Gini Index (CART, IBM IntelligentMiner)

- If a data set D contains examples from n classes, gini index, $gini(D)$ is defined as

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

where p_j is the relative frequency of class j in D

- If a data set D is split on A into two subsets D_1 and D_2, the *gini* index $gini(D)$ is defined as

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

- Reduction in Impurity:

$$\Delta gini(A) = gini(D) - gini_A(D)$$

- The attribute provides the smallest $gini_{split}(D)$ (or the largest reduction in impurity) is chosen to split the node
- The Gini index considers a binary split for each attribute

Binary Splitting using GINI Index

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Age has 3 values, so

- Possible splits are:
  1. {<=30,31-40},{>40}
  2. {<=30, >40},{31-40}
  3. {31-40, >40},{<=30}

$Gini_{\{<=30,31\text{-}40\},\{>40\}} = 9/14\{1-(6/9)^2-(3/9)^2\} + 5/14\{1-(3/5)^2-(2/5)^2\}$

$= 0.48$

$Gini_{\{<=30,>40\},\{31\text{-}40\}} = 10/14\{1-(5/10)^2-(5/10)^2\} + 4/14\{1-(4/4)^2-0\}$

$= 0.357$

$Gini_{\{31\text{-}40,>40\},\{<=30\}} = 9/14\{1-(7/9)^2-(2/9)^2\} + 5/14\{1-(2/5)^2-(3/5)^2\}$

$= 0.39$

- Thus, amongst the three possible splits of age, the gini of {<=30, >40},{31-40} is the lowest.
- So, now there are two possible outcomes of age.
- Let middle_aged = 31-40

  youth = <=30

  senior = >40
- Now the two possible outcomes of age are
  1. {y,s}
  2. {m_a}

age	income	student	credit_rating	buys_computer
{y,s}	high	no	fair	no
{y,s}	high	no	excellent	no
{m}	high	no	fair	yes
{y,s}	medium	no	fair	yes
{y,s}	low	yes	fair	yes
{y,s}	low	yes	excellent	no
{m}	low	yes	excellent	yes
{y,s}	medium	no	fair	no
{y,s}	low	yes	fair	yes
{y,s}	medium	yes	fair	yes
{y,s}	medium	yes	excellent	yes
{m}	medium	no	excellent	yes
{m}	high	yes	fair	yes
{y,s}	medium	no	excellent	no

- Binary splitting for income as income has three possible values:
  1. Low
  2. Medium
  3. High

age	income	student	credit_rating	buys_computer
{y,s}	high	no	fair	no
{y,s}	high	no	excellent	no
{m}	high	no	fair	yes
{y,s}	medium	no	fair	yes
{y,s}	low	yes	fair	yes
{y,s}	low	yes	excellent	no
{m}	low	yes	excellent	yes
{y,s}	medium	no	fair	no
{y,s}	low	yes	fair	yes
{y,s}	medium	yes	fair	yes
{y,s}	medium	yes	excellent	yes
{m}	medium	no	excellent	yes
{m}	high	yes	fair	yes
{y,s}	medium	no	excellent	no

- $Gini_{\{low,high\}.\{medium\}} = 0.458$
- $Gini_{\{low,medium\}.\{high\}} = 0.443$
- $Gini_{\{medium,high\}.\{low\}} = 0.45$

Computation of Gini Index

- Ex. D has 9 tuples in buys_computer = "yes" and 5 in "no"

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

- Gini $_{student}$ = 7/14{1-(6/7)^2-(1/7)^2}+7/14{1-(3/7)^2-(4/7)^2}=0.367
- Gini $_{CR}$ = 8/14{1-(6/8)^2-(2/8)^2}+6/14{1-(3/6)^2-(3/6)^2}=0.429
- Gini$_{income}$ = 0.443
- Gini $_{age}$ = 0.358

age	income	student	credit_rating	buys_computer
{y,s}	high	no	fair	no
{y,s}	high	no	excellent	no
{m}	high	no	fair	yes
{y,s}	medium	no	fair	yes
{y,s}	low	yes	fair	yes
{y,s}	low	yes	excellent	no
{m}	low	yes	excellent	yes
{y,s}	medium	no	fair	no
{y,s}	low	yes	fair	yes
{y,s}	medium	yes	fair	yes
{y,s}	medium	yes	excellent	yes
{m}	medium	no	excellent	yes
{m}	high	yes	fair	yes
{y,s}	medium	no	excellent	no

22

Computing Information-Gain for Continuous-Valued Attributes

- Let attribute A be a continuous-valued attribute
- Must determine the *best split point* for A
  - Sort the value A in increasing order
  - Typically, the midpoint between each pair of adjacent values is considered as a possible *split point*
    - $(a_i+a_{i+1})/2$ is the midpoint between the values of a_i and a_{i+1}
  - The point with the *minimum expected information requirement* for A is selected as the split-point for A
- Split:
  - D1 is the set of tuples in D satisfying A ≤ split-point, and D2 is the set of tuples in D satisfying A > split-point

23

A	Class
100	No
75	No
90	Yes
95	Yes
60	No
120	No
220	No
70	No
85	Yes
125	No

Sorted Values →
Split Positions →

class	No	No	No	Yes	Yes	Yes	No	No	No	No
Taxable Income										
	60	70	75	85	90	95	100	120	125	220
	55	65	72	80	87	92	97	110	122	172	230
	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >
Yes	0 3	0 3	0 3	0 3	1 2	2 1	3 0	3 0	3 0	3 0	3 0
No	0 7	1 6	2 5	3 4	3 4	3 4	3 4	4 3	5 2	6 1	7 0
Gini	0.420	0.400	0.375	0.343	0.417	0.400	0.300	0.343	0.375	0.400	0.420

7 --> no
3 --> yes

A	Class
>97	No
<=97	No
<=97	Yes
<=97	Yes
<=97	No
>97	No
>97	No
<=97	No
<=97	Yes
>97	No

24

Gain Ratio for Attribute Selection (C4.5)

- Information gain measure is biased towards attributes with a large number of values
- C4.5 (a successor of ID3) uses gain ratio to overcome the problem (normalization to information gain)

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

  - GainRatio(A) = Gain(A)/SplitInfo(A)
- Ex.

$$SplitInfo_{income}(D) = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right) = 1.557$$

  - gain_ratio(income) = 0.029/1.557 = 0.019
- The attribute with the maximum gain ratio is selected as the splitting attribute

25

Comparing Attribute Selection Measures

- The three measures, in general, return good results but
  - **Information gain**:
    - biased towards multivalued attributes
  - **Gain ratio**:
    - tends to prefer unbalanced splits in which one partition is much smaller than the others
  - **Gini index**:
    - biased to multivalued attributes
    - has difficulty when # of classes is large
    - tends to favor tests that result in equal-sized partitions and purity in both partitions

Practice with GINI index

Attribute 1	Attribute 2	Attribute 3	Class
A	70	T	C1
A	90	T	C2
A	85	F	C2
A	95	F	C2
A	70	F	C1
B	90	T	C1
B	78	F	C1
B	65	T	C1
B	75	F	C1
C	80	T	C2
C	70	T	C2
C	80	F	C1
C	80	F	C1
C	96	F	C1

Practice with Gain Ratio

Height	Hair	Eyes	Class
Short	Blond	Blue	+
Tall	Blond	Brown	-
Tall	Red	Blue	+
Short	Dark	Blue	-
Tall	Dark	Blue	-
Tall	Blond	Blue	+
Tall	Dark	Brown	-
Short	Blond	Brown	-