Data Mining

Classification IV - Random Forests (Part B)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

/

Where am I?

- Part A explains how the random forests algorithm works.

- Part B presents an example to show how the algorithm grows a tree.

The whole training set

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

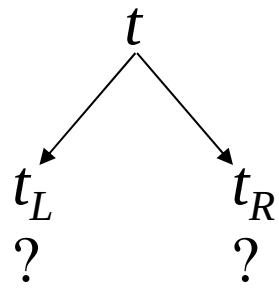
A tree growing example

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

The 6 records above are randomly picked and are used for growing a tree.
There are M = 4 attributes in the training set.

Randomly pick $\sqrt{M}$ = $\sqrt{4}$ = 2 attributes for splitting the root node t.
Assume Income_Source and Current_Carrier are picked for splitting t.

Income_Source	Current_Carrier	Change_Plan
1	AT&T	Yes
1	AT&T	No
2	Sprint	No
2	MCI	Yes
2	Sprint	Yes
1	MCI	Yes

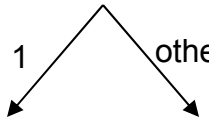


Since there are 4 records with “Yes” and 2 records with “No” in node *t*, the gini impurity measure for t is $1-(4/6)^2-(2/6)^2=0.444$.

There is only one possible split for Income_Source.
There are three possible splits for Current_Carrier.
We are going to find the best split among these four possible splits.

One possible split s_1 for Income_Source:

Income_Source	Change_Plan
1	Yes
1	No
2	No
2	Yes
2	Yes
1	Yes



Income_Source	Change_Plan
1	Yes
1	No
1	Yes

$g(t_L)$=1-(2/3)^2-(1/3)^2=4/9

Income_Source	Change_Plan
2	No
2	Yes
2	Yes

$g(t_R)$=1-(2/3)^2-(1/3)^2=4/9

$\Delta g(s_1,t)$=0.44-(3/6)(4/9)-(3/6)(4/9)=0

One possible split s_2 for Current_Carrier:

Current_Carrier	Change_Plan
AT&T	Yes
AT&T	No
Sprint	No
MCI	Yes
Sprint	Yes
MCI	Yes

AT&T otherwise

Current_Carrier	Change_Plan
AT&T	Yes
AT&T	No

$g(t_L)$=1-(1/2)^2-(1/2)^2=1/2

Current_Carrier	Change_Plan
Sprint	No
MCI	Yes
Sprint	Yes
MCI	Yes

$g(t_R)$=1-(3/4)^2-(1/4)^2=3/8

$\Delta g(s_2,t)$=0.44-(2/6)(1/2)-(4/6)(3/8)=0.028

One possible split s_3 for Current_Carrier:

Current_Carrier	Change_Plan
AT&T	Yes
AT&T	No
Sprint	No
MCI	Yes
Sprint	Yes
MCI	Yes

Sprint / otherwise

Current_Carrier	Change_Plan
Sprint	No
Sprint	Yes

$g(t_L)$=1-(1/2)^2-(1/2)^2=1/2

Current_Carrier	Change_Plan
AT&T	Yes
AT&T	No
MCI	Yes
MCI	Yes

$g(t_R)$=1-(3/4)^2-(1/4)^2=3/8

$\Delta g(s_3,t)$=0.44-(2/6)(1/2)-(4/6)(3/8)=0.028

One possible split s_4 for Current_Carrier:

Current_Carrier	Change_Plan
AT&T	Yes
AT&T	No
Sprint	No
MCI	Yes
Sprint	Yes
MCI	Yes

MCI otherwise

Current_Carrier	Change_Plan
MCI	Yes
MCI	Yes

$g(t_L)$=1-(2/2)^2-(0/2)^2=0

Current_Carrier	Change_Plan
AT&T	Yes
AT&T	No
Sprint	No
Sprint	Yes

$g(t_R)$=1-(2/4)^2-(2/4)^2=1/2

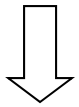
$\Delta g(s_4,t)$=0.44-(2/6)(0)-(4/6)(1/2)=0.11

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

$\Delta g(s_1,t)=0$
$\Delta g(s_2,t)=0.028$
$\Delta g(s_3,t)=0.028$
$\Delta g(s_4,t)=0.11$

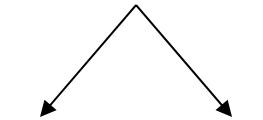
s_4 is the best split.

PU	IS	LP	CC	CP
75...150	2	City	MCI	Yes
75...150	1	Town	MCI	Yes



Yes

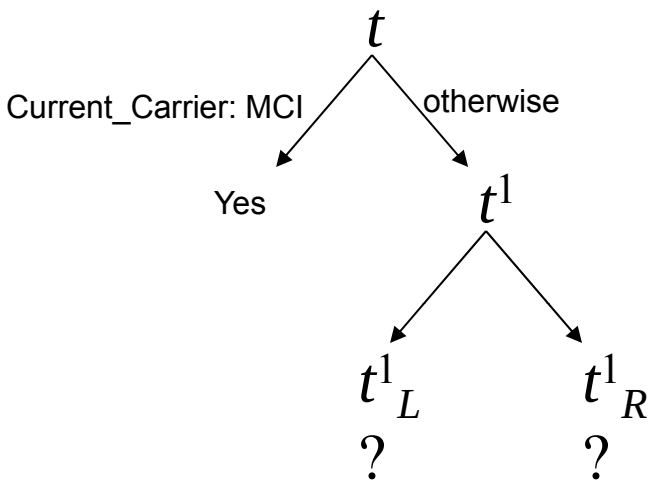
PU	IS	LP	CC	CP
>150	1	Town	AT&T	Yes
<75	1	Town	AT&T	No
<75	2	City	Sprint	No
75...150	2	City	Sprint	Yes



? ?

Randomly pick $\sqrt{M}$ = $\sqrt{4}$ = 2 attributes for splitting node t^1.
Assume Phone_Usage and Living_Place are picked for splitting t^1.

Phone_Usage	Living_Place	Change_Plan
>150	Town	Yes
<75	Town	No
<75	City	No
75...150	City	Yes

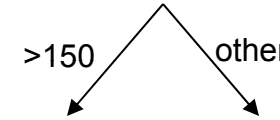


Since there are 2 records with “Yes” and 2 records with “No” in node t^1 , the gini impurity measure for t^1 is $1-(2/4)^2-(2/4)^2=0.5$.

There are three possible splits for Phone_Usage.
There is only one possible split for Living_Place.
We are going to find the best split among these four possible splits.

One possible split s^1_1 for Phone_Usage:

Phone_Usage	Change_Plan
>150	Yes
<75	No
<75	No
75...150	Yes



Phone_Usage	Change_Plan
>150	Yes

$g(t^1_L)=1-(1/1)^2-(0/1)^2=0$

Phone_Usage	Change_Plan
<75	No
<75	No
75...150	Yes

$g(t^1_R)=1-(1/3)^2-(2/3)^2=4/9$

$\Delta g(s^1_1,t^1)=0.5-(1/4)(0)-(3/4)(4/9)=0.167$

One possible split s^1_2 for Phone_Usage:

Phone_Usage	Change_Plan
>150	Yes
<75	No
<75	No
75...150	Yes

75…150 otherwise

Phone_Usage	Change_Plan
75…150	Yes

$g(t^1_L)=1-(1/1)^2-(0/1)^2=0$

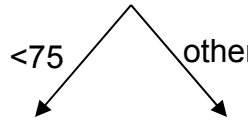
Phone_Usage	Change_Plan
>150	Yes
<75	No
<75	No

$g(t^1_R)=1-(1/3)^2-(2/3)^2=4/9$

$\Delta g(s^1_2,t^1)=0.5-(1/4)(0)-(3/4)(4/9)=0.167$

One possible split s^1_3 for Phone_Usage:

Phone_Usage	Change_Plan
>150	Yes
<75	No
<75	No
75...150	Yes



Phone_Usage	Change_Plan
<75	No
<75	No

$g(t^1_L)=1-(0/2)^2-(2/2)^2=0$

Phone_Usage	Change_Plan
>150	Yes
75…150	Yes

$g(t^1_R)=1-(2/2)^2-(0/2)^2=0$

$\Delta g(s^1_3,t^1)=0.5-(2/4)(0)-(2/4)(0)=0.5$

One possible split s^1_4 for Living_Place:

Living_Place	Change_Plan
Town	Yes
Town	No
City	No
City	Yes

Town otherwise

Living_Place	Change_Plan
Town	Yes
Town	No

Living_Place	Change_Plan
City	No
City	Yes

$g(t^1_L)$=1-(1/2)^2-(1/2)^2=1/2

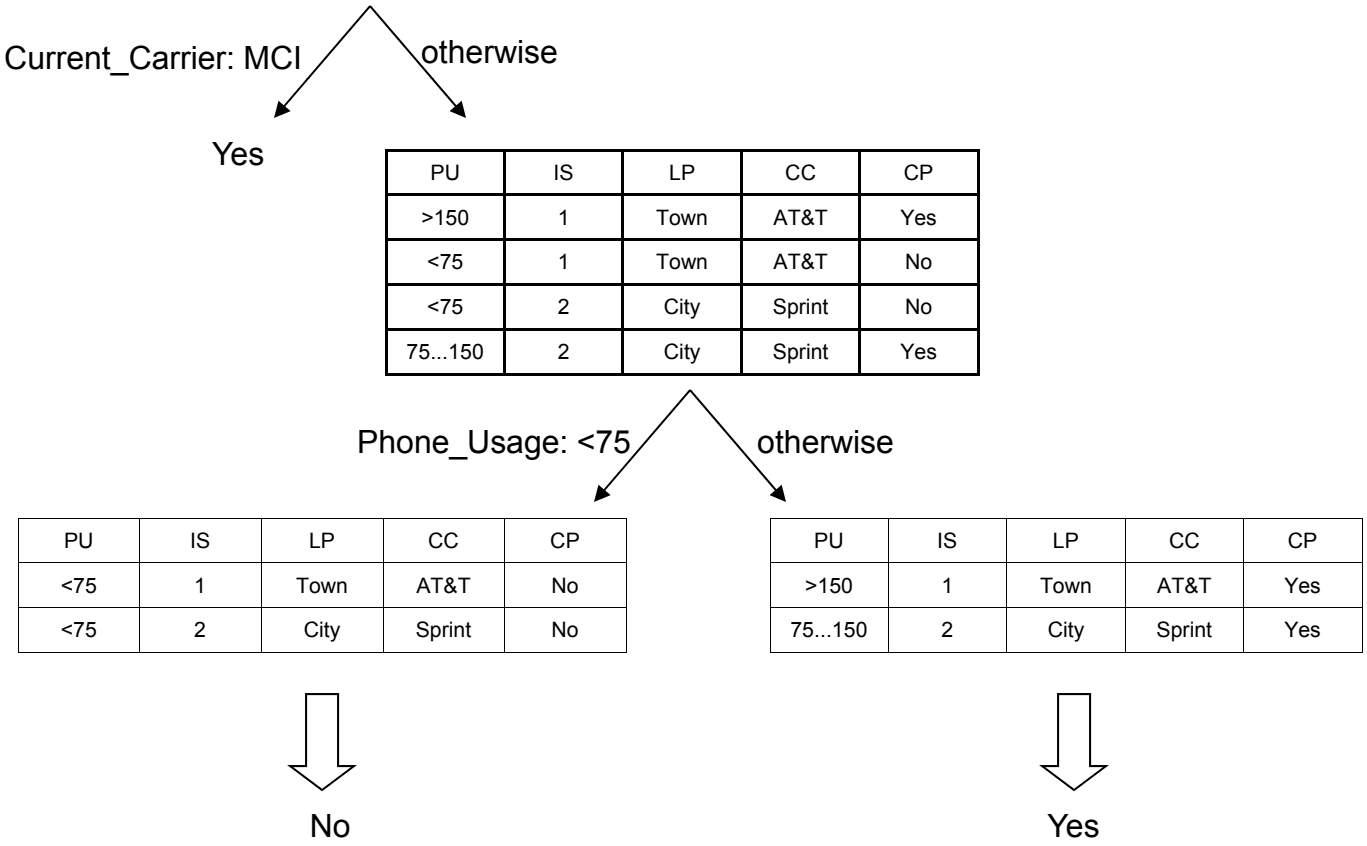
$g(t^1_R)$=1-(1/2)^2-(1/2)^2=1/2

$\Delta g(s^1_4, t^1)$=0.5-(2/4)(1/2)-(2/4)(1/2)=0

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

$\Delta g(s^1_1, t^1)=0.167$
$\Delta g(s^1_2, t^1)=0.167$
$\Delta g(s^1_3, t^1)=0.5$
$\Delta g(s^1_4, t^1)=0$

s^1_3 is the best split.



End of
Random Forests Module (Part B)