Pattern Recognition

- S. S. Samant

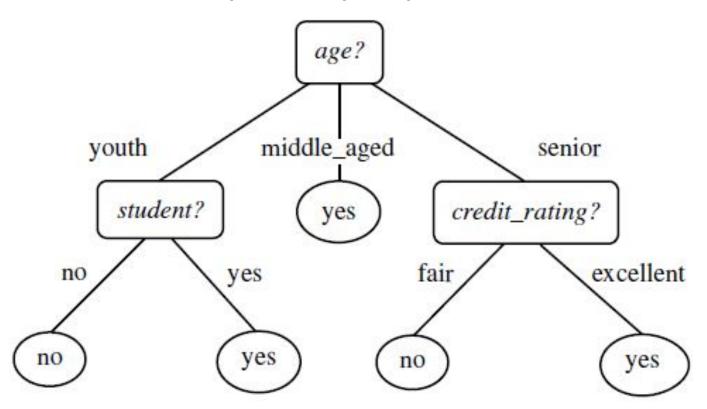


These are *multistage* decision systems in which classes are sequentially rejected until we reach a finally accepted class. To this end, the feature space is split into unique regions, corresponding to the classes, *in a sequential manner*.



These are *multistage* decision systems in which classes are sequentially rejected until we reach a finally accepted class. To this end, the feature space is split into unique regions, corresponding to the classes, *in a sequential manner*.

Will a person buy computer?





These are *multistage* decision systems in which classes are sequentially rejected until we reach a finally accepted class. To this end, the feature space is split into unique regions, corresponding to the classes, *in a sequential manner*.

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

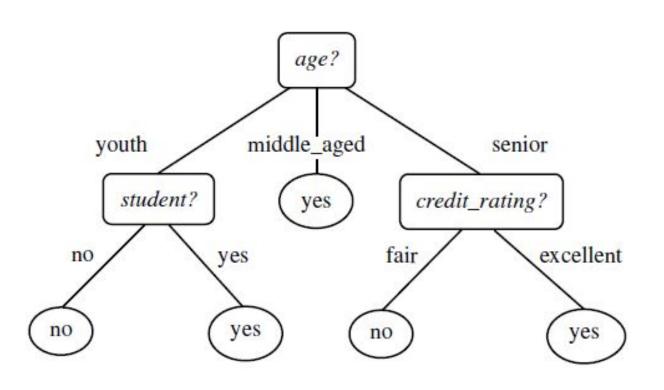


These are *multistage* decision systems in which classes are sequentially rejected until we reach a finally accepted class. To this end, the feature space is split into unique regions, corresponding to the classes, *in a sequential manner*.

- A *splitting criterion* must be adopted according to which the best split from the set of candidate ones is chosen.
- A stop-splitting rule is required that controls the growth of the tree, and a node is declared as a terminal one (*leaf*).
- A rule is required that assigns each leaf to a specific class.

Splitting Criteria (Attribute Selection Measure)







Information Gain also called The *decrease in node impurity*

Expected information needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

where p_i is the probability that an arbitrary tuple in D belongs to class $C_i = \frac{|C_{i,D}|}{|D|}$

If we select attribute A for partition, then the information we still need for classification is given by:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

Attribute A can be used to split D into ν partitions or subsets

$$Gain(A) = Info(D) - Info_A(D)$$

RID	age	income	student	credit_rating	Class: buys_compute	er
1	youth	high	no	fair	no	_
2	youth	high	no	excellent	no	
3	middle_aged	high	no	fair	yes	0 0 5
4	senior	medium	no	fair	yes	$Info(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$
5	senior	low	yes	fair	yes	$Injo(D) = -\frac{1}{14}\log_2(\frac{1}{14}) - \frac{1}{14}\log_2(\frac{1}{14})$
6	senior	low	yes	excellent	no	
7	middle_aged	low	yes	excellent	yes	
8	youth	medium	no	fair	no	m
9	youth	low	yes	fair	yes	$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$
10	senior	medium	yes	fair	yes	$Injo(D) = -\sum p_i \log_2(p)$
11	youth	medium	yes	excellent	yes	i=1
12	middle_aged	medium	no	excellent	yes	
13	middle_aged	high	yes	fair	yes	
14	senior	medium	no	excellent	no	



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 I
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

$$Info(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

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 In
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

What if we split on age?

high

low

medium

medium

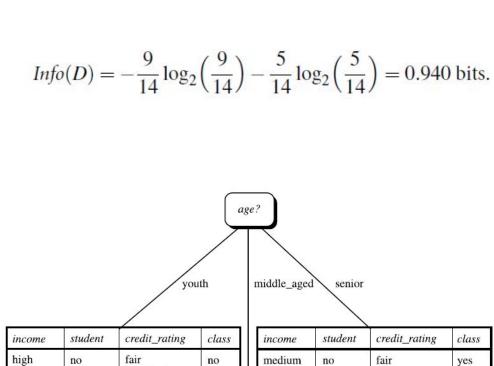
yes

excellent

excellent

fair

fair



income	student	credit_rating	class
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

low

medium

medium

no

no

yes

yes

fair

fair

excellent

excellent

yes

no

yes

no

yes

yes

yes



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 I
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

$$nfo(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

$$Info_{age}(D) = \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)$$

$$+\frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right)$$

$$+\frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$$

$$= 0.694 \text{ bits.}$$

$$\mathit{Info}_A(D) = \sum_{j=1}^v rac{|D_j|}{|D|} imes \mathit{Info}(D_j)$$

ctudent

credit rating

RID



age	income	student	creatt_rating	Class: buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent G	Gain(age) = Info
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no
•	youth youth middle_aged senior senior senior middle_aged youth youth senior youth middle_aged middle_aged	youth high youth high middle_aged high senior low senior low middle_aged low youth medium youth low senior medium youth medium youth medium middle_aged medium middle_aged medium middle_aged high	youth high no youth high no middle_aged high no senior medium no senior low yes senior low yes middle_aged low yes youth medium no youth low yes senior medium yes senior medium yes middle_aged medium yes middle_aged medium no middle_aged high yes	youth high no fair youth high no excellent middle_aged high no fair senior medium no fair senior low yes fair senior low yes excellent middle_aged low yes excellent youth medium no fair youth low yes fair senior medium yes fair youth wedium yes fair youth medium yes fair youth medium yes fair middle_aged medium no excellent middle_aged high yes fair

Class hive combuter

$$Info(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

$$\begin{split} Info_{age}(D) &= \frac{5}{14} \times (-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}) \\ &+ \frac{4}{14} \times (-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}) \\ &+ \frac{5}{14} \times (-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}) \\ &= 0.694 \text{ bits.} \end{split}$$

$$\mathit{Info}_A(D) = \sum_{j=1}^{v} rac{|D_j|}{|D|} imes \mathit{Info}(D_j)$$



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

$$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246$$
 bits.

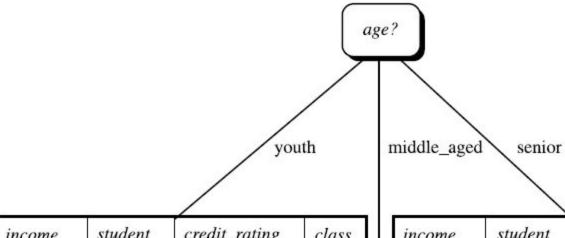
Gain(income) = 0.029 bits

$$Gain(student) = 0.151$$
 bits

 $Gain(credit_rating) = 0.048$ bits.

$$\mathit{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \mathit{Info}(D_j)$$





income	student	credit_rating	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	student	credit_rating	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	student	credit_rating	class
high low medium high	no yes no yes	fair excellent excellent fair	yes yes yes



Thank You!