

- In Decision Tree induction we need to find the best attribute with which we can move down in best way.
- We have
 - Information gain (Already discussed in previous lecture)
 - Gini Index

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Introduction

- A way of attribute selection measure (Selects the best attribute).
- It is a measure of the impurity (inequality) of D.

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2,$$

P_i = count of specific class level / total count of D

$$Gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

- Attribute whose impurity is less will be selected.

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

It uses a binary split of each attribute.

Finds all possible subsets using all possible values.

– If attribute A have v possible values then there are 2^v possible subsets.

– Attribute -: Income

- Values-: {low , medium , high}

– Subsets-: $2^3 = 8$

= {low , medium , high} ,

{low , medium},

{low , high} , { medium , high} , {low , high} , {low} , {medium}, {high}, {}

RID	age	income	student	credit_rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.aged	medium	no	excellent	yes
13	middle.aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

If a binary split on income, partitions D into D1 and D2 the gini index of D given that partitioning is

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

$$\begin{aligned}
 &Gini_{income \in \{low, medium\}}(D) \\
 &= \frac{10}{14} Gini(D_1) + \frac{4}{14} Gini(D_2) \\
 &= \frac{10}{14} \left(1 - \left(\frac{6}{10} \right)^2 - \left(\frac{4}{10} \right)^2 \right) + \frac{4}{14} \left(1 - \left(\frac{1}{4} \right)^2 - \left(\frac{3}{4} \right)^2 \right)
 \end{aligned}$$

- Four Steps

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2,$$

1. Find the impurity of D ,using formula 1
2. Find the impurity of each resulting partition using formula 2.=

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2).$$

3. Find reduction in Impurity using formula

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

Whichever split best minimizes gini index in that attribute.

4. Now select the best attribute which gives the minimum gini index overall

Induction of Decision tree using Gini Index

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2,$$

- Step 1 -: Compute the impurity of D.
- Total tuples are 14
- 9 tuples belonging to Class buys_computer = yes
- 5 tuples belonging to class buys_computer=no
- Using formula 1

$$Gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Find the splitting criterion for tuples in D

- We need to compute the gini index of each attribute (age , income, credit_rating, student)

- Lets take income first

Now consider each possible splitting subsets

({low , medium}, {low , high} , {medium , high} , {low} , {medium}, {high})

- Lets take {low,medium} first
- Total tuples where income \in {low , medium} = 10 (D1)
- rest left =4 (D2)
- Now compute gini index based on this partitioning

$$\begin{aligned} &Gini_{income \in \{low, medium\}}(D) \\ &= \frac{10}{14} Gini(D_1) + \frac{4}{14} Gini(D_2) \\ &= \frac{10}{14} \left(1 - \left(\frac{6}{10} \right)^2 - \left(\frac{4}{10} \right)^2 \right) + \frac{4}{14} \left(1 - \left(\frac{1}{4} \right)^2 - \left(\frac{3}{4} \right)^2 \right) \end{aligned}$$

=.450

It will be same for
{high}

	Tuples in D1		Tuples in D2		Gini index
{low, medium} or {High}	Tuples in D1 {low,medium}		Tuples in D2 (high)		.450
	10		4		
	Buys_computer (yes)	Buys_computer (no)	Buys_computer (yes)	Buys_computer (no)	
	6	4	2	2	

	Tuples in D1		Tuples in D2		Gini index
{low,high} or {medium}	Tuples in D1 {low,high}		Tuples in D2 (medium)		.315
	8		6		
	Buys_computer(yes)	Buys_computer(no)	Buys_computer(yes)	Buys_computer(no)	
	6	5	4	2	

← ↗ ≡ ↘

	Tuples in D1		Tuples in D2		Gini index
{medium,high} or {low}	Tuples in D1 {medium,high}		Tuples in D2 (low)		.300
	10		4		
	Buys_computer(yes)	Buys_computer(no)	Buys_computer(yes)	Buys_computer(no)	
	6	4	3	1	

← ↗ ≡ ↘

- Best binary split for income is {medium,high} or {low} with minimum gini index.
- Now do the same for attribute age , student , and credit_rating.

Attribute	Split	Gini index	Reduction in impurity $\Delta G = \text{gini}(D) - \text{gini}_A(D)$
income	{medium,high} or {low}	.300	.459-.300 = .159
age	{youth_senior} or {middle aged}	.375	.459-.375 = .084
Student	Binary	.367	.459-.367 = .092
Credit_rating	binary	.429	.459-.429 = .03

Income is selected with minimum gini index and highest reduction in impurity

