1. **Information gain**

ID3 uses information gain as its attribute selection measure. Let node *N* represents or hold the tuples of partition *D*. The attribute with the highest information gain is chosen as the splitting attribute for node *N*.

*Info(D)* is just the average amount of information needed to identify the class label of a tuples in *D*. The expected information needed to classify a tuples in *D* is given by

 (1)

How much more information would we still need (after the partitioning) in order to arrive at an exact classification? This amount is measured by

**** (2)

Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on *A*). That is,

 (3)

**EXAMPLE:**



**Question**: Find the information gain for all attributes

**AGE Attribute**

First find *Info (D)* using formula 1



Find Info for particular Attribute using formula 2



Gain of age using formula 3

Gain (age) =Info (D)-Infoage (D) =0.940-0.694=0.246

**INCOME Attribute**

First find *Info (D)* using formula 1



Find Info for particular Attribute using formula 2



Gain of income using formula 3

Gain (income) =Info (D)-Infoincome (D) =0.940-0.911=0.029

**STUDENT Attribute**

First find *Info (D)* using formula 1



Find Info for particular Attribute using formula 2



Gain of student using formula 3

Gain (student) =Info (D)-Infostudent (D) =0.940-0.789=0.151

**Credit Attribute**



Find Info for particular Attribute using formula 2



Gain of credit rating using formula 3

Gain (credit rating) =Info (D)-Infocredit (D) =0.940-0.892=0.048

**Answer:**

**Information gain for**

Age=0.246

Income=0.029

Student=0.151

Credit=0.048

1. **Gain Ratio**

C4.5, a successor of ID3, uses an extension to information gain known as *gain ratio*, which attempts to overcome this bias. It applies a kind of normalization to information gain using a “split information” value defined analogously with *Info*(*D*) as

 (4)

Gain ratio is defined as

 (5)

**Question**: Find the Gain Ratio for all attributes

**AGE Attribute**

Find split-info for age attribute using formula 4



Find gain ratio using formula 5



**Income Attribute**

Find split-info for income attribute using formula 4



Find gain ratio using formula 5



**Student Attribute**

Find split-info for student attribute using formula 4



Find gain ratio using formula 5



**Credit Attribute**

Find split-info for credit attribute using formula 4



Find gain ratio using formula 5



**Answer:**

**Gain ratio for**

Age=0.155

Income=0.019

Student=0.151

Credit=0.0487

1. **Gain Ratio**

Gini index measures the impurity of *D*, a data partition or set of training tuples, as

 (6)

Gini of particular attribute

 (7)

 (8)

Question: Gini index for the entire attribute

**Income Attribute {low, medium}**

First find *Gini (D)* using formula 6



Find Info for particular Attribute using formula 7



Gini index of income using formula 8

Gini index (income) =Gini (D)-Giniincome(D) =0.459-0.443=0.016

**Age (youth, senior}**

First find *Gini (D)* using formula 6



Find Info for particular Attribute using formula 7



Gini index of income using formula 8

Gini index (income) =Gini (D)-Giniage(D) =0.459-0.357=0.102

**student Attribute**

First find *Gini (D)* using formula 6



Find Info for particular Attribute using formula 7



Gini index of student using formula 8

Gini index (student) =Gini (D)-Ginistudent (D) =0.459-0.367=0.092

**student Attribute**

First find *Gini (D)* using formula 6



Find Info for particular Attribute using formula 7



Gini index of income using formula 8

Gini index (income) =Gini (D)-Gini credit(D) =0.459-0.429=0.03