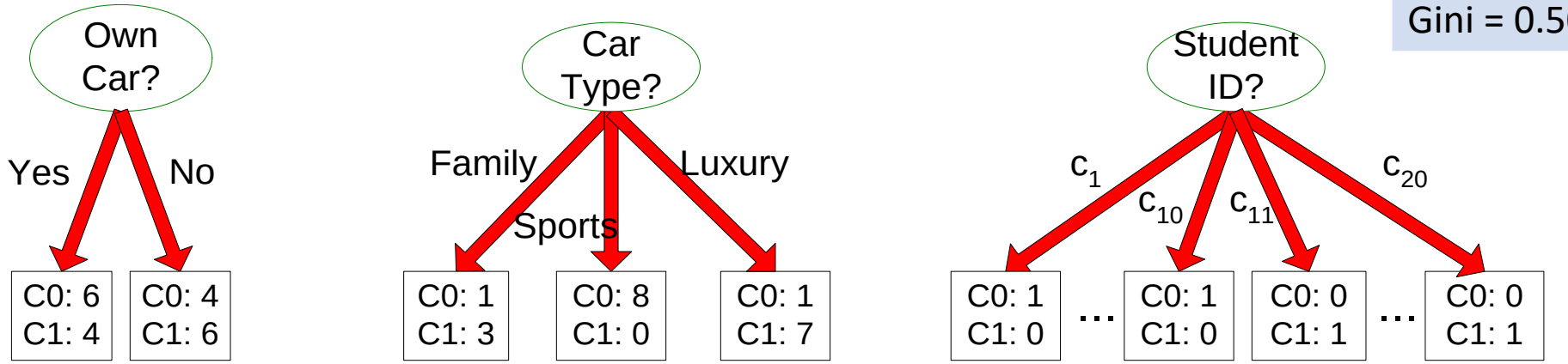
Categorical Attributes: Computing Gini Index

Before Splitting: 10 records of class 0,
10 records of class 1

	Parent
C0	10
C1	10
Gini = 0.50	

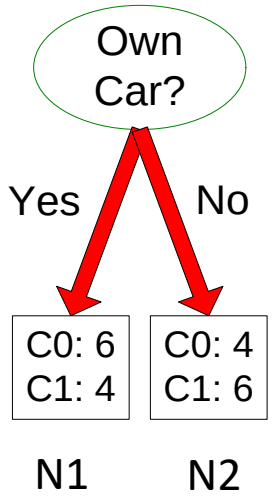


Which test condition is the best?

Categorical Attributes: Computing Gini Index

Before Splitting: 10 records of class 0,
10 records of class 1

	Parent
C0	10
C1	10
Gini = 0.50	



Gini(N1)
= 1 – (6/10)² – (4/10)²
= 0.48

Gini(N2)
= 1 – (4/10)² – (6/10)²
= 0.48

	N1	N2
C0	6	4
C1	4	6
Gini = 0.48		

Gini(Children)

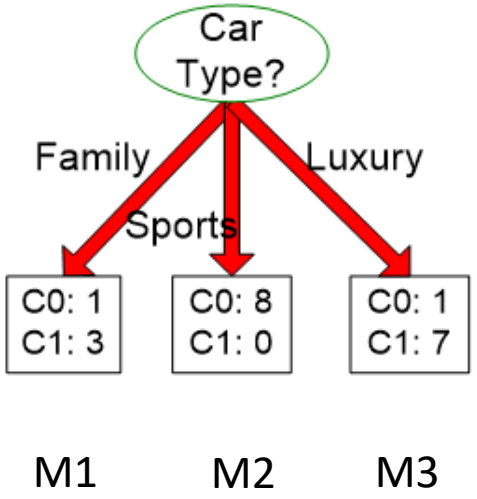
= (10/20 * 0.48) + (10/20 * 0.48)
= 0.48

Gain = 0.50 – 0.48 = 0.02

Categorical Attributes: Computing Gini Index

Before Splitting: 10 records of class 0,
10 records of class 1

	Parent
C0	10
C1	10
Gini = 0.50	



Gini(M1)
$= 1 - (1/4)^2 - (3/4)^2$
$= 0.375$

Gini(M2)
$= 1 - (8/8)^2 - (0/8)^2$
$= 0$

Gini(M3)
$= 1 - (1/8)^2 - (7/8)^2$
$= 0.219$

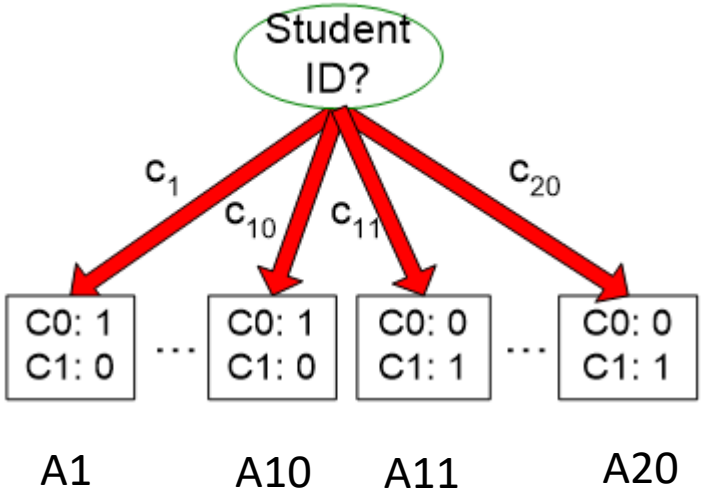
	M1	M2	M3
C0	1	8	1
C1	3	0	7
Gini = 0.163			

Gini(Children)

$= (4/20 * 0.375) + (8/20 * 0) + (8/20 * 0.219)$
$= 0.163$

Gain = 0.50 – 0.48 = 0.337

Categorical Attributes: Computing Gini Index



Before Splitting: 10 records of class 0,
10 records of class 1

	Parent
C0	10
C1	10
Gini = 0.50	

Gini(A1)
= 1 – (1/1)² – (0/1)²
= 0

. . .

Gini(A20)
= 1 – (0/1)² – (1/1)²
= 0

	A1	…	A20
C0	1	…	0
C1	0	…	1
Gini = 0			

Gini(Children)

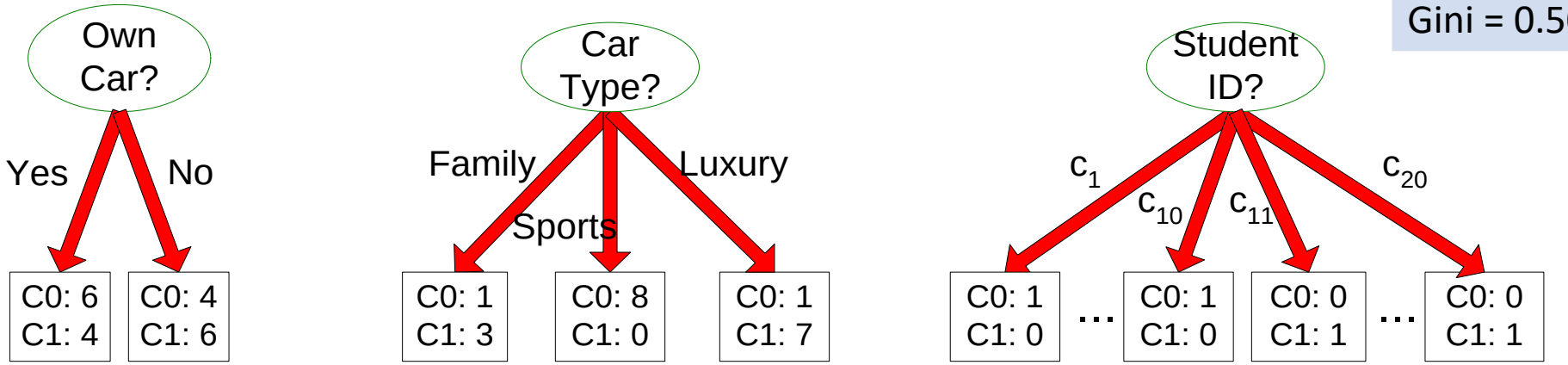
= $\sum_{1}^{20}((1/10) * 0)$
= 0

Gain = 0.50 – 0 = 0.5

Categorical Attributes: Computing Gini Index

Before Splitting: 10 records of class 0,
10 records of class 1

	Parent
C0	10
C1	10
Gini = 0.50	



Which test condition is the best?

ถึงแม้ว่า Student ID? จะมีความสมบูรณ์ในการแยกมากกว่าใคร คือ Gain = 0.5 แต่จำนวน เงื่อนไข (test condition) มีมากเกินไปถึง 20 เงื่อนไข ดังนั้น เพื่อให้ต้นไม้ตัดสินใจที่สร้างทำงานได้ดี และป้องกันการเกิดปัญหา Overfiting จึงควรเลือก Car Type? ครับ