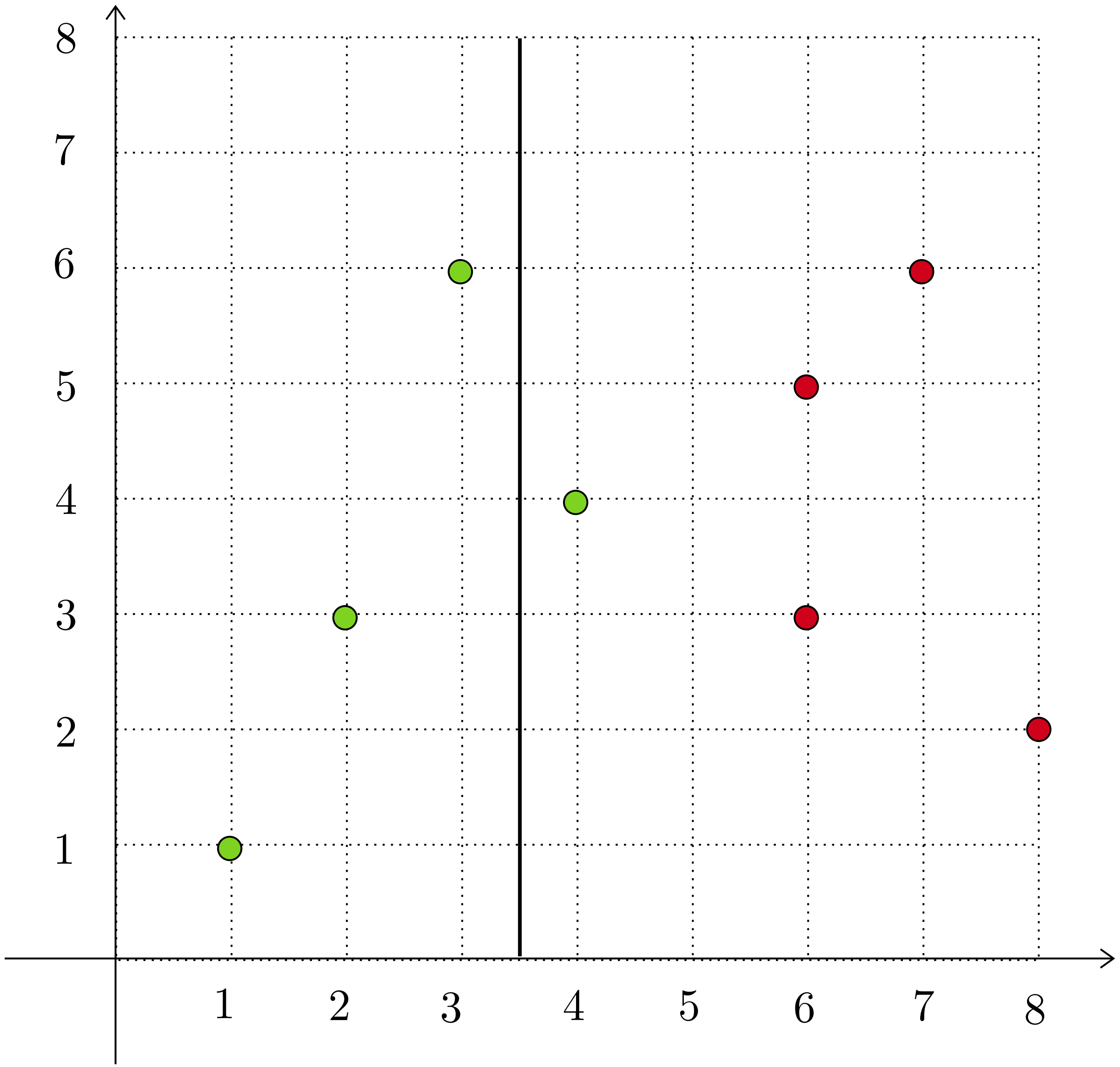
Information Gain

$$P = \{1,1,1,1,0,0,0,0\}$$
$$L = \{1,1,1\}$$
$$R = \{1,0,0,0,0\}$$

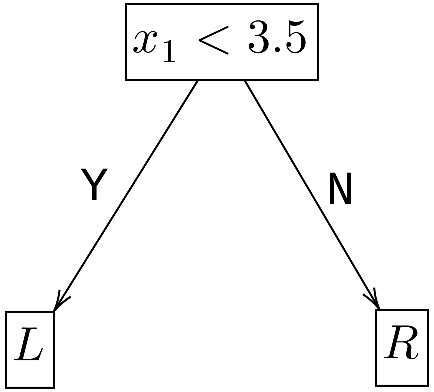
$$p_{_P} = {}^1\!/_2, p_{_L} = 1,\ p_{_R} = {}^1\!/_5$$

$$\begin{aligned} E_{_P} &= -p_{_P}\log p_{_P} - (1-p_{_P})\log(1-p_{_P}) \\ &= 1 \end{aligned}$$

Growing a Tree



Information Gain



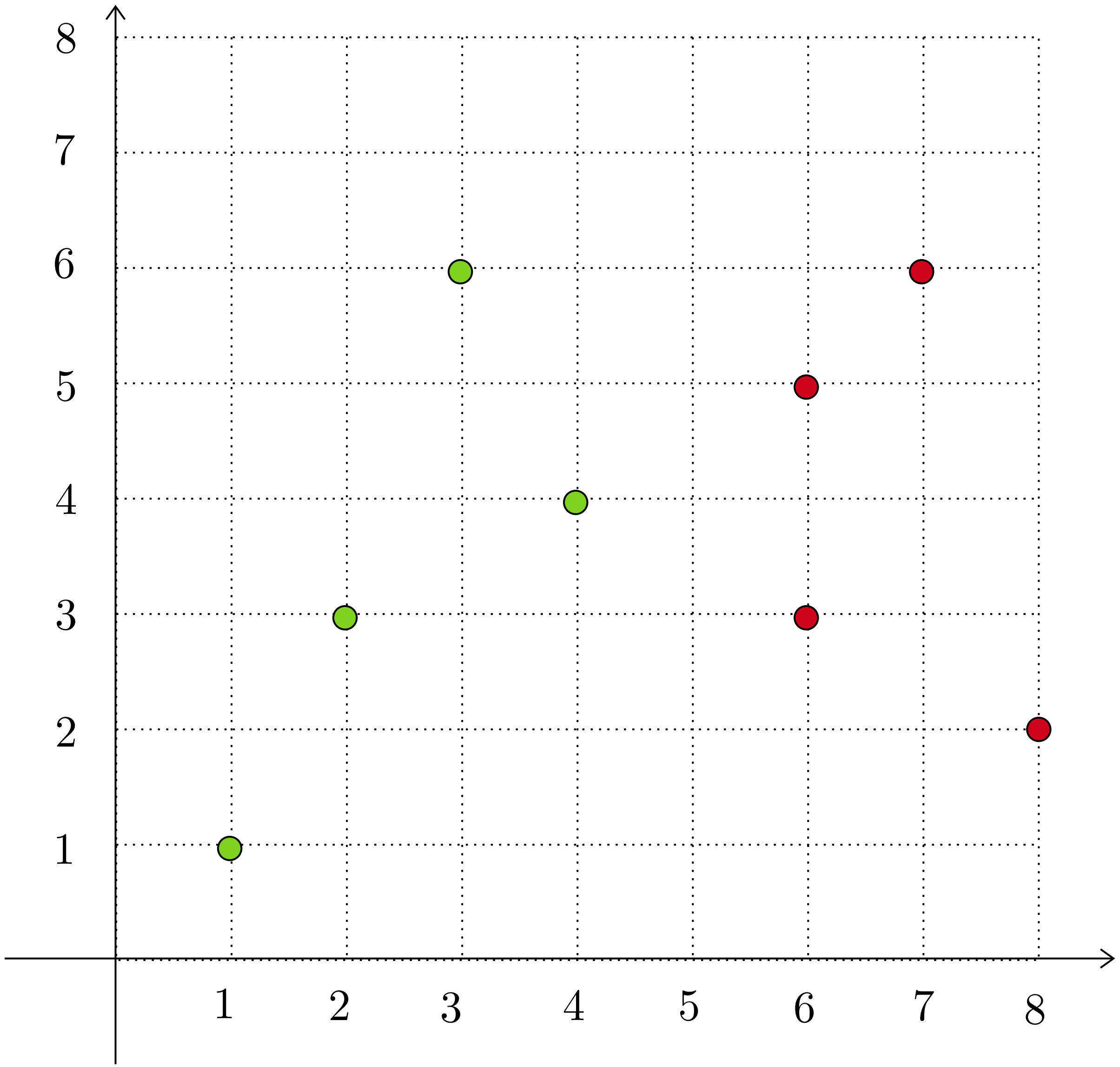
$$P = \{1,1,1,1,0,0,0,0\}$$
$$L = \{1,1,1\}$$
$$R = \{1,0,0,0,0\}$$

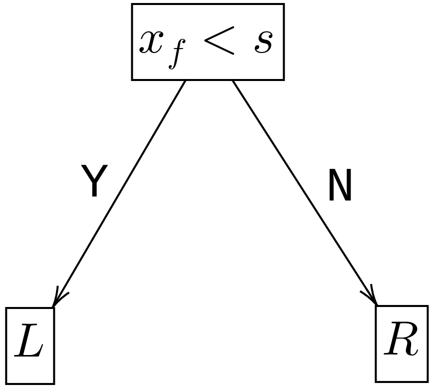
$$p_{_P} = {}^1\!/_2, p_{_L} = 1,\ p_{_R} = {}^1\!/_5$$

$$\begin{aligned} E_{_P} &= -p_{_P} \log p_{_P} - (1-p_{_P})\log(1-p_{_P}) \\ &= 1 \end{aligned}$$

$$\begin{aligned} IG &= E_{_P} - E_{_{LR}} \\ &= 1 - 0.451 \\ &= 0.549 \end{aligned}$$

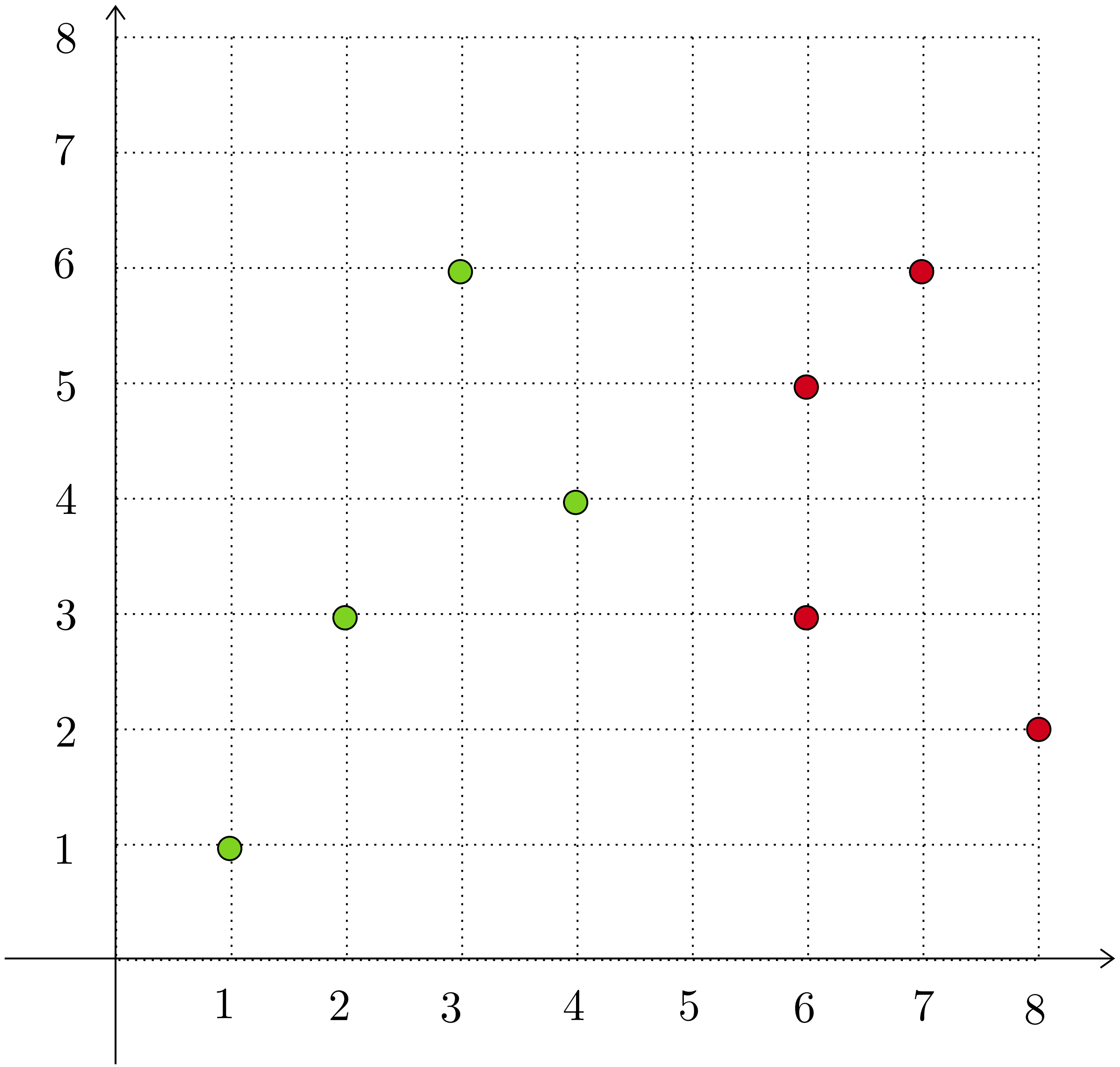
Growing a Tree



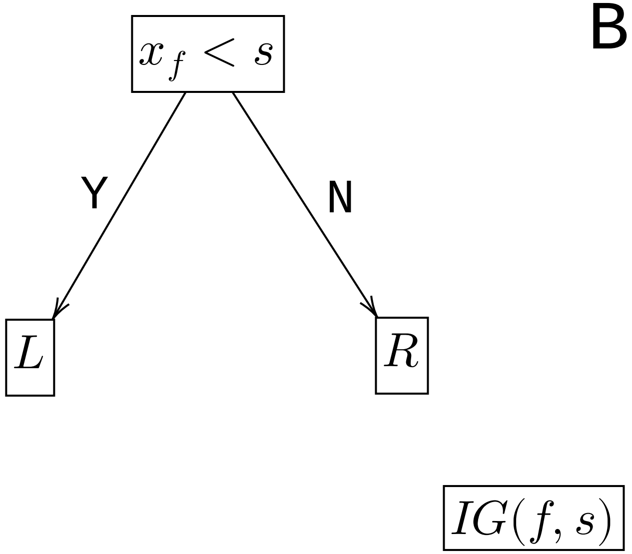


Best question

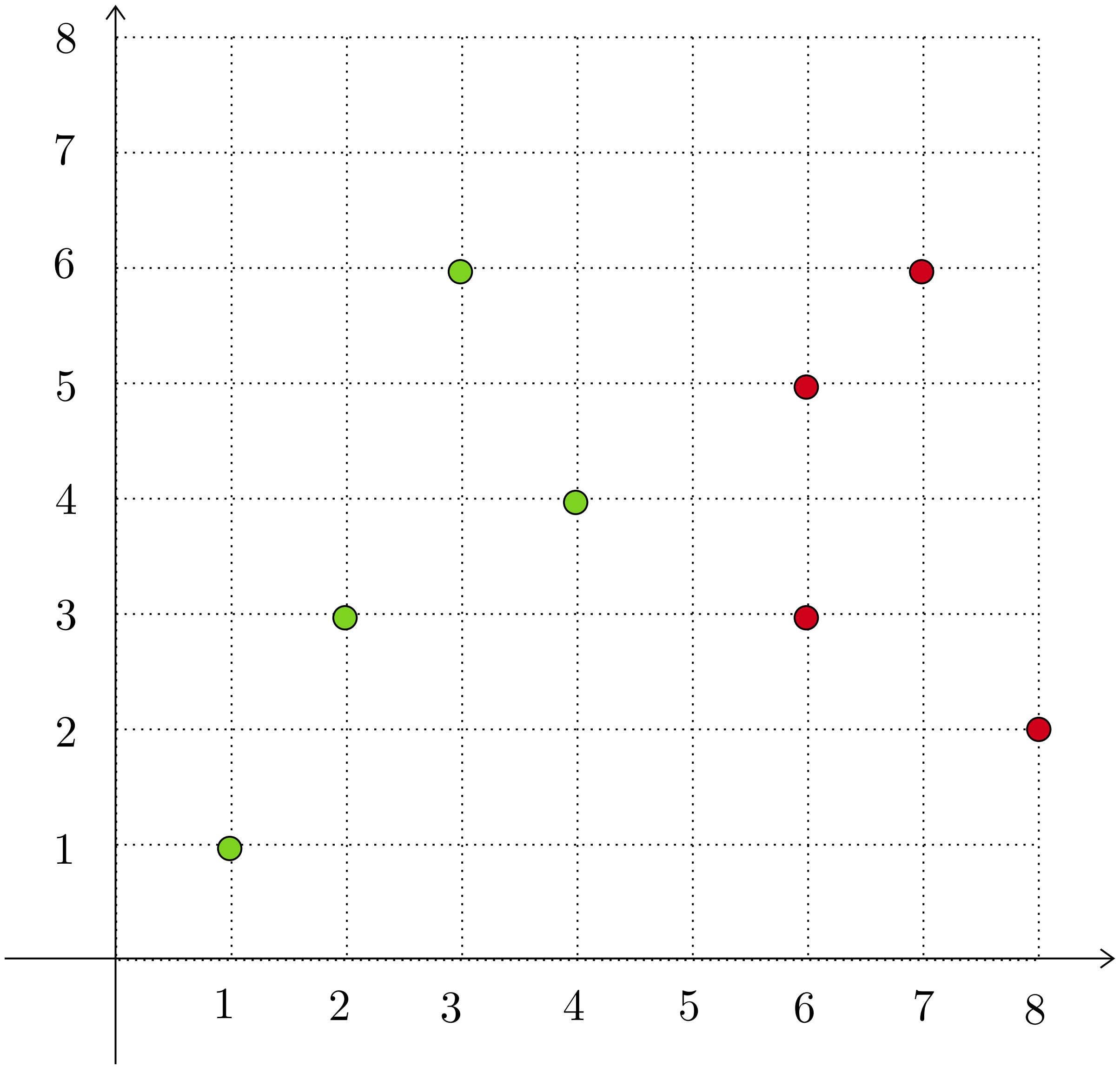
Growing a Tree



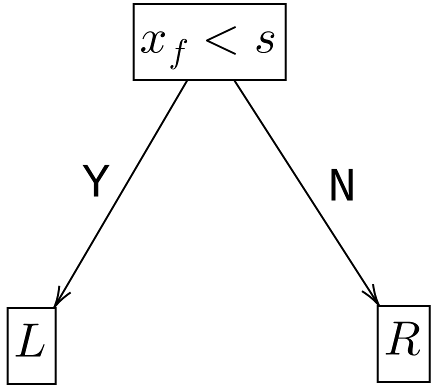
Best question



Growing a Tree



Best question

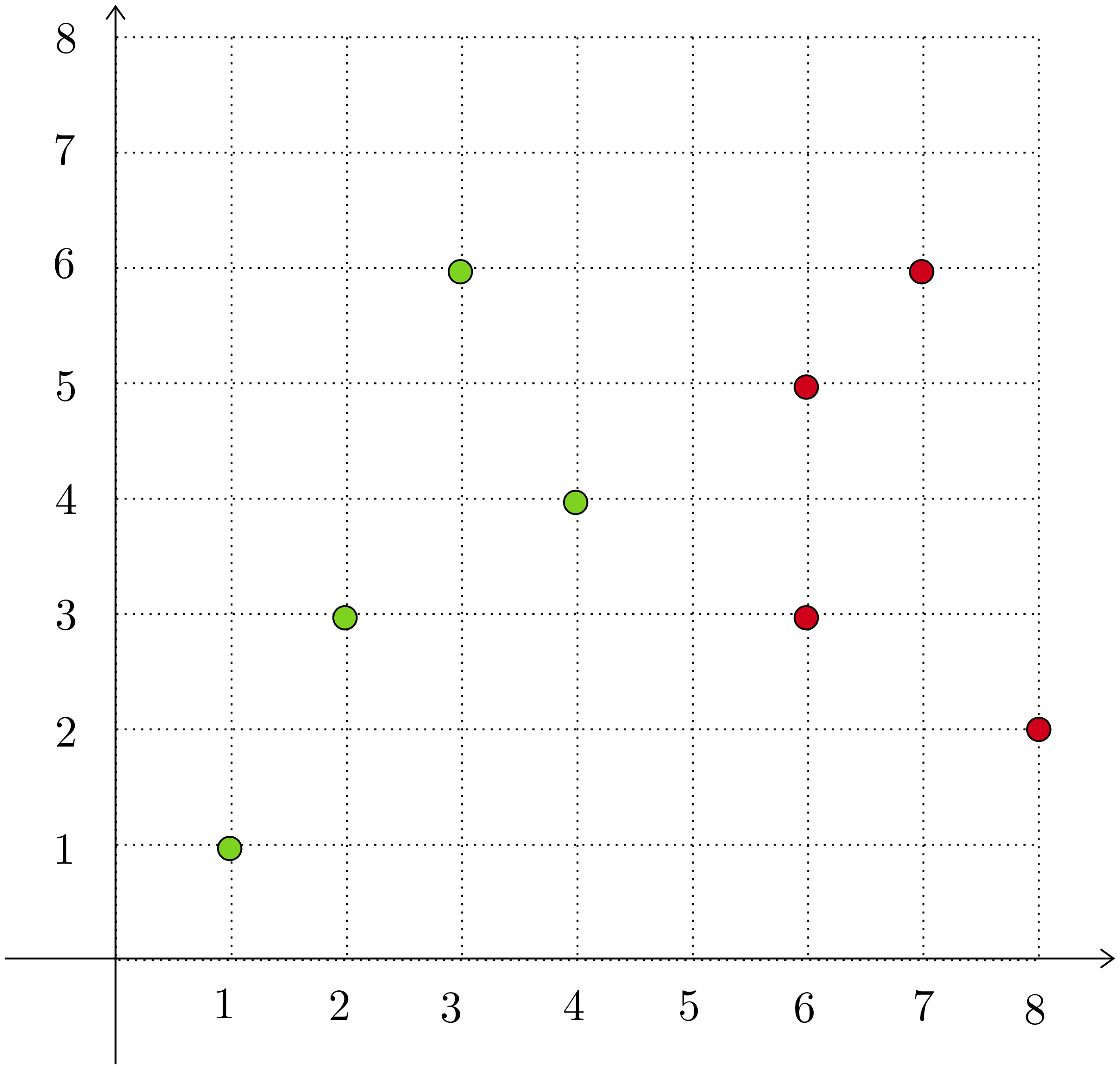


$IG(f, s)$

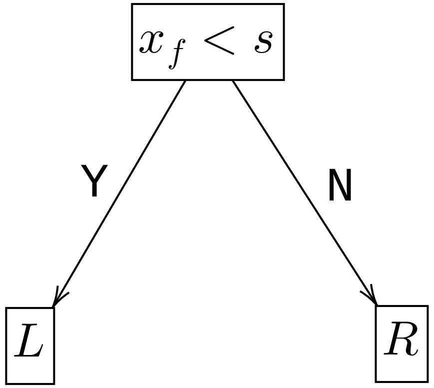
s	$IG(1, s)$

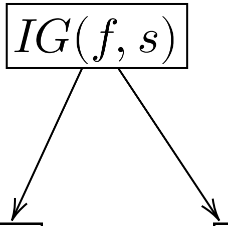
s	$IG(2, s)$

Growing a Tree



Best question

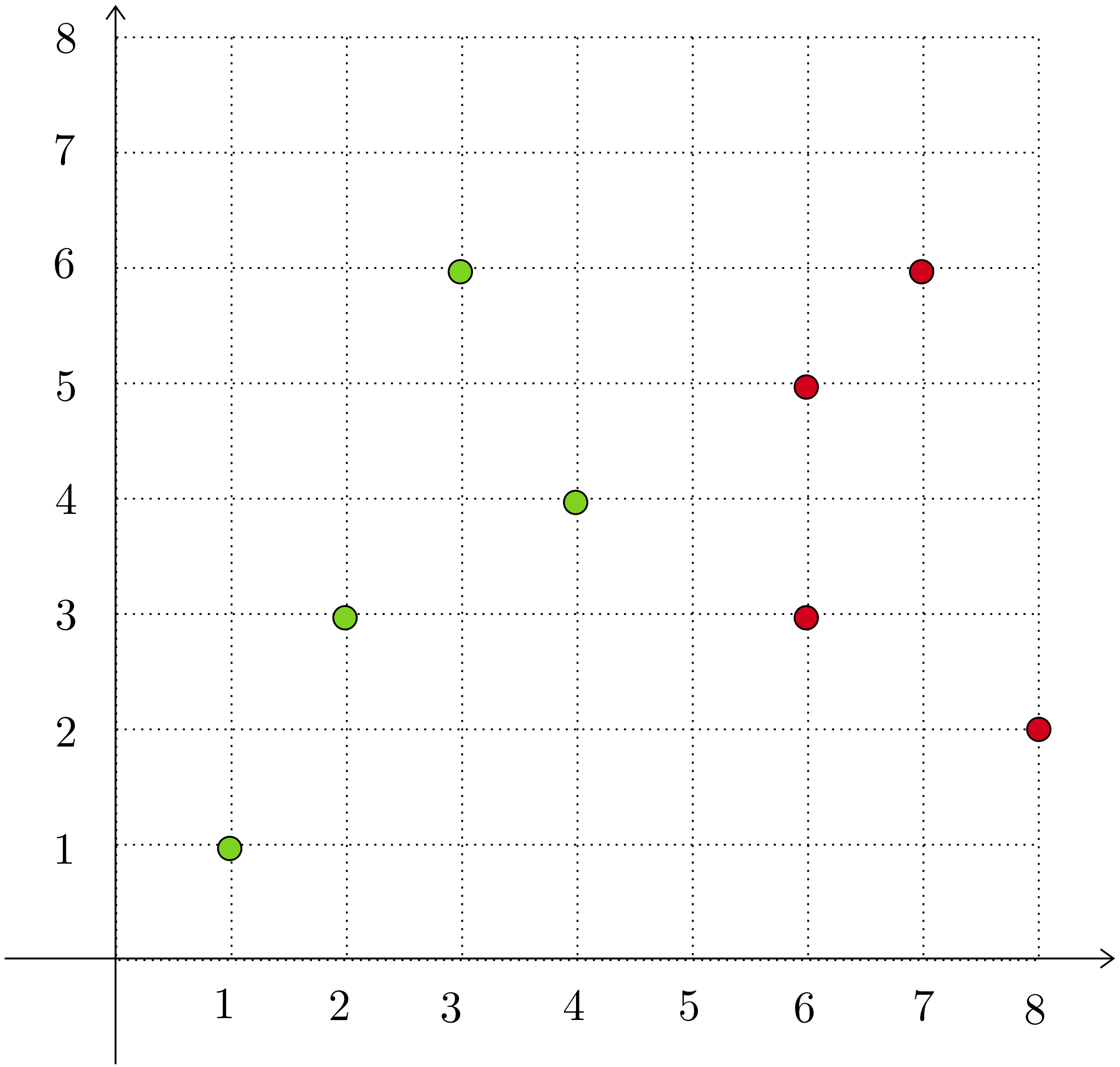




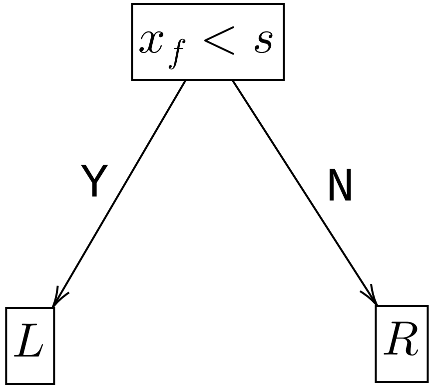
s	$IG(1, s)$
1.5	
2.5	
3.5	
5	
6.5	
7.5	

s	$IG(2, s)$
1.5	
2.5	
3.5	
4.5	
5.5	

Growing a Tree



Best question

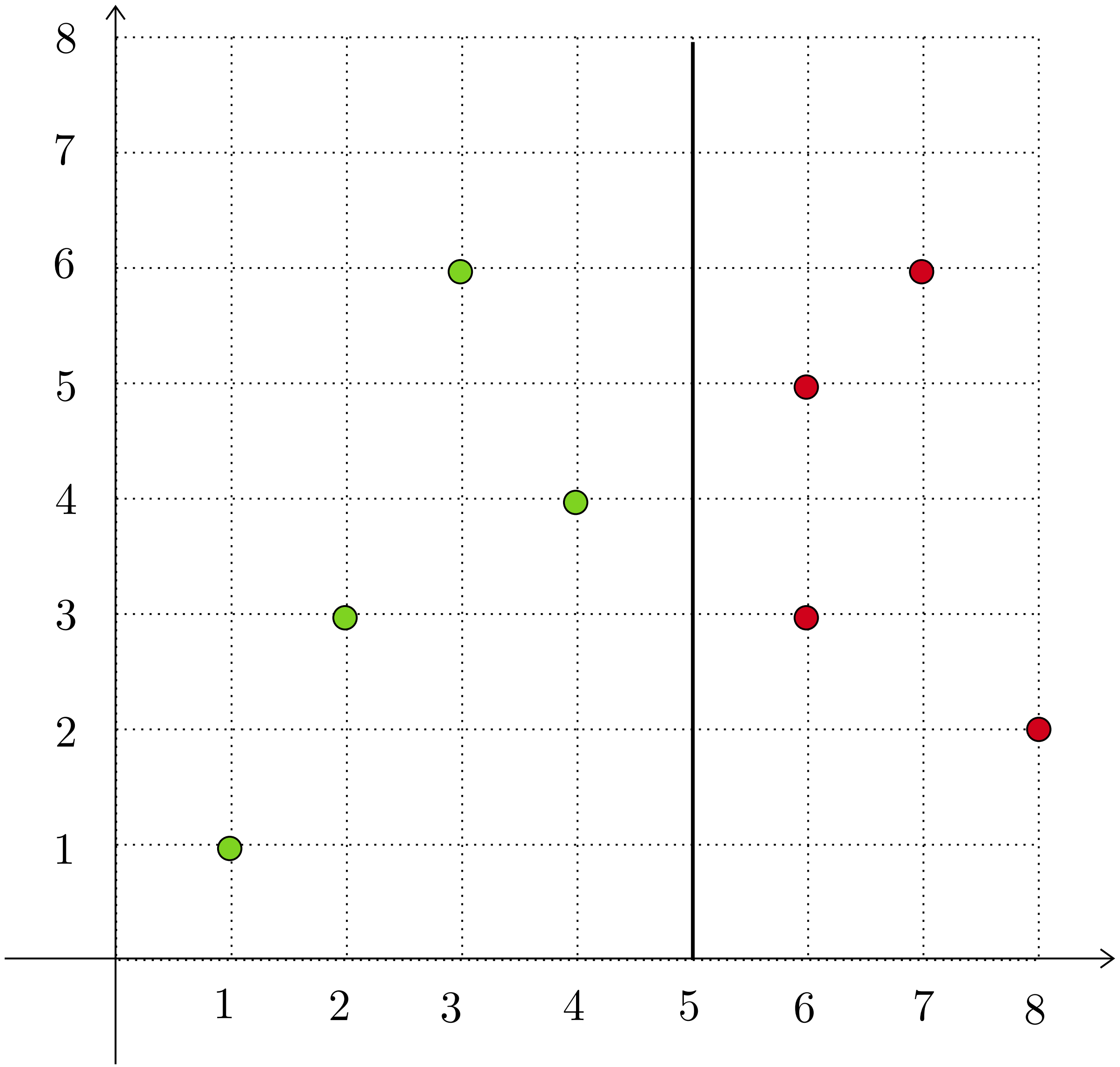


$IG(f, s)$

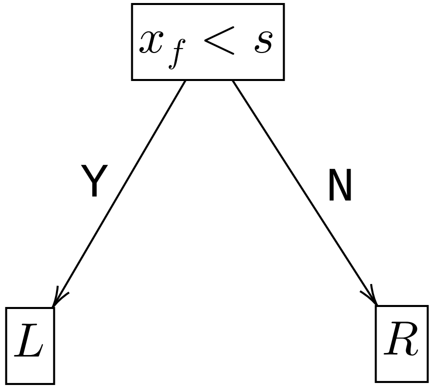
s	$IG(1, s)$
1.5	0.138
2.5	0.311
3.5	0.549
5	1
6.5	0.311
7.5	0.138

s	$IG(2, s)$
1.5	0.138
2.5	0
3.5	0
4.5	0.049
5.5	0

Growing a Tree



Best question



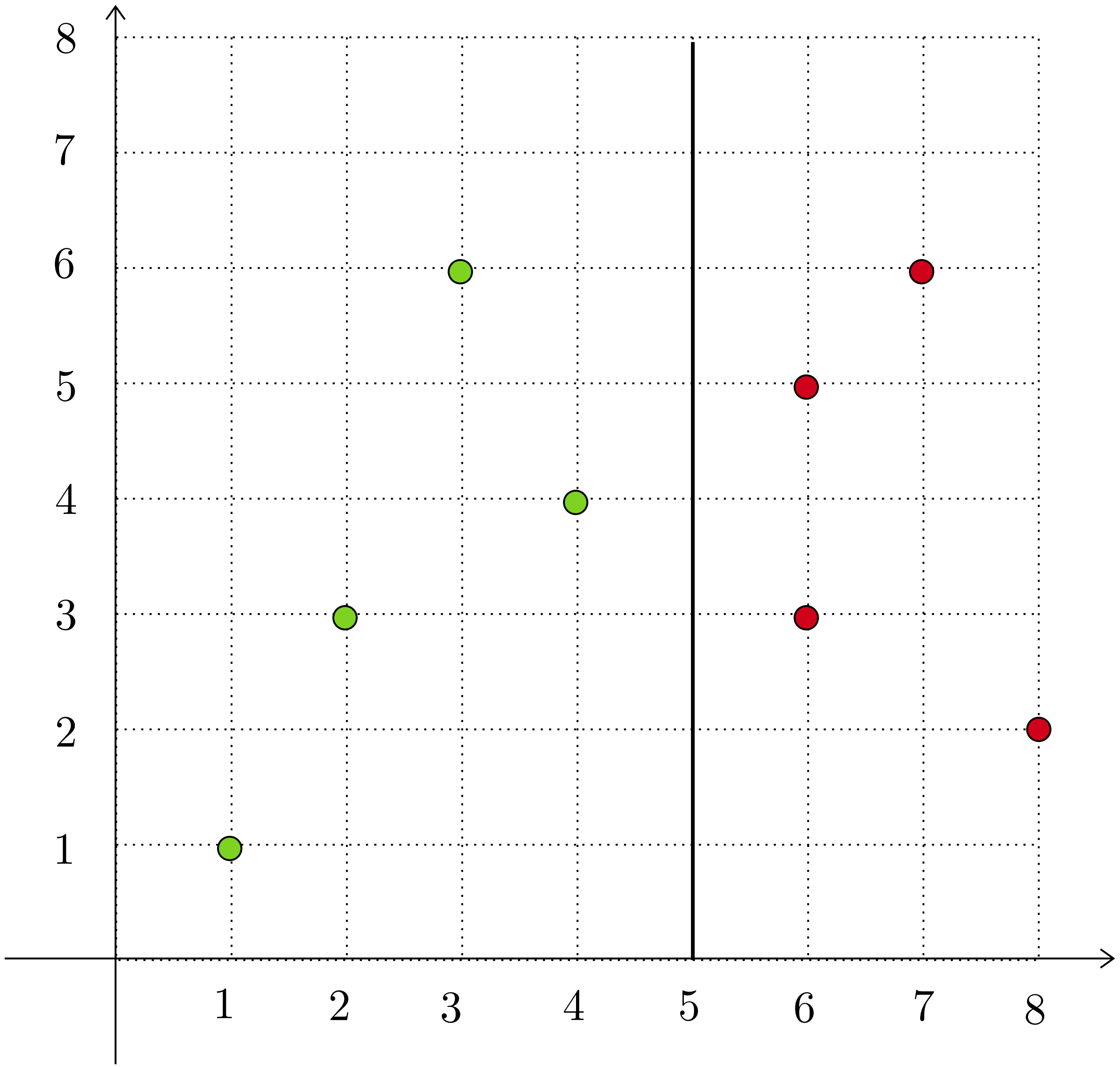
$\boxed{x_1 < 5}$

$\boxed{IG(f, s)}$

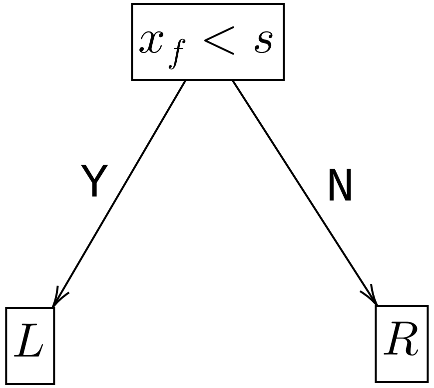
s	$IG(1, s)$
1.5	0.138
2.5	0.311
3.5	0.549
5	**1**
6.5	0.311
7.5	0.138

s	$IG(2, s)$
1.5	0.138
2.5	0
3.5	0
4.5	0.049
5.5	0

Growing a Tree



Best question



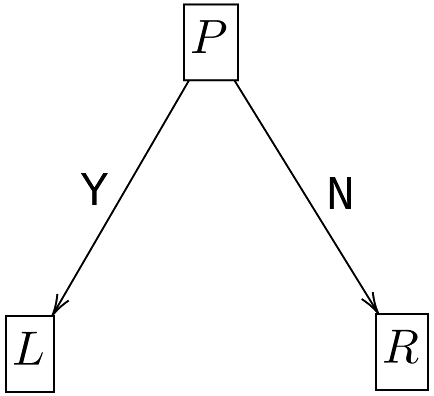
$x_1 < 5$

$IG(f, s)$

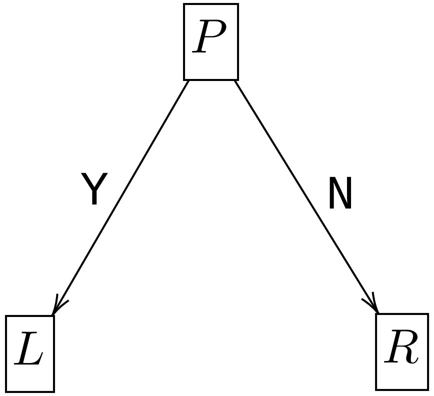
s	$IG(1, s)$
1.5	0.138
2.5	0.311
3.5	0.549
5	**1**
6.5	0.311
7.5	0.138

s	$IG(2, s)$
1.5	0.138
2.5	0
3.5	0
4.5	0.049
5.5	0

Growing a Tree

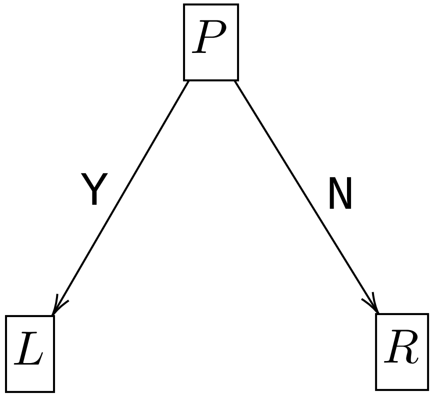


Growing a Tree



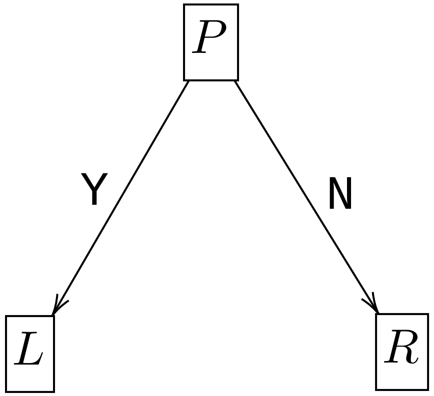
- D: dataset at the parent

Growing a Tree



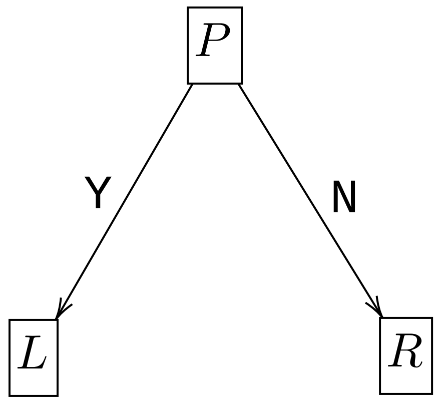
- D: dataset at the parent
- $x_f < s$: question

Growing a Tree



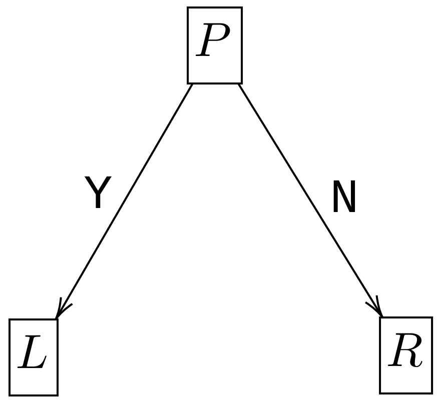
- D: dataset at the parent
- $x_f < s$: question
- D_L and D_R: partitions

Growing a Tree



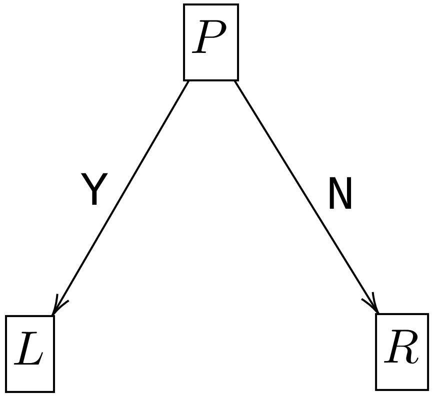
- D: dataset at the parent
- $x_f < s$: question
- D_L and D_R: partitions
- p_P, p_L, p_R: proportions at P, L, R

Growing a Tree



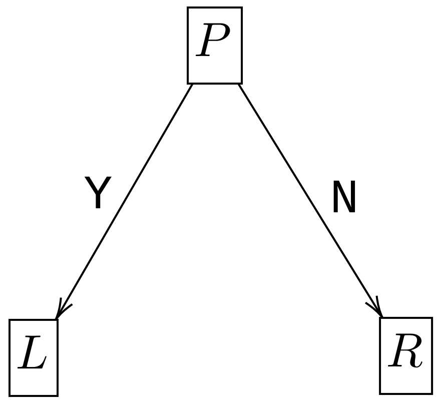
- D: dataset at the parent
- $x_f < s$: question
- D_L and D_R: partitions
- p_P, p_L, p_R: proportions at P, L, R
- γ: proportions of points in L

Growing a Tree



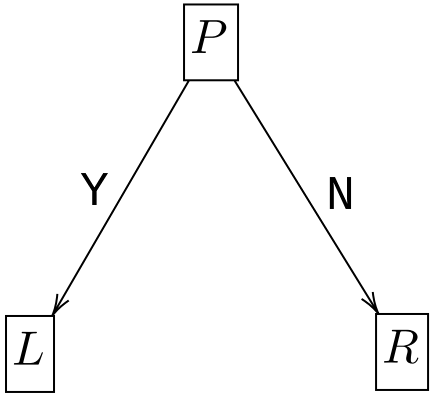
- D: dataset at the parent
- $x_f < s$: question
- D_L and D_R: partitions
- p_P, p_L, p_R: proportions at P, L, R
- γ: proportions of points in L
- E_P: entropy of P
- E_L: entropy of L
- E_R: entropy of R

Growing a Tree



- D: dataset at the parent
- $x_f < s$: question
- D_L and D_R: partitions
- p_P, p_L, p_R: proportions at P, L, R
- γ: proportions of points in L
- E_P: entropy of P
- E_L: entropy of L
- E_R: entropy of R
- IG: information gain

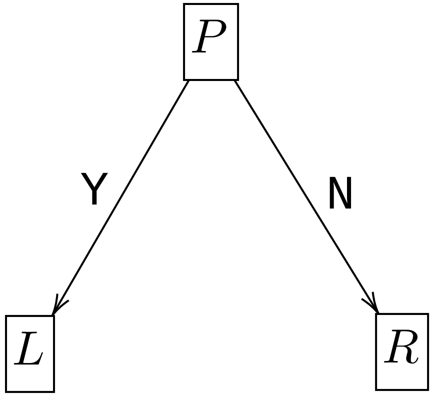
Growing a Tree



- D: dataset at the parent
- $x_f < s$: question
- D_L and D_R: partitions
- p_P, p_L, p_R: proportions at P, L, R
- γ: proportions of points in L
- E_P: entropy of P
- E_L: entropy of L
- E_R: entropy of R
- IG: information gain

$$n_P = n_L + n_R$$

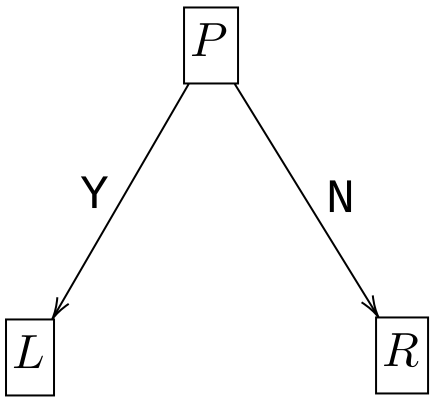
Growing a Tree



- D: dataset at the parent
- $x_f < s$: question
- D_L and D_R: partitions
- p_P, p_L, p_R: proportions at P, L, R
- γ: proportions of points in L
- E_P: entropy of P
- E_L: entropy of L
- E_R: entropy of R
- IG: information gain

$$n_P = n_L + n_R \qquad \gamma = \frac{n_L}{n_P}$$

Growing a Tree

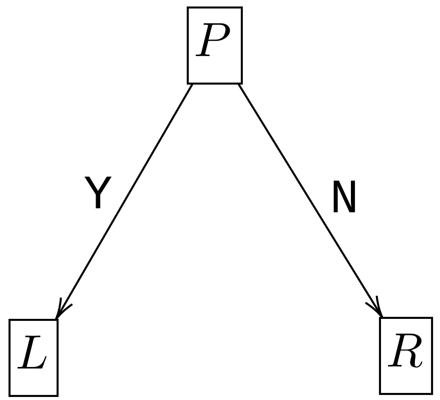


- D: dataset at the parent
- $x_f < s$: question
- D_L and D_R: partitions
- p_P, p_L, p_R: proportions at P, L, R
- γ: proportions of points in L
- E_P: entropy of P
- E_L: entropy of L
- E_R: entropy of R
- IG: information gain

$$n_P = n_L + n_R \qquad \gamma = \frac{n_L}{n_P}$$

$$E = -p\log p - (1-p)\log(1-p)$$

Growing a Tree



- D: dataset at the parent
- $x_f < s$: question
- D_L and D_R: partitions
- p_P, p_L, p_R: proportions at P, L, R
- γ: proportions of points in L
- E_P: entropy of P
- E_L: entropy of L
- E_R: entropy of R
- IG: information gain

$$n_P = n_L + n_R \qquad \gamma = \frac{n_L}{n_P}$$

$$E = -p \log p - (1-p)\log(1-p)$$

$$IG = E_P - [\gamma E_L + (1-\gamma) E_R]$$