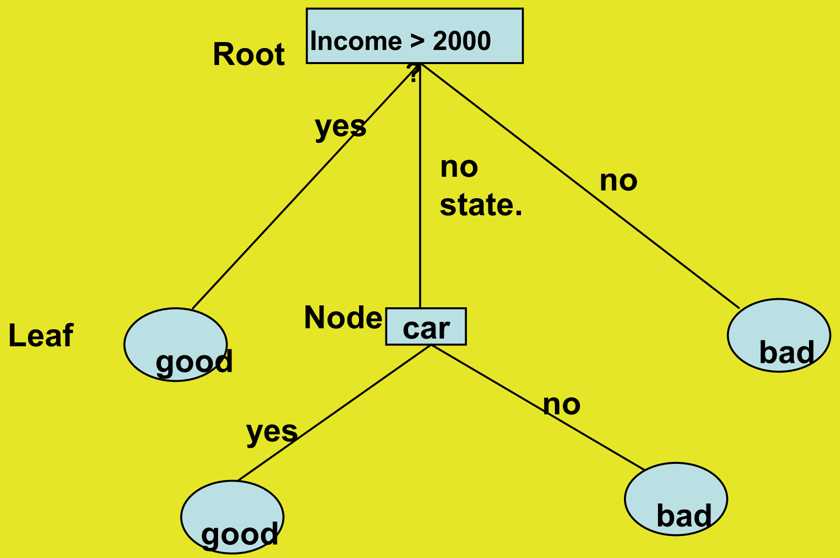
Statistical Methods in Data Mining



Decision Trees

Professor Dr. Gholamreza Nakhaeizadeh

content

Decision Trees

- Introduction
- Example: Credit Rating
- Example: Computer buyers
- Attribute selection measure in Decision Trees
- Construction of Decision Trees
- Gain Ratio
- Gini Index
- Overfitting
- Pruning

Decision Trees (DT)

Introduction

- DT are classification tools
- Class Variable (Target Variable): Nominal
- Attributes: Nominal or continuous-valued
- Top Down construction based on heuristic methods by using training data (Greedy instead completely search : tends to find good solutions quickly, but not always optimal ones)

Simple fictive example; Credit Rating in a Bank

	Income >2000	Car	Gender	Credit Rating
Customer 1	no	yes	F	bad
Customer 2	no statement	no	F	bad
Customer 3	no statement	yes	M	good
Customer 4	no	yes	M	bad
Customer 5	yes	yes	M	good
Customer 6	yes	yes	F	good
Customer 7	no statement	yes	F	good
Customer 8	yes	no	F	good
Customer 9	no statement	no	M	bad
Customer 10	no	no	F	bad

Simple fictive example; Credit Rating in a Bank

	Income >2000	Car	Gender	Credit Rating
Customer 1	no	yes	F	bad
Customer 2	no statement	no	F	bad
Customer 3	no statement	yes	M	good
Customer 4	no	yes	M	bad
Customer 5	yes	yes	M	good
Customer 6	yes	yes	F	good
Customer 7	no statement	yes	F	good
Customer 8	yes	no	F	good
Customer 9	no statement	no	M	bad
Customer 10	no	no	F	bad

Simple fictive example; Credit Rating in a Bank

	Income >2000	Car	Gender	Credit Rating
Customer 1	no	yes	F	bad
Customer 2	no statement	no	F	bad
Customer 3	no statement	yes	M	good
Customer 4	no	yes	M	bad
Customer 5	yes	yes	M	good
Customer 6	yes	yes	F	good
Customer 7	no statement	yes	F	good
Customer 8	yes	no	F	good
Customer 9	no statement	no	M	bad
Customer 10	no	no	F	bad

Simple fictive example; Credit Rating in a Bank

	Income >2000	Car	Gender	Credit Rating
Customer 1	no	yes	F	bad
Customer 2	no statement	no	F	bad
Customer 3	no statement	yes	M	good
Customer 4	no	yes	M	bad
Customer 5	yes	yes	M	good
Customer 6	yes	yes	F	good
Customer 7	no statement	yes	F	good
Customer 8	yes	no	F	good
Customer 9	no statement	no	M	bad
Customer 10	no	no	F	bad

Simple fictive example; Credit Rating in a Bank

	Income >2000	Car	Gender	Credit Rating
Customer 1	no	yes	F	bad
Customer 2	no statement	no	F	bad
Customer 3	no statement	yes	M	good
Customer 4	no	yes	M	bad
Customer 5	yes	yes	M	good
Customer 6	yes	yes	F	good
Customer 7	no statement	yes	F	good
Customer 8	yes	no	F	good
Customer 9	no statement	no	M	bad
Customer 10	no	no	F	bad

Simple fictive example; Credit Rating in a Bank

	Income >2000	Car	Gender	Credit Rating
Customer 1	no	yes	F	bad
Customer 2	no statement	no	F	bad
Customer 3	no statement	yes	M	good
Customer 4	no	yes	M	bad
Customer 5	yes	yes	M	good
Customer 6	yes	yes	F	good
Customer 7	no statement	yes	F	good
Customer 8	yes	no	F	good
Customer 9	no statement	no	M	bad
Customer 10	no	no	F	bad

Simple fictive example; Credit Rating in a Bank

Classifier

If income> 2000 = yes Credit Rate=good

If income> 2000 = no Credit Rate=bad

If income= no statement & car=yes Credit Rate=good

If income= no statement & car=no Credit Rate=bad

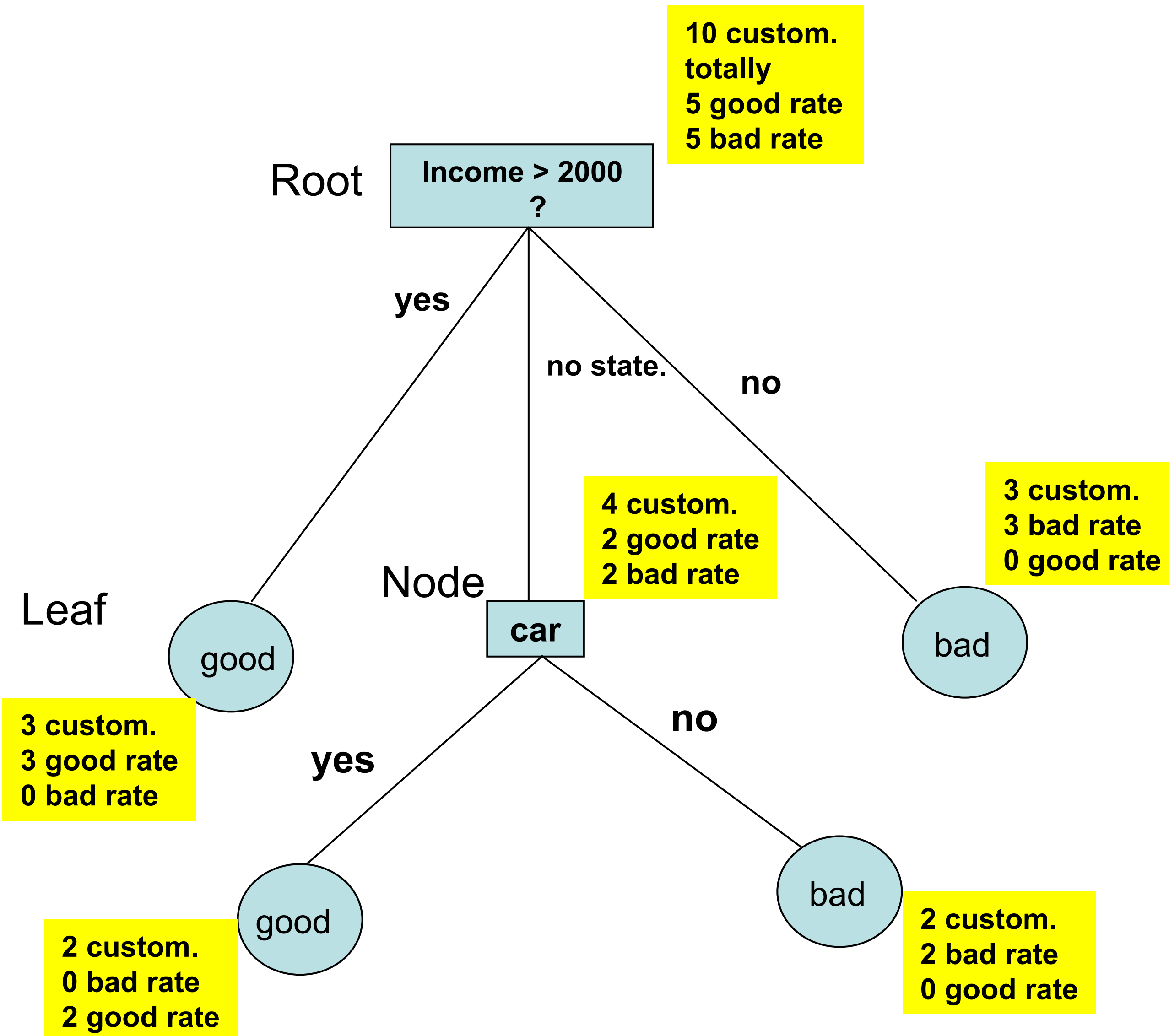
This classifier can be regarded as an Inductive expert systems

	Income >2000	Car	Gender	Credit Rating
Customer 1	no	yes	F	bad
Customer 2	no statement	no	F	bad
Customer 3	no statement	yes	M	good
Customer 4	no	yes	M	bad
Customer 5	yes	yes	M	good
Customer 6	yes	yes	F	good
Customer 7	no statement	yes	F	good
Customer 8	yes	no	F	good
Customer 9	no statement	no	M	bad
Customer 10	no	no	F	bad

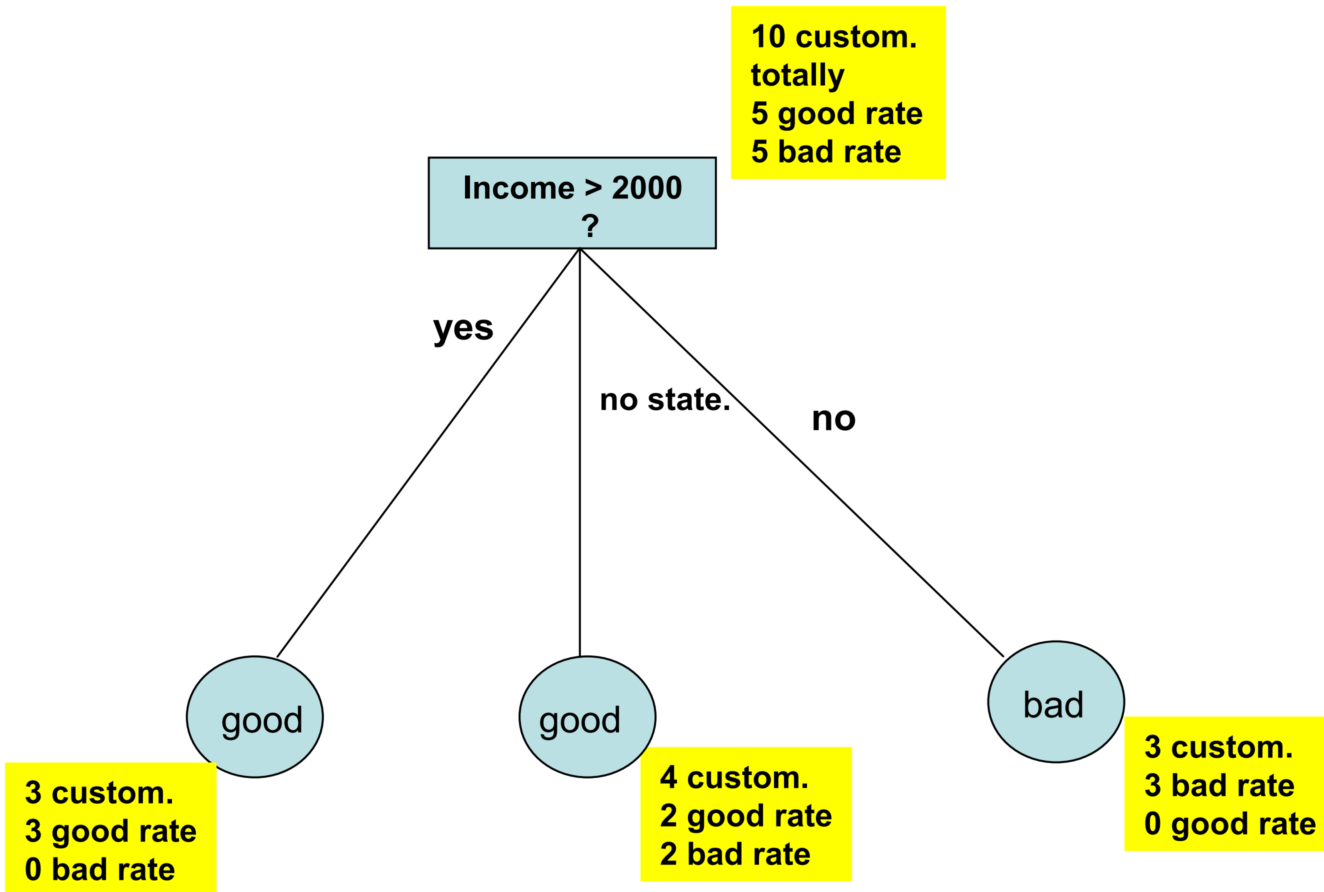
Rating new Customers

- Rating a new customer with income 3000 = good
- Rating a new customer who has no car and made no income statement = bad
- .....

Credit rating: decision tree construction



Credit rating: pruned decision tree



Perhaps due to background knowledge of credit officer

- Clementine Demo

Credit_toy2.str

- Clementine Demo

German-credit1.str

Example: Computer buyers

- Source: Jiawei Han, et al. 2006

Nr.	Age	Income	Student?	Credit Rating	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

Table:1
Computer buyers

Attribute selection measure in Decision Trees

information Gain (IG)

Introduction

IG is based on information theory due to Shannon

Example 1 :
Finding a certain number between 1 and 1000 by asking question

1. Alternative

Choose randomly a number between 1 and 1000 and ask whether it is the right one. No optimal method, because in the worst case 999 questions are needed to find the right number.

In this case, if the first answer is “no”, the IG has been very little. Because, there are still 999 alternative numbers between them the number we are looking for is.

Attribute selection measure in Decision Trees

information Gain (IG) (continues)

Second alternative

The first question should be: is the number ≤ 500?

The IG of this question is too higher because after the answer we have to search between 500 numbers instead of 1000.

If the answer of this question is positive, the next question is, as it may be expected: Is the number ≤ 250? and so on .

In this example, IG of each new question is equal to the amount of information one gains by asking this question.

Higher the IG of a question (attribute) → quicker to reach the goal.

Attribute selection measure in Decision Trees

information Gain (IG) (continues)

Definition of Entropy

Y: a random variable and $\mathbf{P(Y=b_1)=p_1, \ldots\ldots\ldots\ldots P(Y=b_m)=p_m}$

Entropy of Y:

$$I(Y) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \cdots\cdots\cdots\cdots - p_m \log_2 p_m = -\sum_{i=1}^{m} p_i \log_2 p_i \quad (1)$$

I(Y) is the expected information needed to find a certain value in distribution of Y

Attribute selection measure in Decision Trees

information Gain (IG) (continues)

Remark1

Definition (1) is due to Shannon in conjunction with information theory and aims to find the number of needed bits to communicate a messages (for this reason the base of used logarithm is 2)

Remark 2

(In example for m=1000 we get pi = 1/1000) and

I (Y) in (1) is equal nearly to 10 which is the average number of the

question that one needs to find a certain number

between 1 and 1000

Attribute selection measure in Decision Trees

information Gain (IG) (continues)

Example 2: Computer buyers

- Two classes: Buys Computer (yes or no)
- C1: class 1 (yes), C2: class 2 (no)
- N1: Numbers of tuples in C1= 9
- N2: Numbers of tuples in C2 = 5
- N=N1+N2=14
- p1: probability that a tuple belongs to C1
- p2: probability that a tuple belongs to C2
- Probability should be approximated by the portions

•Thus: p1=N1/N = 9/14 and p2= N2/N = 5/14

Nr.	Buys Computer
1	no
2	no
3	yes
4	yes
5	yes
6	no
7	yes
8	no
9	yes
10	yes
11	yes
12	yes
13	yes
14	no

Using relation **(1)** results to

$$I(Y) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.94$$

This is the Expected information (entropy) needed to classify a tuple.

Attribute selection measure in Decision Trees

information Gain (IG) (continues)

Conditional Entropy

Example: X Income Y Football Fan

Using relation **(1)** results to:
I(X) = 1.5 and I(Y) = 1
Moreover:
P (Football Fan = yes) = 0.5
P (Income= high) = 0.5
P (Income high and Football Fan = no) = 0.25
P (Football Fan = yes | Income = medium) = 0

X	Y
high	yes
medium	no
low	yes
high	no
high	no
low	yes
medium	no
high	yes

Table 2 : Football Fan

Entropy of Y for X = b: I (Y | X = b)

Using this definition and **(1)** leads to:

$$\left.\begin{aligned} I (Y \mid X = high) &= 1 \\ I (Y \mid X = medium) &= 0 \\ I (Y \mid X = low) &= 0 \end{aligned}\right\} \quad (2)$$

Attribute selection measure in Decision Trees

information Gain (IG) (continues) **Conditional Entropy**

Generally:

$$I(Y|X) = \sum_{i} p\,(X = b_i)\ I(\,Y \mid X = b_i\,) \qquad \textbf{(3)}$$

Called average conditional entropy

X= bi	P(X=bi)	I(Y| X=bi)
high	0.5	1
medium	0.25	0
low	0.25	0

From **Table 2** and relation **(2)** we can get:
And from this table:
$I(Y|X) = 0.5*1+0.25*0+0.25*0 = 0.5$

we have seen already $I(Y) = 1$ and now by using the values of X we have got $I(Y|X) = 0.5$ $\longrightarrow$ the needed information reduced to half and we have got $1- 0.5 = 0.5$ **"Information Gain"**

Attribute selection measure in Decision Trees

information Gain (IG) (continues)

Generally:

IG(Y|X) = I(Y) – I (Y|X) (4)

(4) called information gained by using X

Inserting (3) in (4) leads to:

$$IG(Y|X) = I(Y) - \sum_i p_i \, (X = bi)\, I(Y \mid X = bi) \qquad (5)$$

In the relation (5), like before p (x=bi) can be approximated by Ni/N, where Ni is the frequently of the value xi in X.

(5) Is one of the measures that has been used for attribute selection in Decision trees. The Decision Tree algorithms ID3 e.g. uses this measure

Construction of Decision Trees

Root selection: using the attribute with highest information gain

Example: Computer Buyers

In the following we show the target variable (computer Buyers) with Y and the attributes (Age, Income..) with X

Now we calculate IG (Y) regarding attribute **age**

$$IG_{age}(Y) = I(Y) - I(Y|\,age) \quad \textbf{(6)}$$

with

I(Y| age) = p (youth)* I (Y | age=youth) +
p (middle aged)* I (Y | age=middle aged) +
p (senior)* I (Y | age=senior)

we have seen already that I(Y)=0,94 and for Attribute age we have
p (youth) = 5/14, p (senior)= 5/14 and
p (middle aged) = 4/14

(7)

X (Age, Income, Student?, Credit Rating) — Y (Buys Computer)

Nr.	Age	Income	Student?	Credit Rating	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

Root selection: continues

On the other hand :

p (Y=no | age=youth)=3/5

p (Y= yes | age = youth) = 2/5

It means.

I (Y | age=youth) = **(8)**

-3/5 Log (-3/5) -2/5 Log (-2/5)= 0,968

From (7) and (8) we get:

P(youth)* I (Y | age= youth) = 5/14 * 0.968 = 0,346

In the same way we can calculate the other components of (6) :

$IG_{age}(Y) = I(Y) – I(Y \mid age) = 0.246$

and for the other attributes:

$IG_{income}(Y) = 0.029$

$IG_{student}(Y) = 0.151$

$IG_{Credit\ Rating}(Y) = 0.048$

Nr.	Age	Income	Student?	Credit Rating	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

IG of the attribute *age* is at the highest; Splitting of the DT starts by using this attribute as the root and its values (senior, middle aged, and youth) as the first branches of the tree

Construction of Decision Trees

splitting

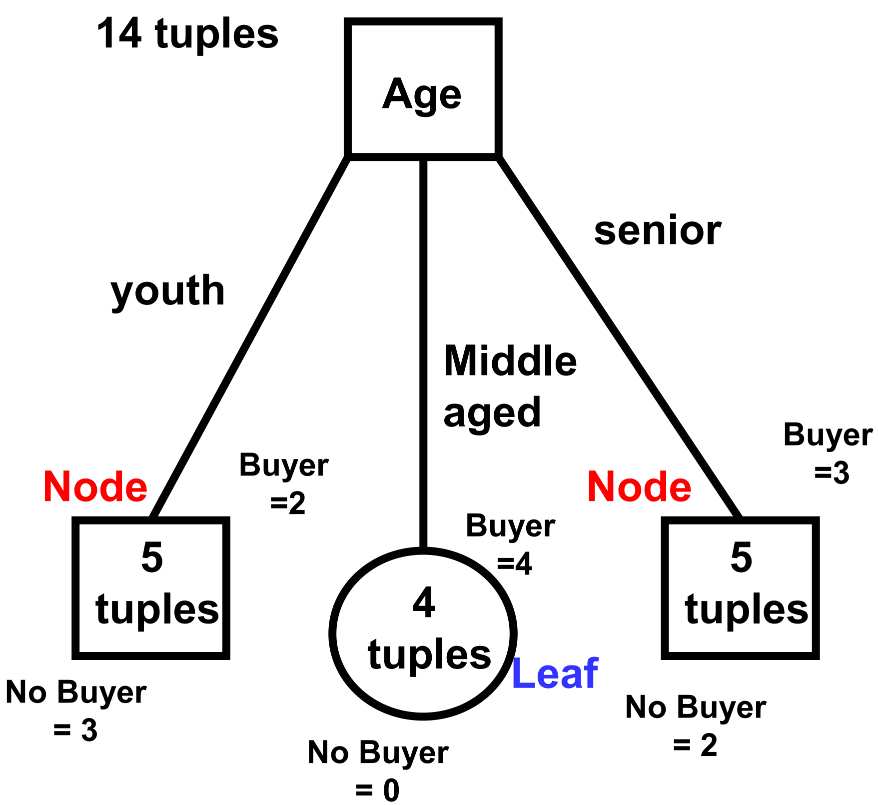
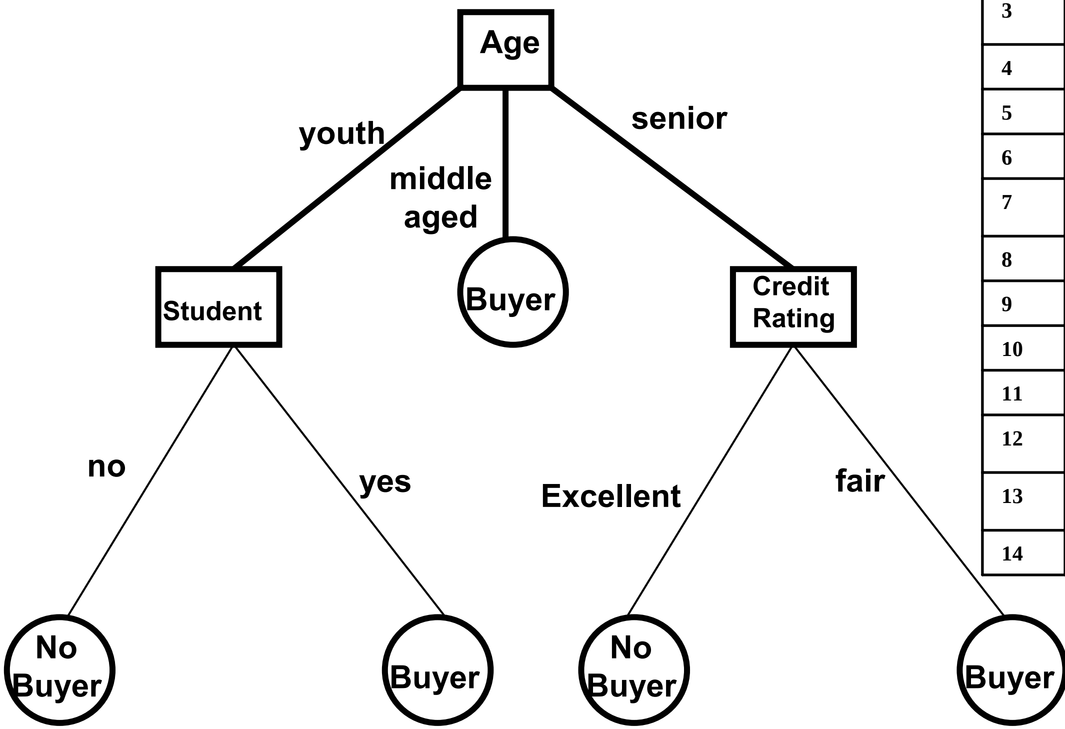


Fig 1, first splitting of DC

Nr.	Age	Income	Student?	Credit Rating	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

Construction of Decision Trees

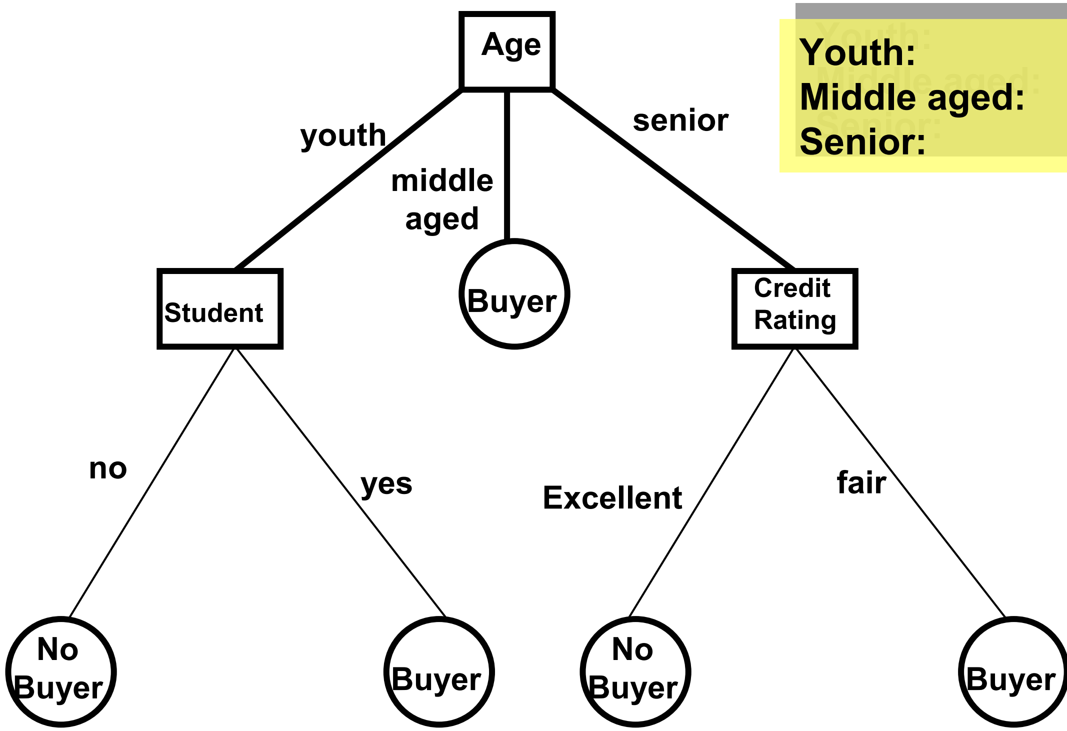
Splitting (continues)



Nr.	Age	Income	Student?	Credit Rating	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

Notice: Attribute income wasn't use

Decision Trees as classifiers for new tuples



Youth:	<= 30
Middle aged:	30=< <=40
Senior:	> 40

1. Age 37 = buyer
2. Age 55 with excellent credit rating = no buyer
3. Age 18 but no student = no buyer
4. …..

Construction of Decision Trees

Splitting of continuous - valued attributes

Example: monthly income of 10 individuals

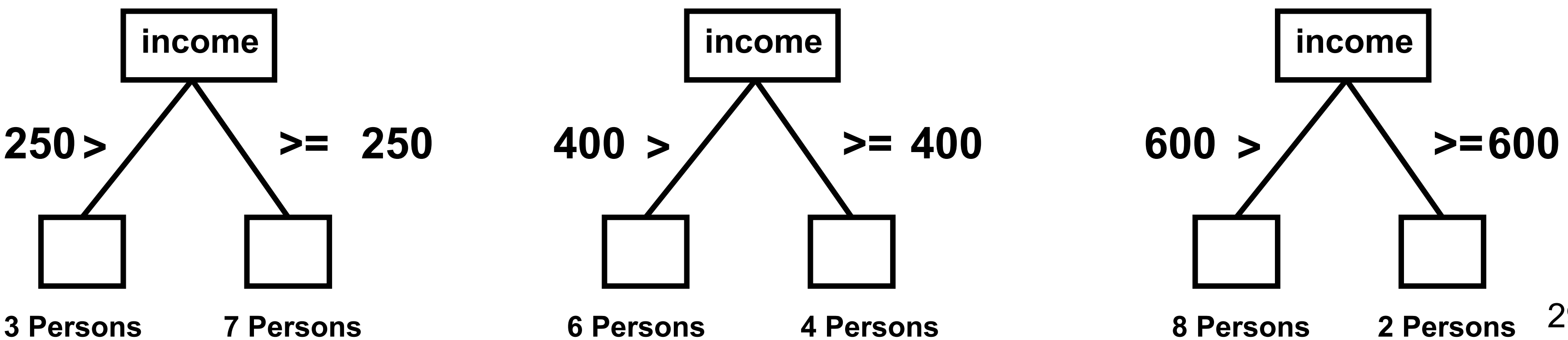
Income	700	300	200	200	300	200	500	300	700	500

Step 1: Sort the attribute values increasing and calculate the average of each two neighbors as possible threshold

Attribute value	200	300	500	700
Average	**250**	**400**	**600**	

Step 2: Split the node using the averages as alternative thresholds

Step 3: Calculate for each split the IG (income) and choose the one with highest IG

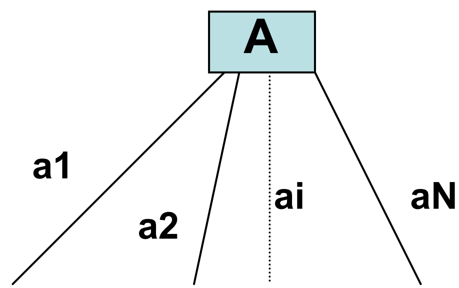


Attribute selection measure in Decision Trees

Other selection measure Gain Ratio

IG has a significant drawback:
it does not take into account the number of attribute values

Suppose that we have just N tuples and between the attributes we have a discrete-valued attribute A with values a1, a2,…ai,.. aj,….aN with ai ≠ aj for ∀ i and j In this case we would have by splitting using A, so many partitions as tuples namely N:



For this case $I(Y \mid A) = -\sum_{i}^{N} 1/N \; Log\,(Y \mid a_i) = 0$; **Regarding:** $IG(A) = I(Y) - I(Y|A)$

means A would have the maximal IG .
This extreme case shows very well that the IG prefers selection attributes with a large number of partitions.

Attribute selection measure in Decision Trees

Other selection measure

Gain Ratio (continues)

To overcome this problem Quinlan suggests for C4.5 (extension of ID3 algorithm) using Gain Ratio instead of Information Gain.

Gain Ratio is defined as: GR(A) = IG (A)/I (A)

with

$$I(A) = - \sum_{i}^{n} n_i/N * \log_2 (n_i/N)$$

n_i/N is the portion of the tuples with attribute value a_i

Nr.	Age	Income	Student?	Credit Rating	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

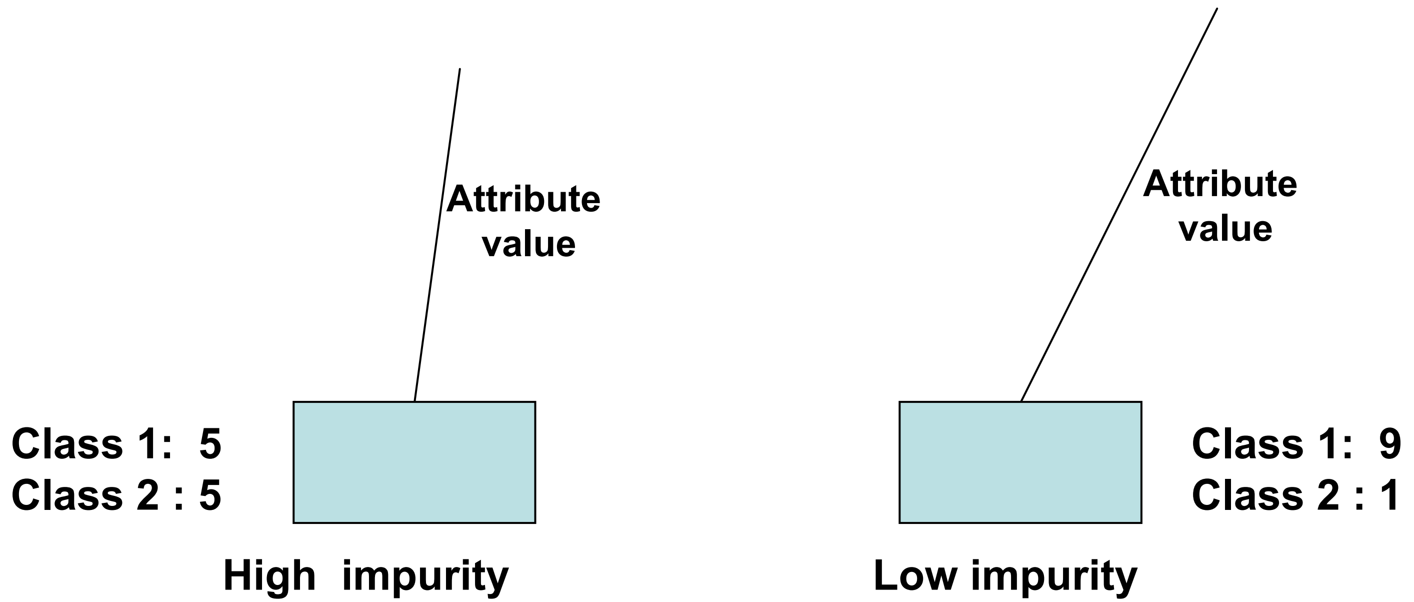
a_1 = high $n_1/N = 4/14$
a_2 = medium $n_2/N = 6/14$
a_3 = low $n_3/N = 4/14$

$\rightarrow$ I (A) = - 4/14 * Log 4/14 – 6/14 * Log 6/14 – 4/14 Log 4/14 = 0.926

we had calculated IG (income) = 0. 029 thus : GR(A) = 0.029/0.926 = 0.031

Attribute selection measure in Decision Trees

Gini Index (Gini) — **Impurity in the nodes**



GINI Index: Measure of Impurity

Attribute selection measure in Decision Trees

Gini Index (Gini) | **Gini Index of a node**

n = Number of tuples at the node k

n_j = Number of the tuples belong to the class j at the node k

n_j / n = relative frequently of the class $_j$ at the node k

$$\text{Gini}\,(k) = 1 - \sum_j \left(n_j / n|_k\right)^2$$

For two classes:

$n_1 = 0 \quad n_2 = n$ → Gini = 0 → lowest impurity

$n_1/n = ½ \quad n_2/n = ½$ → Gini = 1/2 → highest impurity

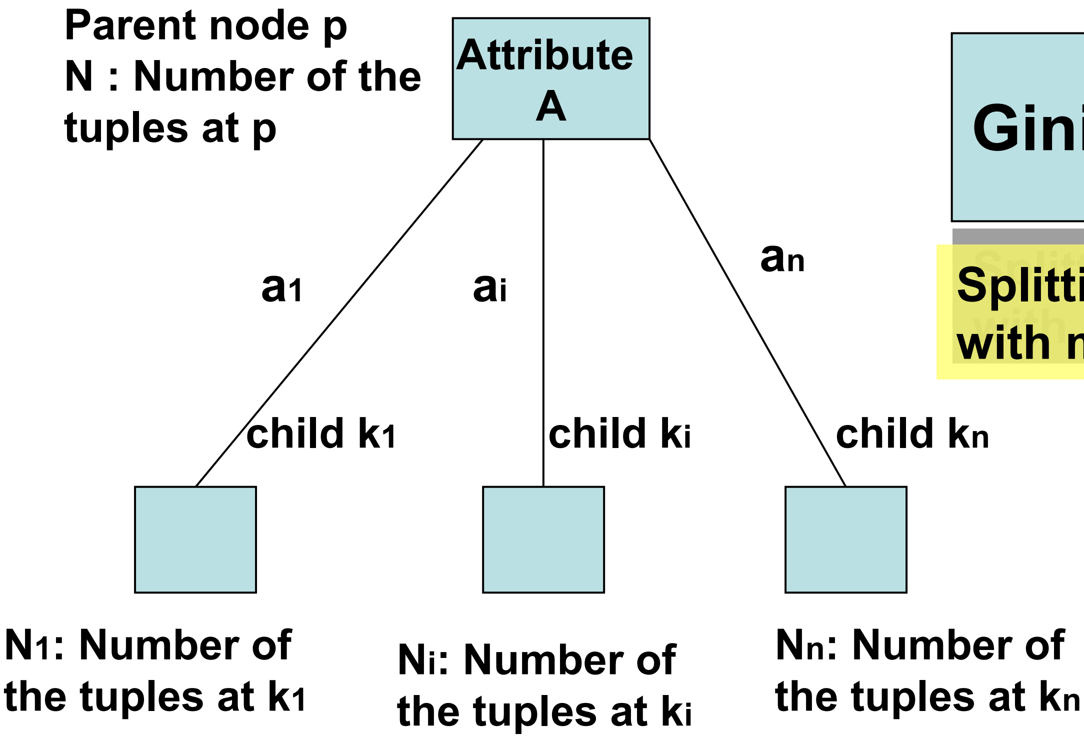
Further examples:

$n_1/n=1/6 \quad n_2/n=5/6 \qquad \text{Gini} = 1 - (1/6)^2 - (5/6)^2 = 0.278$

Attribute selection measure in Decision Trees

Gini Index

Gini Index as attribute selection measure



$$\text{Gini}(A) = \sum_{i}^{n} (N_i/N)\ \text{Gini}(k_i)$$

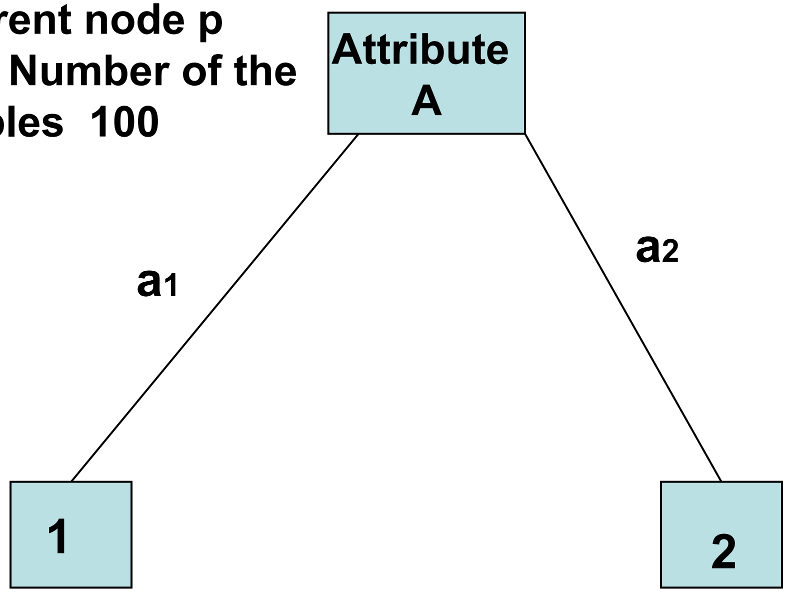
Splitting will be done for the attribute with minimal GINI

Some of DT algorithms among them CART use Gini Index

Attribute selection measure in Decision Trees

Gini Index | **Example**

Parent node p
N : Number of the tuples 100



$N_1 = 40$
$C1 = 20,\ C2 = 20$

$N_2 = 60$
$C1 = 10\ \ C2 = 50$

$N_1 / N = 40/100 = 0.4$

$N_2 / N = 60/100 = 0.6$

$$\text{Gini}(A) = \sum_{i}^{n} (N_i/N)\ \text{Gini}(k_i)$$

$\text{Gini}(A) = 0.4 * 0.5 + 0.6 * 0.278$
$= 0.367$

$$\text{Gini}(k) = 1 - \sum_{j} (n_j / n|\ k)^2 \longrightarrow$$

$\text{Gini}(1) = 1-(20/40)^2 - (20/40)^2 = 0,5$

$\text{Gini}(2) = 1-(10/60)^2 - (50/60)^2 = 0,278$

Construction of Decision Trees

Overfitting

Overfitting means: DT can classifies the training data with a relative high accuracy rate but not the test data. It means the DC is not able to generalize

Solution: Tree Pruning

- Pre-pruning
- Post-pruning

- Pre-pruning: Stop growing of the tree in the early stages

  Stop Criteria:
  - at pure nodes or nodes with high degree of purity
  - small number of tuples at a node
  - no more improving of accuracy rate by more growing

  ……

Construction of Decision Trees

Overfitting

Post-pruning :

- **Produce a full grown tree**
- **Prune this tree in different depths to produce a set of pruned trees**
- **Select the best one using a “validation” data set**

Weakness and Strength of Decision Trees

- **Strength**
  - Produce understandable classification rules with reasonable accuracy rates
  - Decision trees can be constructed relatively fast
  - Decision trees indicate clearly which attributes are most important for classification

- **Weakness**
  - By using of decision trees only descriptive analysis of data is possible
  - Discretization of continuous-valued is necessary
  - They are not appropriate for time series analysis and prediction

Short Review

Part Four: Decision Trees

- Introduction
- Example: Credit Rating
- Example: Computer buyers
- Attribute selection measure in Decision Trees
- Construction of Decision Trees
- Gain Ratio
- Gini Index
- Overfitting
- Pruning