Gini Index

✓ There are nine tuples belonging to the class buys_Computer = yes and the remaining five tuples belong to the class buys_Comp = no.

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

$$= 0.459$$

✓ To find the splitting criterian for the tuples in D, we need to compute Gini Index for each attribute

✓ let's start with the attribute income and Consider each of the possible splitting subset

Incomes

possible Values are {low, medium, high}

possible Subsets are

{low, medium, high}

{low, medium}

{low, high}

{medium, high}

{low},

{medium},

{high}

{ }

⟹ From this we can exclude powerset {low, medium, high} and empty set { } since Conceptually they do not represent split.

Consider the Subset

{low, medium}

Income	Yes	No	No. of Instances
high	2	2	4
medium	4	2	6
low	3	1	4
			14

$$Gini^{(D)}_{income \in \{low, medium\}}$$

$$= \frac{10}{14} Gini(D_1) + \frac{4}{14} Gini(D_2)$$

$$= \frac{10}{14}\left(1-\left(\frac{7}{10}\right)^2-\left(\frac{3}{10}\right)^2\right) + \frac{4}{14}\left(1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right)$$

$$= 0.443$$

$$= Gini^{(D)}_{income \in \{high\}}$$

$$G^{(D)}_{income \in \{low, high\}}$$

$$= \frac{8}{14}\left(\left(1-\left(\frac{5}{8}\right)^2-\left(\frac{3}{8}\right)^2\right)\right) + \frac{6}{14}\left(1-\left(\frac{4}{6}\right)^2-\left(\frac{2}{6}\right)^2\right)$$

$$= 0.458$$

$$= Gini^{D}_{income \in \{medium\}}$$

=

$$\text{Gini}_{\text{income} \in \{\text{medium, high}\}}(D)$$

$$= \frac{10}{14}\left(1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right)$$

$$= 0.450$$

$$= \text{Gini}_{\text{income} \in \{\text{low}\}}(D)$$

Gini for subsets

$\{\text{low, medium}\} = 0.443$ ✓

$\{\text{high}\} \neq \; = 0.443$ ✓

$\{\text{low, high}\} = 0.458$

$\{\text{medium}\} = 0.458$

$\{\text{medium, high}\} = 0.450$

$\{\text{low}\} = 0.450$

→ Best binary split for attribute income is on {low, medium} or {high} because it minimizes Gini Index

<u>Age</u>

Gini
age

proper subsets

{youth, middle_aged, senior}

{youth, middle_aged}

{youth, senior}

{middle_aged, senior}

{youth}, {middle_aged}, {senior}

{ }

Exclude powerset and nullset so we will get only 6

age	yes	no	no of Instances
youth	2	3	5
middleaged	4	0	4
Senior	3	2	5
			14

$gini_{age \in \{youth, middle\_aged\}}^{D}$

$$= \frac{9}{14}\left(1-\left(\frac{6}{9}\right)^2-\left(\frac{3}{9}\right)^2\right)+\frac{5}{14}\left(1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2\right)$$

$= 0.4571$

$= gini_{age \in \{senior\}}(D)$

$gini_{age \in \{youth, senior\}}^{D}$

$$= \frac{10}{14}\left(1-\left(\frac{5}{10}\right)^2-\left(\frac{5}{10}\right)^2\right)+\frac{4}{14}\left(1-\left(\frac{4}{4}\right)^2-\left(\frac{0}{4}\right)^2\right)$$

$= 0.3571$

$= gini_{age \in \{middle\_aged\}}^{D}$

$$\text{gini}_{age \in \{middle\_aged,\ senior\}}(D)$$

$$= \frac{9}{14}\left(1-\left(\frac{7}{9}\right)^2-\left(\frac{2}{9}\right)^2\right)+\frac{5}{14}\left(1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2\right)$$

$$= 0.3936$$

$$= \text{gini}_{age \in \{youth\}}(D)$$

✓ possible subsets

Gini

{youth, middle_aged}	= 0.4571
{senior}	= 0.4571
{youth, senior}	= 0.3571 ✓
{middle_aged}	= 0.3571 ✓
{middle_aged, senior}	= 0.3936
{youth}	= 0.3936

✓ We obtain {youth, senior} or {middle_aged} as the best split for age with a gini index of 0.357

Binary

Student

Values are {Yes, no}

Student	Yes	No	No of Instances
Yes	6	1	7
No	3	4	7

14

$$gini_{student}(D) = \frac{7}{14}\left(1-\left(\frac{6}{7}\right)^2-\left(\frac{1}{7}\right)^2\right)+ \left(\frac{7}{14}\right)\left(1-\left(\frac{3}{7}\right)^2-\left(\frac{4}{7}\right)^2\right)$$

$$= 0.367$$

Credit_rating

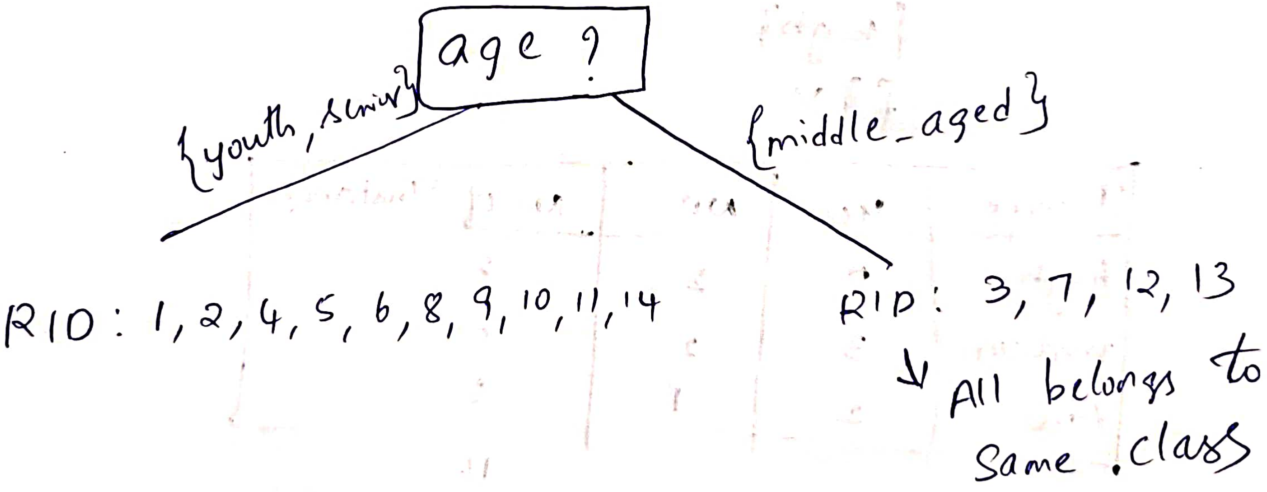
Values are {Fair, Excellent}

Credit_rating	Yes	No	No of Instances
Fair	6	2	8
Excellent	3	3	6

$$gini_{credit\_rating}(D) = \frac{8}{14}\left(1-\left(\frac{6}{8}\right)^2-\left(\frac{2}{8}\right)^2\right)+\frac{6}{14}\left(1-\left(\frac{3}{6}\right)^2-\left(\frac{3}{6}\right)^2\right)$$

$$= 0.428$$

Attribute	split	Gini Index	Reduction in impurity $\Delta gini = gini(D) - gini_A(D)$
age	{youth, senior} or {middle_aged}	0.3571	0.459 − 0.357 = 0.102
Income	{low, medium} or {high}	0.443	0.459 − 0.443 = 0.016
student	Binary	0.367	0.459 − 0.367 = 0.092
credit_rating	Binary	0.428	0.459 − 0.428 = 0.031


age ?
{youth, senior}
{middle_aged}
RID: 1, 2, 4, 5, 6, 8, 9, 10, 11, 14
RID: 3, 7, 12, 13
All belongs to same class