

$$p \downarrow l = N \downarrow l / N = 5/6$$

$$N \downarrow l / N = 1/6$$

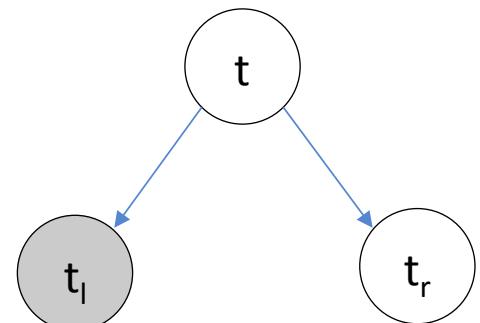
## Impurity, Splits and the Gini Index

- Gini Index:  $i \downarrow G(t) = \sum_k p(c \downarrow k | t)(1 - p(c \downarrow k | t))$

- Impurity decrease due to a split  $s$  applied to a node  $t$ :

$$\Delta i(s, t) = i(t) - p_{\downarrow l} i(t \downarrow l) - p_{\downarrow r} i(t \downarrow r)$$

- Choose splits to **maximize** the decrease in impurity.



# Exercise

The features are all binary, and there  
2 classes (c1 and c2).  
Which feature should I split on first?

X1	X2	X3	Y
0	0	0	c1
0	0	1	c1
0	1	0	c2
0	1	1	c2
0	1	1	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	1	1	c2

- Gini Index:  $i \downarrow G(t) = \sum_{c \downarrow k} p(c | t))$
- Impurity decrease due to a split s applied to a

$$\Delta i(s,t) = i(t) - p \downarrow l i(t \downarrow l)$$

# Exercise

The features are all binary, and there are 2 classes (c1 and c2).

Which feature should I split on first?

Step 1: Calculate Gini impurity of the original node.

$$i \downarrow G(t) = \sum_{k=1}^2 p(c \downarrow k | t)(1 - p(c \downarrow k | t))$$

$$i \downarrow G(t) = 1/5 \cdot 4/5 + 4/5 \cdot 1/5$$

x1	x2	x3	y
0	0	0	c1
0	0	1	c1
0	1	0	c2
0	1	1	c2
0	1	1	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	1	1	c2

- Gini Index:  $i \downarrow G(t) = \sum_{k=1}^2 p(c \downarrow k | t))$

- Impurity decrease due to a split s applied to a

$$\Delta i(s, t) = i(t) - p \downarrow l i(t \downarrow l)$$

# Exercise

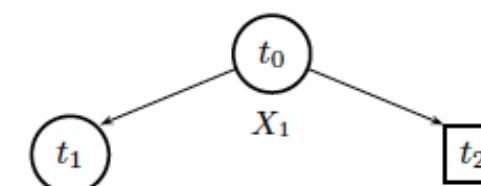
The features are all binary, and there are 2 classes (c1 and c2).  
Which feature should I split on first?

Step 1: Calculate Gini impurity of the original node.

$$i\downarrow G(t) = \sum p(c\downarrow k|t)(1 - p(c\downarrow k|t))$$

$$i\downarrow G(t) = 1/5 \cdot 4/5 + 4/5 \cdot 1/5$$

Step 2: Calculate Gini impurity for the new nodes.



$$\begin{aligned} p(y = c_1|t_1) &= \frac{2}{5} \\ p(y = c_2|t_1) &= \frac{3}{5} \end{aligned}$$

$$\begin{aligned} p(y = c_1|t_2) &= \frac{0}{5} \\ p(y = c_2|t_2) &= \frac{5}{5} \end{aligned}$$

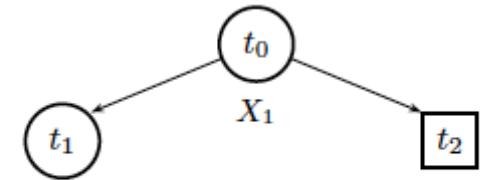
- Gini Index:  $i\downarrow G(t) = \sum p(c\downarrow k|t))$
- Impurity decrease due to a split s applied to a node t

$$\Delta i(s,t) = i(t) - p\downarrow l i(t\downarrow l)$$

X1	X2	X3	Y
0	0	0	c1
0	0	1	c1
0	1	0	c2
0	1	1	c2
0	1	1	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	1	1	c2

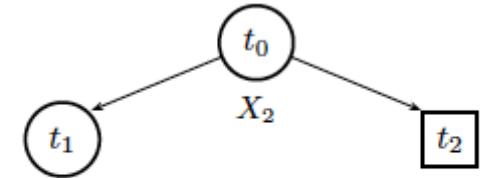
# Exercise

X1	X2	X3	Y
0	0	0	c1
0	0	1	c1
0	1	0	c2
0	1	1	c2
0	1	1	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	0	0	c2
1	1	1	c2



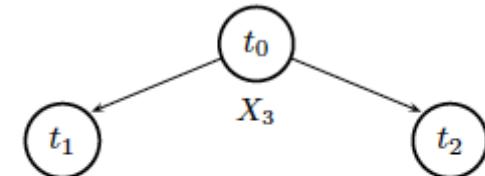
$$p(y = c_1|t_1) = \frac{2}{5} \quad p(y = c_1|t_2) = \frac{0}{5}$$

$$p(y = c_2|t_1) = \frac{3}{5} \quad p(y = c_2|t_2) = \frac{5}{5}$$



$$p(y = c_1|t_1) = \frac{2}{6} \quad p(y = c_1|t_2) = \frac{0}{4}$$

$$p(y = c_2|t_1) = \frac{4}{6} \quad p(y = c_2|t_2) = \frac{4}{4}$$



$$p(y = c_1|t_1) = \frac{1}{6} \quad p(y = c_1|t_2) = \frac{1}{4}$$

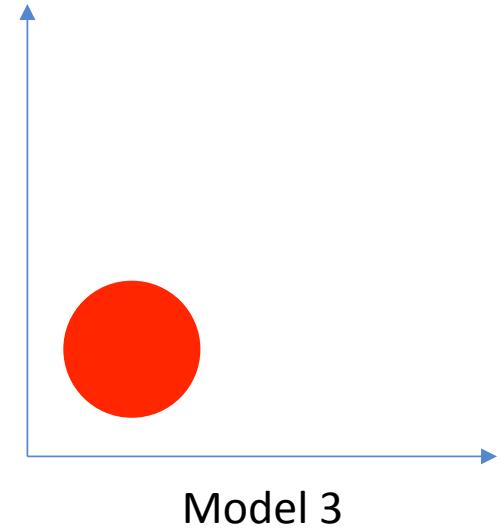
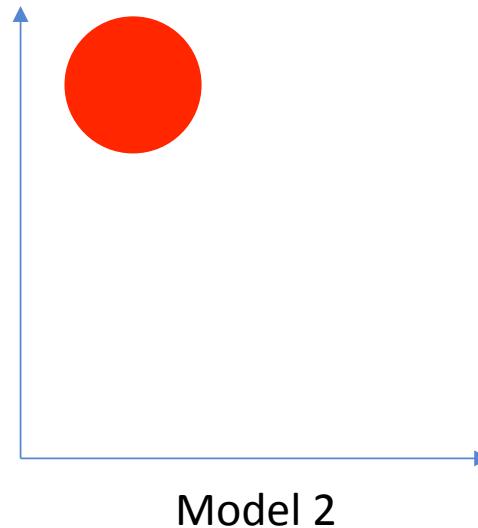
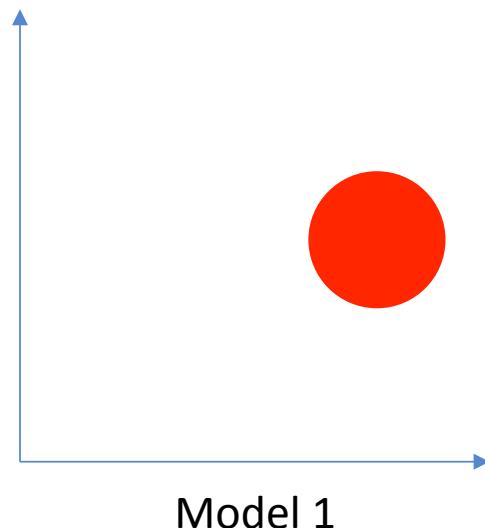
$$p(y = c_2|t_1) = \frac{5}{6} \quad p(y = c_2|t_2) = \frac{3}{4}$$

- Gini Index:  $i \downarrow G(t) = \sum k p(c_k | t)$

- Impurity decrease due to a split s applied to a

$$\Delta i(s,t) = i(t) - p \downarrow l i(t \downarrow l)$$

## Random Forests: Intuition



# Random Forests: Intuition

