Data Mining

Classification I – Decision Tree (Part B)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

/

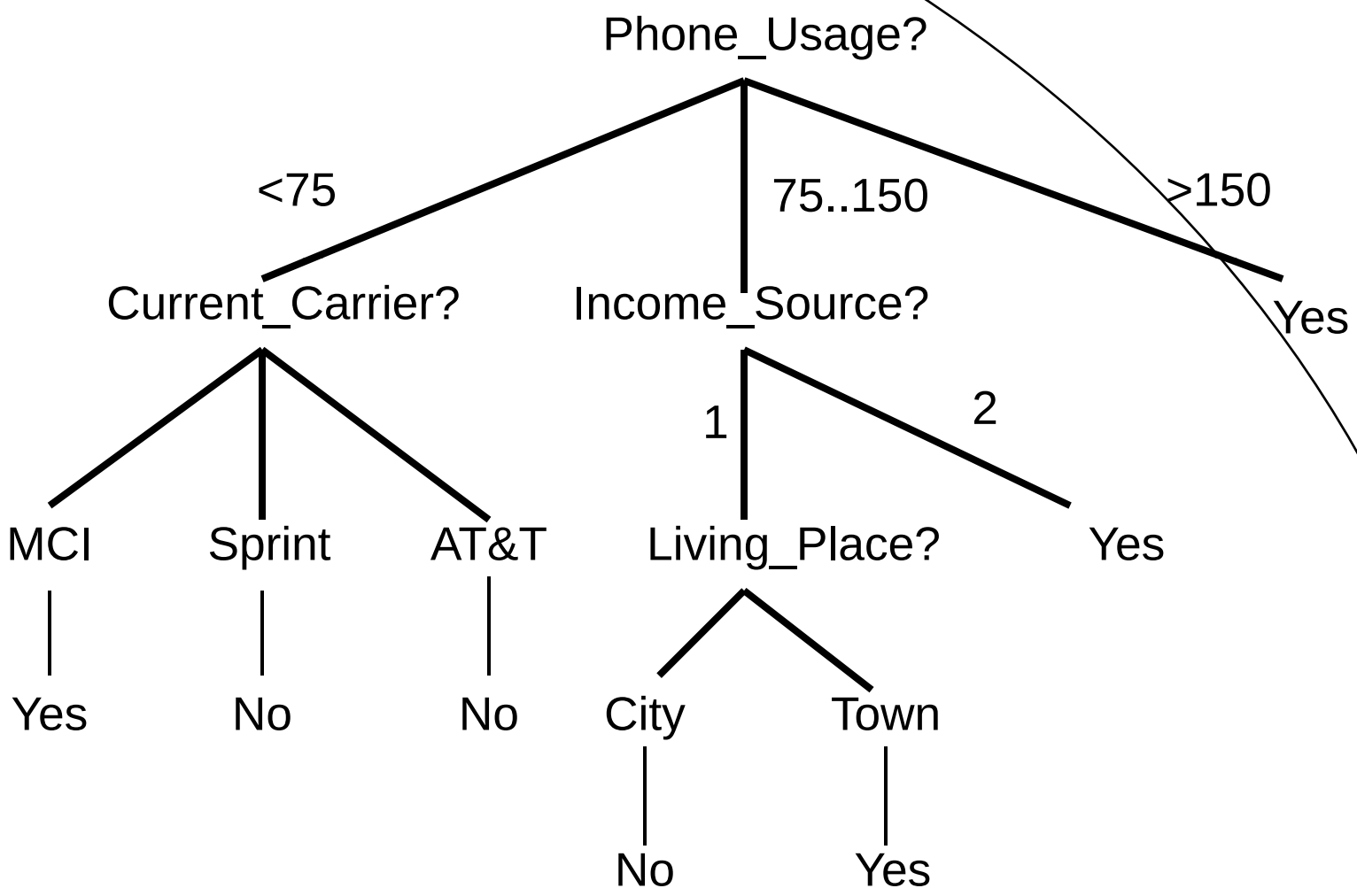
Input: A Training Dataset for

"Change to new AT&T calling plan"

Phone_Usage	Income_Source	Living_Place	Current_Carrier	Change_Plan
<75	1	City	MCI	Yes
75..150	2	Town	MCI	Yes
<75	1	City	Sprint	No
>150	2	Town	AT&T	Yes
75..150	1	City	MCI	No
75..150	2	Town	AT&T	Yes
<75	2	Town	AT&T	No
>150	2	City	Sprint	Yes
<75	1	City	AT&T	No
75..150	1	Town	MCI	Yes
75..150	2	City	Sprint	Yes
>150	1	Town	AT&T	Yes
<75	1	Town	AT&T	No
<75	2	City	Sprint	No
75..150	2	City	MCI	Yes

Output: A Decision Tree for

“Change to new AT&T calling plan”



Attribute Selection by Information Gain Computation

- Class P: change_plan = “yes”
- Class N: change_plan = “no”
- Compute the expected information needed to classify a given sample:

$$I(p,n) = I(9,6) = -\frac{9}{15}\log_2\frac{9}{15} - \frac{6}{15}\log_2\frac{6}{15} = 0.971$$

Information Gain Computation
(1) Phone Usage

- Compute the entropy and information gain

P_Usage	p_i	n_i	$I(p_i, n_i)$
<75	1	5	0.65
75..150	5	1	0.65
>150	3	0	0

$$E(P\_Usage) = \frac{6}{15} I(1,5) + \frac{6}{15} I(5,1) + \frac{3}{15} I(3,0) = 0.52$$

$$Gain(P\_Usage) = I(p,n) - E(P\_Usage) = 0.971 - 0.520 = 0.451$$

Information Gain Computation
(2) Income Source

- Compute the entropy and information gain

I_Source	p_i	n_i	$I(p_i, n_i)$
1	3	4	0.985
2	6	2	0.811

$$E(I\_Source) = \frac{7}{15} I(3,4) + \frac{8}{15} I(6,2) = 0.892$$

$$Gain(I\_Source) = I(p,n) - E(I\_Source)$$
$$= 0.971 - 0.892 = 0.079$$

Information Gain Computation
(3) Living Place

- Compute the entropy and information gain

L_Place	p_i	n_i	$I(p_i, n_i)$
City	4	4	1
Town	5	2	0.864

$$E(L\_Place) = \frac{8}{15} I(4,4) + \frac{7}{15} I(5,2) = 0.936$$

$$Gain(L\_Place) = I(p,n) - E(L\_Place)$$
$$= 0.971 - 0.936 = 0.035$$

Information Gain Computation
(4) Current Carrier

- Compute the entropy and information gain

C_Carrier	p_i	n_i	$I(p_i, n_i)$
MCI	4	1	0.722
Sprint	2	2	1
AT&T	3	3	1

$$E(C\_Carrier) = \frac{5}{15} I(4,1) + \frac{4}{15} I(2,2) + \frac{6}{15} I(3,3) = 0.908$$

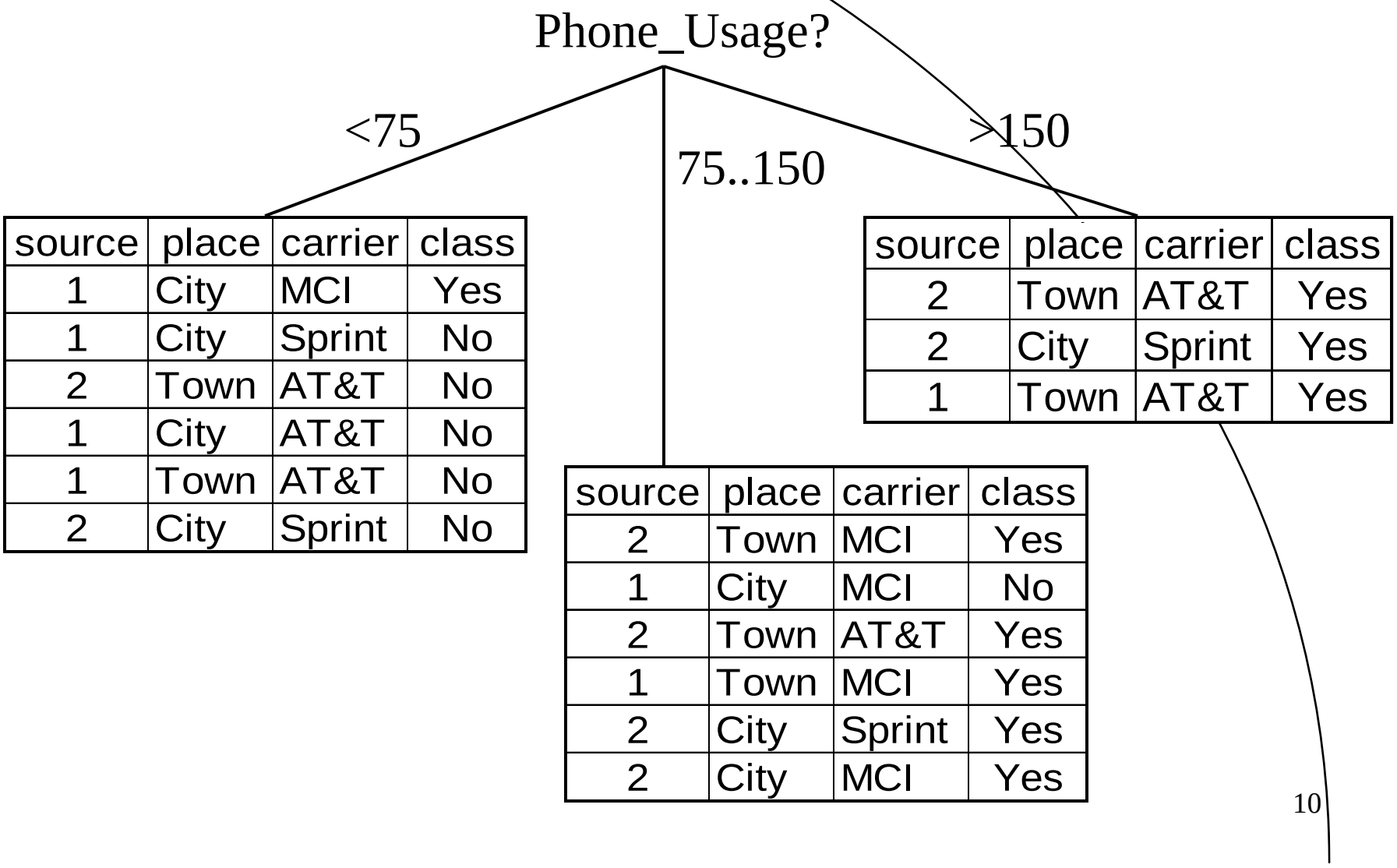
$$Gain(C\_Carrier) = I(p,n) - E(C\_Carrier) = 0.971 - 0.908 = 0.063$$

Information Gain Computation

- Select Highest Gain

$$Gain\ (Phone\ \_\ Usage\ ) = 0.451$$
$$Gain\ (Income\ \_\ Source\ ) = 0.079$$
$$Gain\ (Living\ \_\ Place\ ) = 0.035$$
$$Gain\ (Current\ \_\ Carrier\ ) = 0.063$$

Partitioned Training Data Set by "Phone Usage"

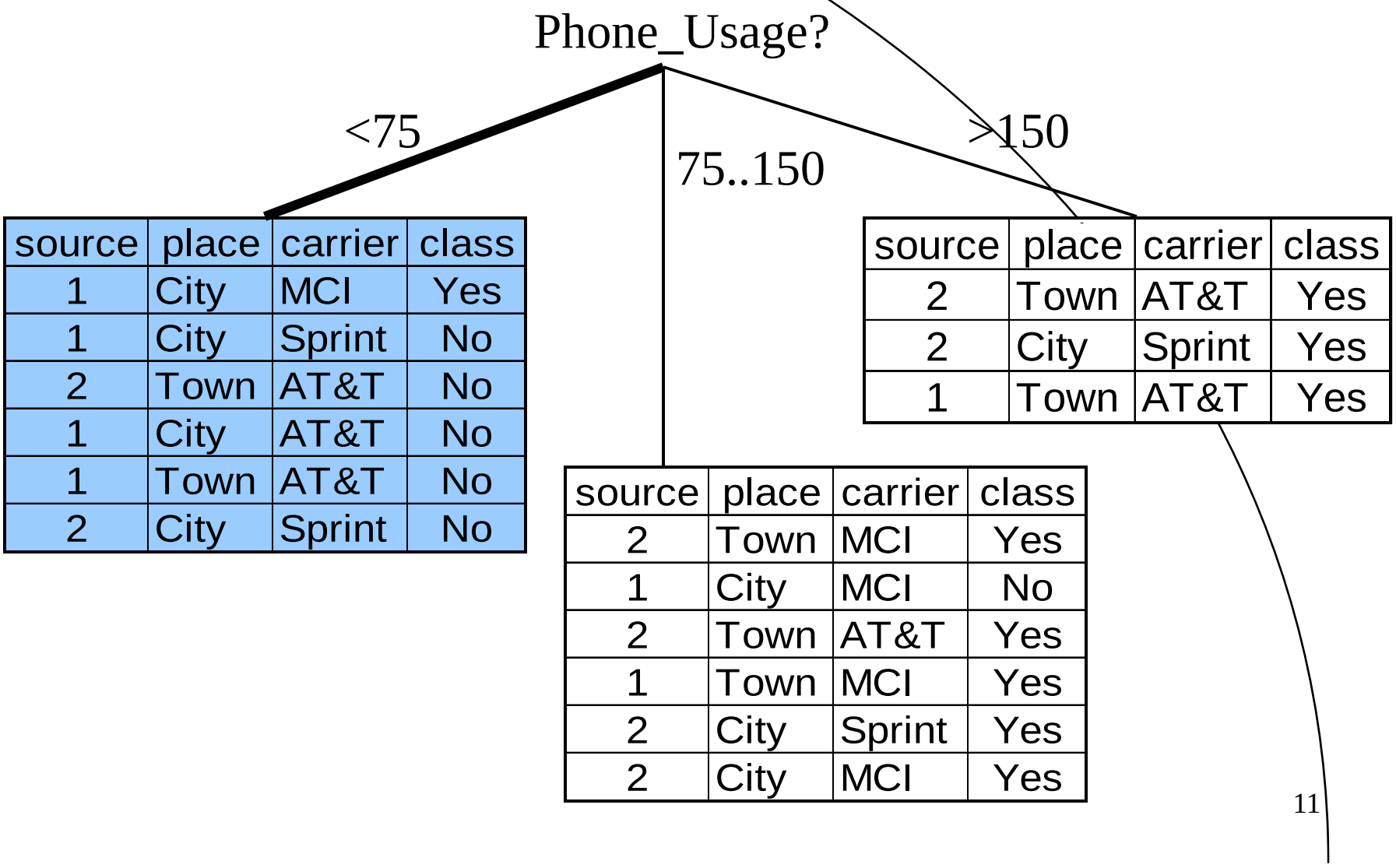


source	place	carrier	class
1	City	MCI	Yes
1	City	Sprint	No
2	Town	AT&T	No
1	City	AT&T	No
1	Town	AT&T	No
2	City	Sprint	No

source	place	carrier	class
2	Town	MCI	Yes
1	City	MCI	No
2	Town	AT&T	Yes
1	Town	MCI	Yes
2	City	Sprint	Yes
2	City	MCI	Yes

source	place	carrier	class
2	Town	AT&T	Yes
2	City	Sprint	Yes
1	Town	AT&T	Yes

Partitioned Training Data Set by “Phone Usage”



source	place	carrier	class
1	City	MCI	Yes
1	City	Sprint	No
2	Town	AT&T	No
1	City	AT&T	No
1	Town	AT&T	No
2	City	Sprint	No

source	place	carrier	class
2	Town	MCI	Yes
1	City	MCI	No
2	Town	AT&T	Yes
1	Town	MCI	Yes
2	City	Sprint	Yes
2	City	MCI	Yes

source	place	carrier	class
2	Town	AT&T	Yes
2	City	Sprint	Yes
1	Town	AT&T	Yes

Information Gain Computation for < 75 (1) Income Source

$$I(p,n) = I(1,5) = -\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6} = 0.650$$

- Compute the entropy and information gain

I_Source	p_i	n_i	$I(p_i, n_i)$
1	1	3	0.811
2	0	2	0

$$E(I\_Source) = \frac{4}{6}I(1,3) + \frac{2}{6}I(0,2) = 0.541$$

$$Gain(I\_Source) = I(p,n) - E(I\_Source)$$
$$= 0.650 - 0.541 = 0.109$$

Information Gain Computation for < 75 (2) Living Place

- Compute the entropy and information gain

L_Place	p_i	n_i	$I(p_i, n_i)$
City	1	3	0.811
Town	0	2	0

$$E(L\_Place) = \frac{4}{6} I(1,3) + \frac{2}{6} I(0,2) = 0.541$$

$$Gain(L\_Place) = I(p,n) - E(L\_Place)$$
$$= 0.650 - 0.541 = 0.109$$

Information Gain Computation for < 75 (3) Current Carrier

- Compute the entropy and information gain

C_Carrier	p_i	n_i	$I(p_i, n_i)$
MCI	1	0	0
Sprint	0	2	0
AT&T	0	3	0

$$E(C\_Carrier) = \frac{1}{6} I(1,0) + \frac{2}{6} I(0,2) + \frac{3}{6} I(0,3) = 0$$

$$Gain(C\_Carrier) = I(p,n) - E(C\_Carrier) = 0.650$$

Information Gain Computation for < 75

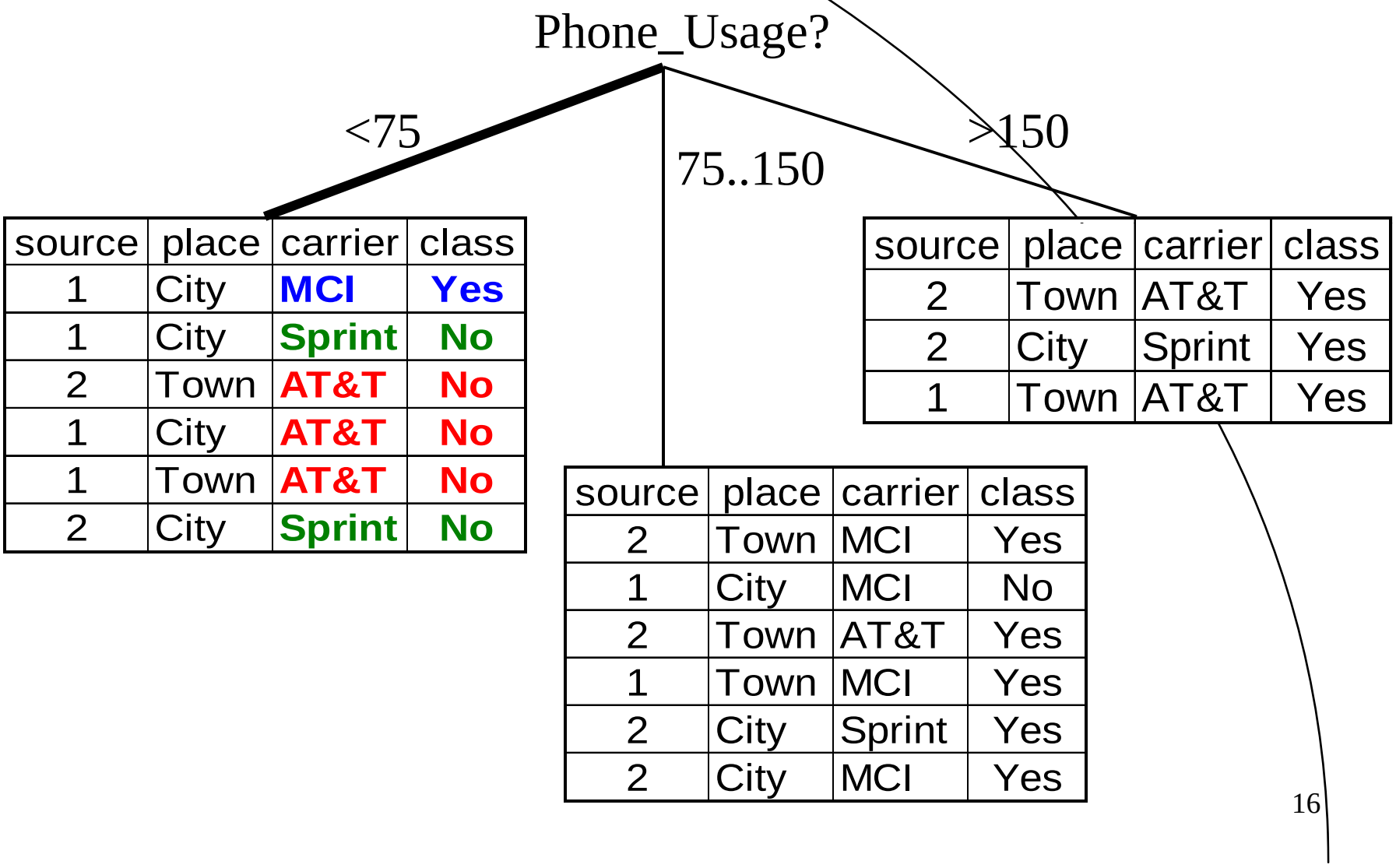
- Select Highest Gain

$$Gain\ (Income\ \_\ Source\ ) = 0.109$$

$$Gain\ (Living\ \_\ Place\ ) = 0.109$$

$$Gain\ (Current\ \_\ Carrier\ ) = 0.650$$

Partitioned Training Data Set by “Phone Usage”

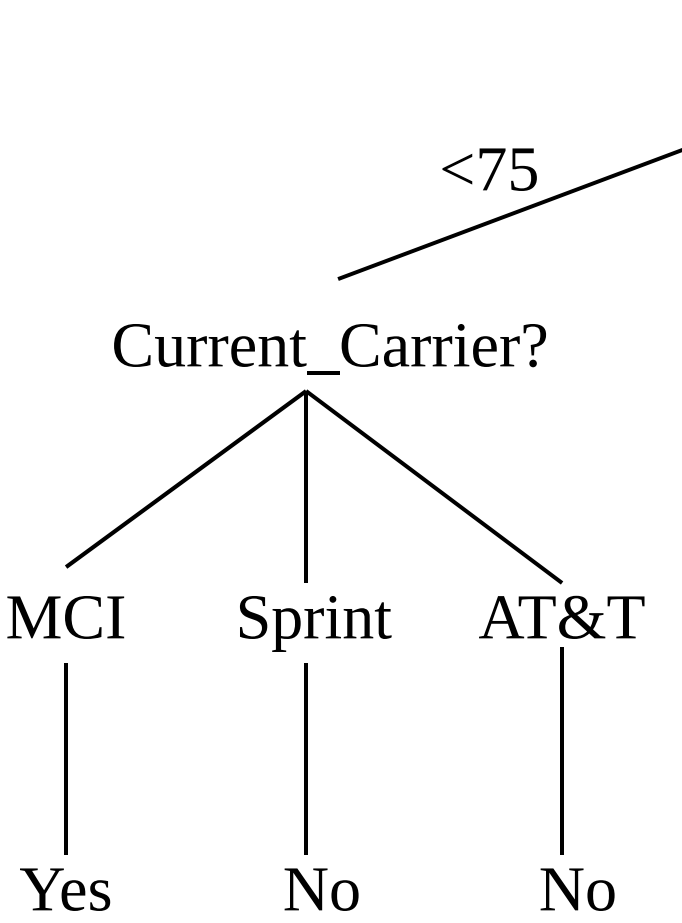


source	place	carrier	class
1	City	**MCI**	**Yes**
1	City	**Sprint**	**No**
2	Town	**AT&T**	**No**
1	City	**AT&T**	**No**
1	Town	**AT&T**	**No**
2	City	**Sprint**	**No**

source	place	carrier	class
2	Town	MCI	Yes
1	City	MCI	No
2	Town	AT&T	Yes
1	Town	MCI	Yes
2	City	Sprint	Yes
2	City	MCI	Yes

source	place	carrier	class
2	Town	AT&T	Yes
2	City	Sprint	Yes
1	Town	AT&T	Yes

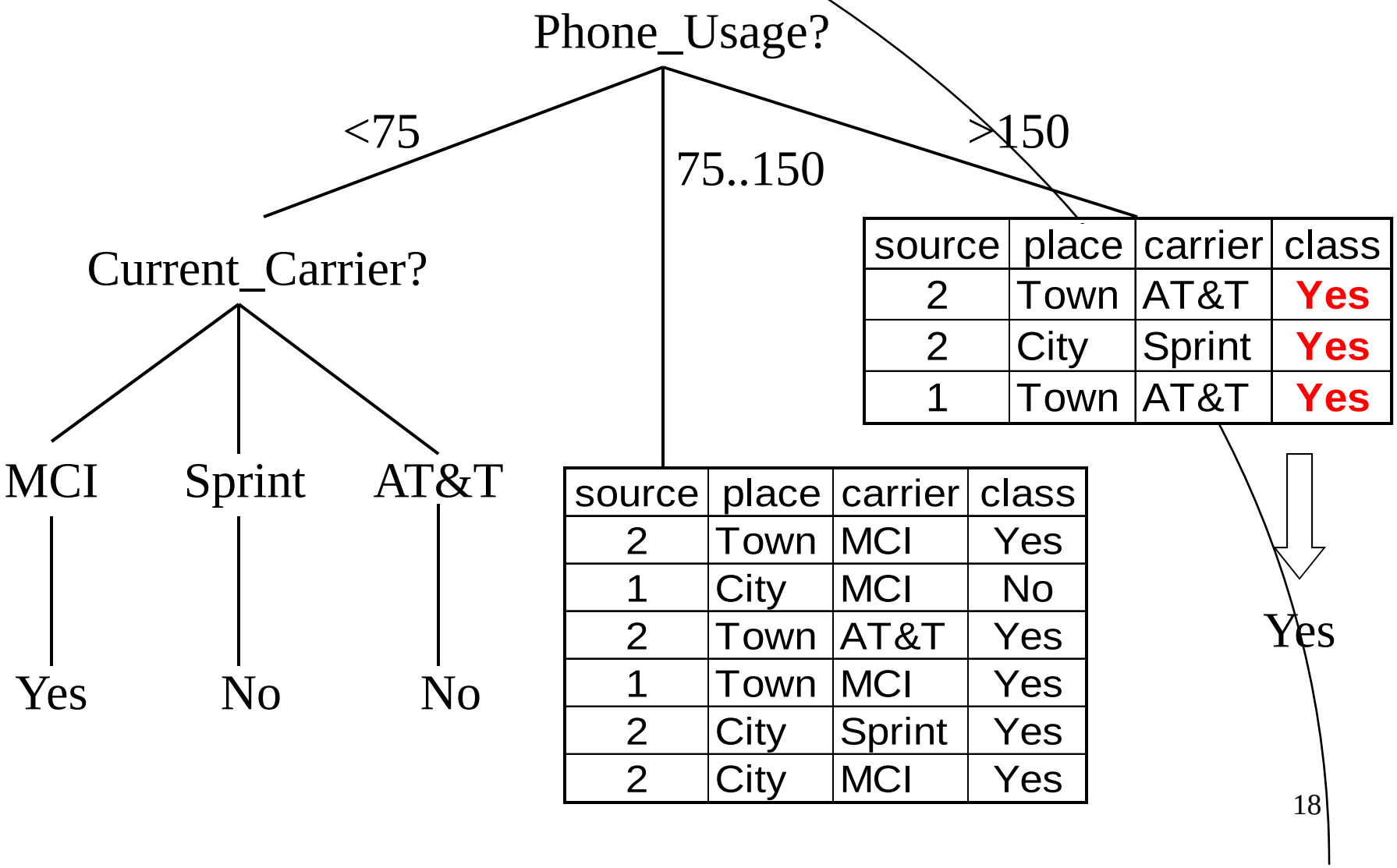
Partitioned Training Data Set and Partial Decision Tree



source	place	carrier	class
2	Town	AT&T	Yes
2	City	Sprint	Yes
1	Town	AT&T	Yes

source	place	carrier	class
2	Town	MCI	Yes
1	City	MCI	No
2	Town	AT&T	Yes
1	Town	MCI	Yes
2	City	Sprint	Yes
2	City	MCI	Yes

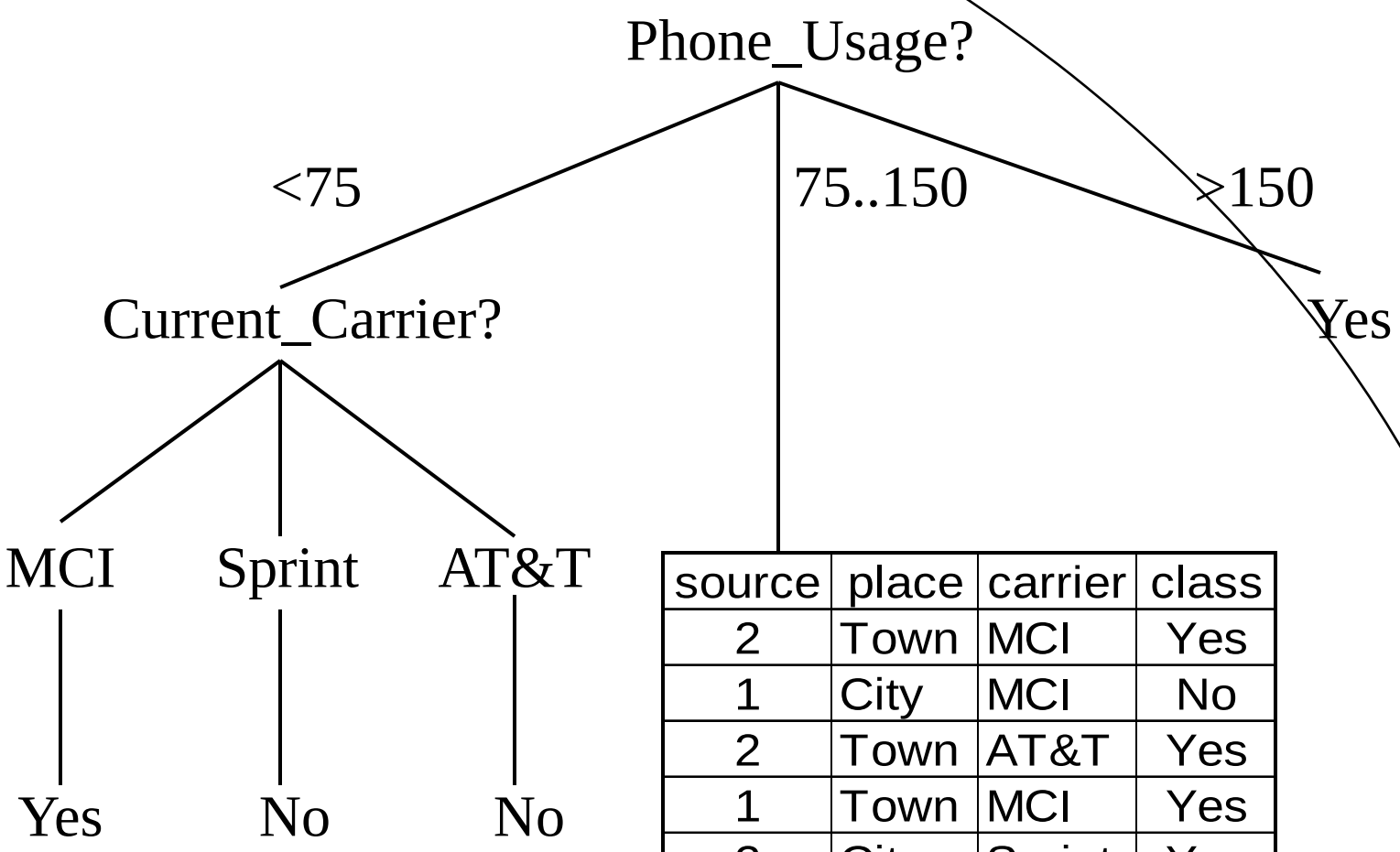
Partitioned Training Data Set and Partial Decision Tree



source	place	carrier	class
2	Town	AT&T	**Yes**
2	City	Sprint	**Yes**
1	Town	AT&T	**Yes**

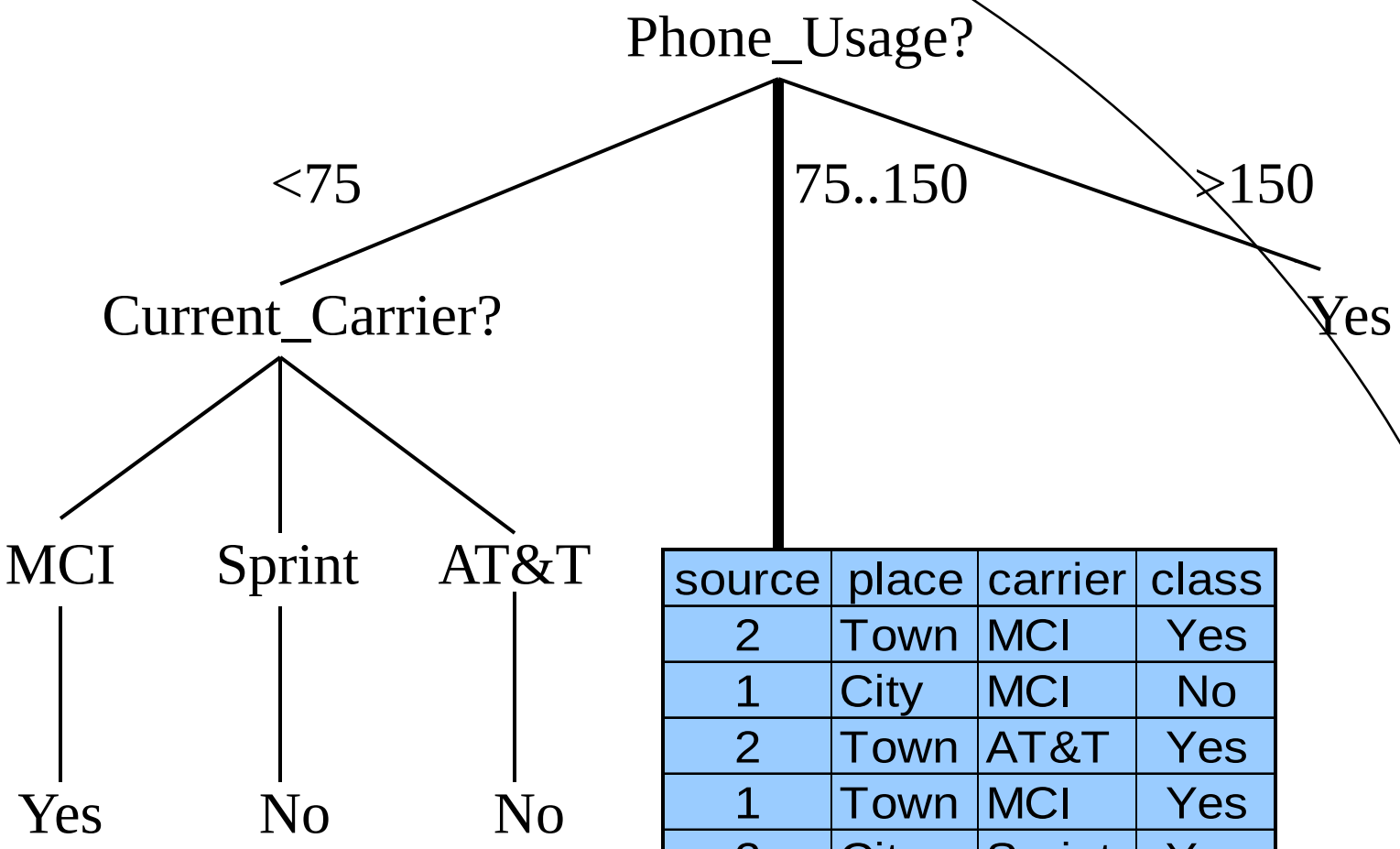
source	place	carrier	class
2	Town	MCI	Yes
1	City	MCI	No
2	Town	AT&T	Yes
1	Town	MCI	Yes
2	City	Sprint	Yes
2	City	MCI	Yes

Partitioned Training Data Set and Partial Decision Tree



source	place	carrier	class
2	Town	MCI	Yes
1	City	MCI	No
2	Town	AT&T	Yes
1	Town	MCI	Yes
2	City	Sprint	Yes
2	City	MCI	Yes

Partitioned Training Data Set and Partial Decision Tree



source	place	carrier	class
2	Town	MCI	Yes
1	City	MCI	No
2	Town	AT&T	Yes
1	Town	MCI	Yes
2	City	Sprint	Yes
2	City	MCI	Yes

Information Gain Computation for 75..150 (1) Income Source

$$I(p,n) = I(5,1) = -\frac{5}{6}\log_2\frac{5}{6} - \frac{1}{6}\log_2\frac{1}{6} = 0.650$$

- Compute the entropy and information gain

I_Source	p_i	n_i	$I(p_i, n_i)$
1	1	1	1
2	4	0	0

$$E(I\_Source) = \frac{2}{6}I(1,1) + \frac{4}{6}I(4,0) = 0.333$$

$$Gain(I\_Source) = I(p,n) - E(I\_Source)$$
$$= 0.650 - 0.333 = 0.317$$

Information Gain Computation for 75..150 (2) Living Place

- Compute the entropy and information gain

L_Place	p_i	n_i	$I(p_i, n_i)$
City	2	1	0.919
Town	3	0	0

$$E(L\_Place) = \frac{3}{6} I(2,1) + \frac{3}{6} I(3,0) = 0.460$$

$$Gain(L\_Place) = I(p,n) - E(L\_Place)$$
$$= 0.650 - 0.460 = 0.190$$

Information Gain Computation for 75..150 (3) Current Carrier

- Compute the entropy and information gain

C_Carrier	p_i	n_i	$I(p_i, n_i)$
MCI	3	1	0.811
Sprint	1	0	0
AT&T	1	0	0

$$E(C\_Carrier) = \frac{4}{6} I(3,1) + \frac{1}{6} I(1,0) + \frac{1}{6} I(1,0) = 0.541$$

$$Gain(C\_Carrier) = I(p,n) - E(C\_Carrier)$$
$$= 0.650 - 0.541 = 0.109$$

Information Gain Computation for 75..150

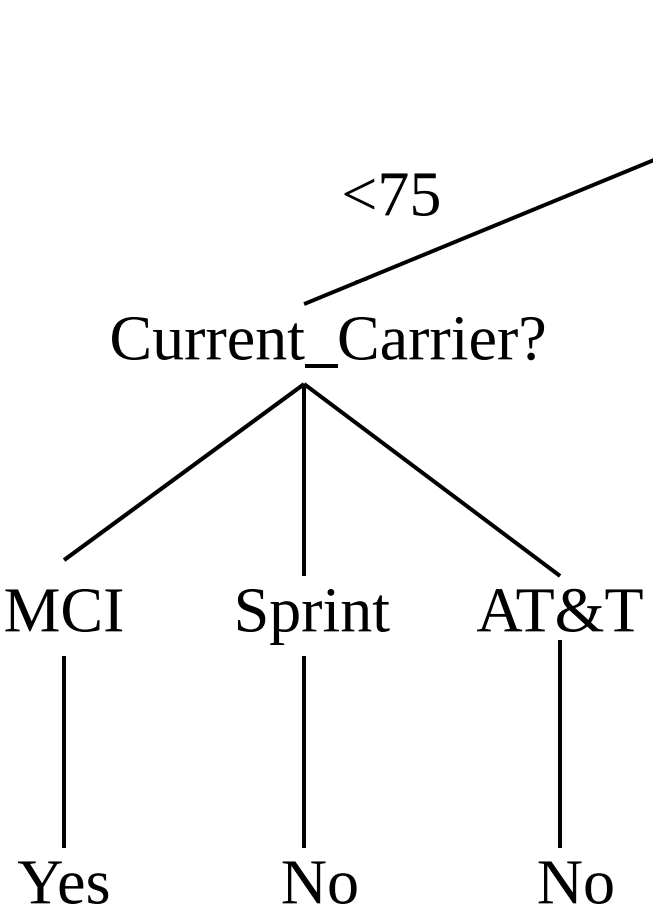
- Select Highest Gain

$$Gain\ (Income\ \_\ Source\ ) = 0.317$$

$$Gain\ (Living\ \_\ Place\ ) = 0.190$$

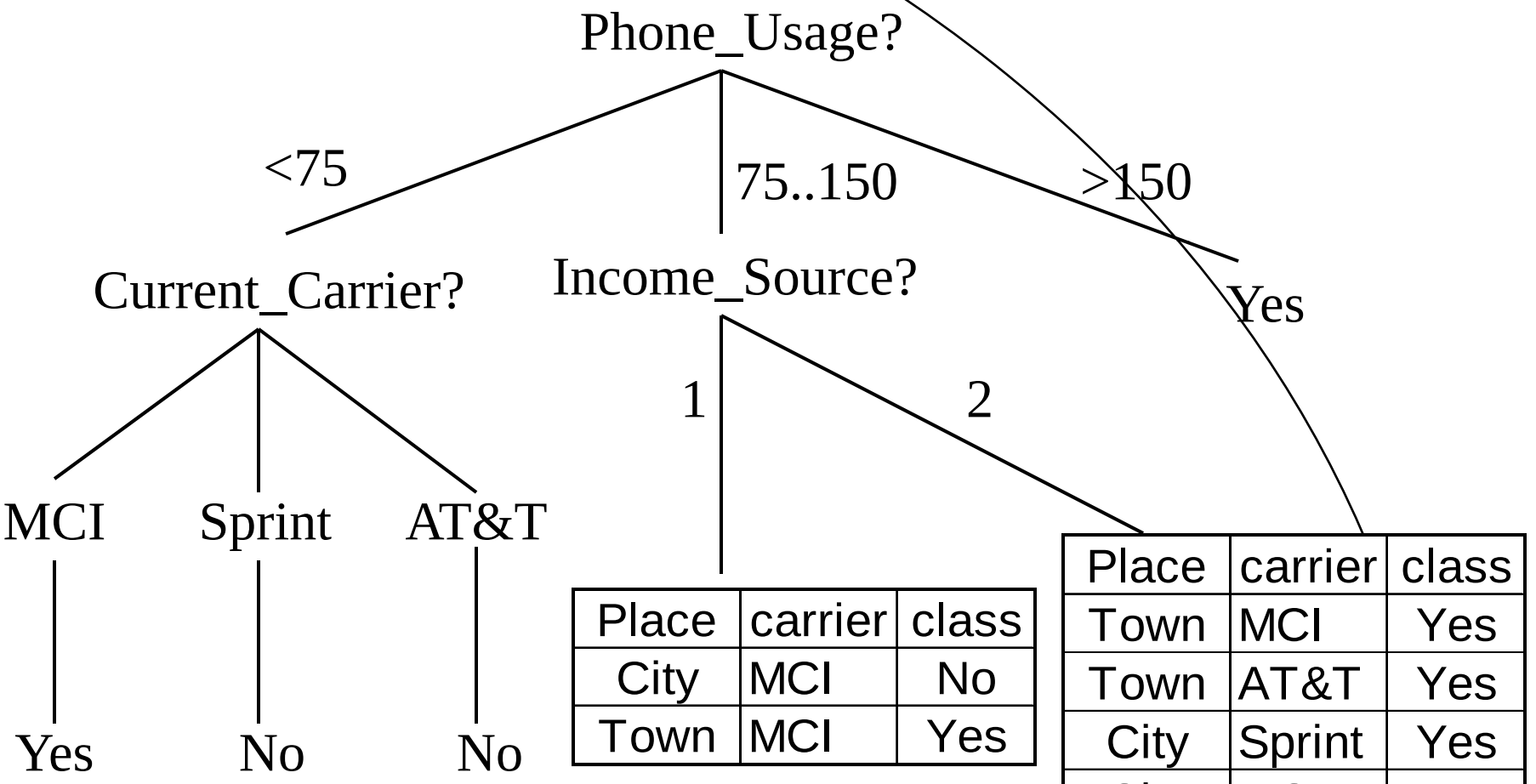
$$Gain\ (Current\ \_\ Carrier\ ) = 0.109$$

Partitioned Training Data Set and Partial Decision Tree



source	place	carrier	class
2	Town	MCI	Yes
1	City	MCI	No
2	Town	AT&T	Yes
1	Town	MCI	Yes
2	City	Sprint	Yes
2	City	MCI	Yes

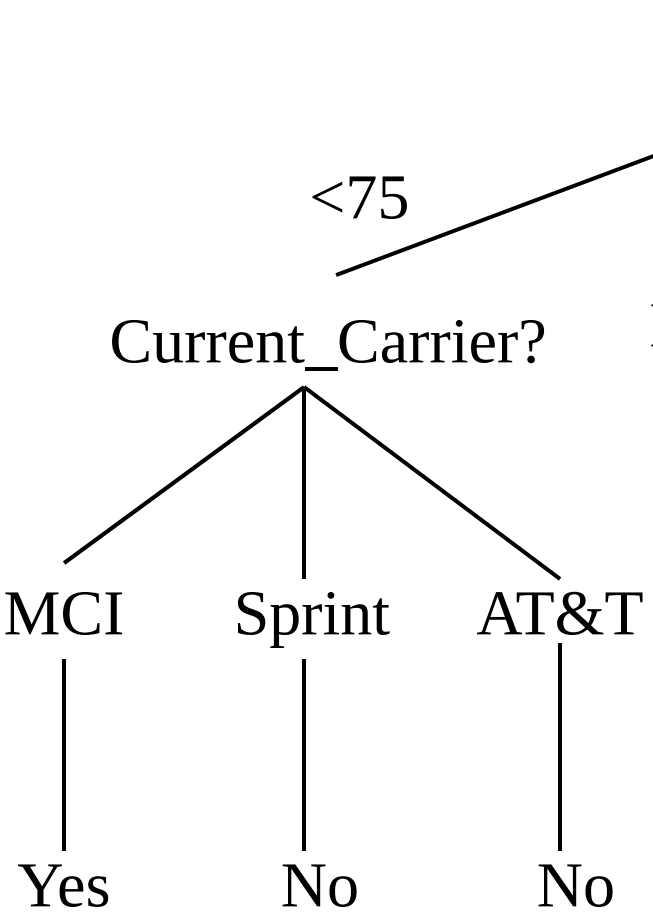
Partitioned Training Data Set and Partial Decision Tree



Place	carrier	class
City	MCI	No
Town	MCI	Yes

Place	carrier	class
Town	MCI	Yes
Town	AT&T	Yes
City	Sprint	Yes
City	MCI	Yes

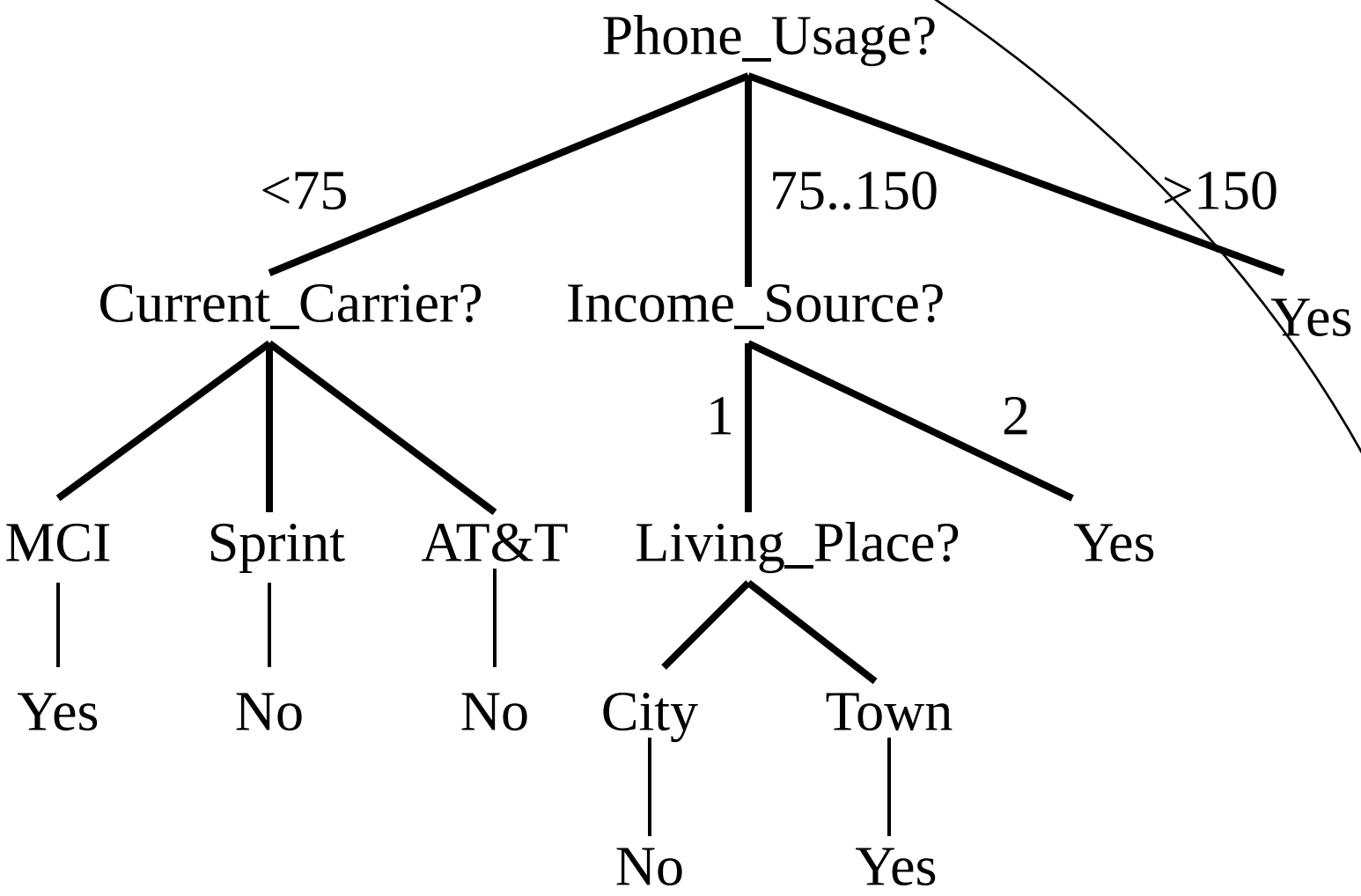
Partitioned Training Data Set and Partial Decision Tree



Place	carrier	class
City	MCI	No
Town	MCI	Yes

Place	carrier	class
Town	MCI	Yes
Town	AT&T	Yes
City	Sprint	Yes
City	MCI	Yes

Final Decision Tree



Extracting Classification Rules from Tree

IF usage = “<75” AND carrier = “MCI” THEN *change_plan* = “Yes”

IF usage = “<75” AND carrier = “Sprint” THEN *change_plan* = “No”

IF usage = “<75” AND carrier = “AT&T” THEN *change_plan* = “No”

IF usage = “75..150” AND source = “1” AND place = “City” THEN *change_plan* = “No”

IF usage = “75..150” AND source = “1” AND place = “Town” THEN *change_plan* = “Yes”

IF usage = “75..150” AND source = “2” THEN *change_plan* = “Yes”

IF usage = “>150” THEN *change_plan* = “Yes”

End of
Decision Tree Module (Part B)