TUTORIAL

CLASSIFICATION

CZ4032 Data Analytics and Mining
(Data Mining)

Q1

- Given the following training data of building materials, train a decision tree model using the entropy-based impurity measure at a node t, i.e.,

$$Entropy(t) = -\sum_j p(j|t)\log_2 p(j|t)$$

where p(j | t) is the relative frequency of class j at node t.

- Use the trained decision tree model to determine if a material instance (Size="small", Color="green", Shape= "wedge") is appropriate to be used for construction?

- each distinct value of a categorical value will become a child node at splitting.

Id	Size	Color	Shape	Can be used?
1	Medium	blue	brick	Yes
2	Small	red	sphere	Yes
3	Large	green	pillar	Yes
4	Large	green	sphere	Yes
5	Small	red	wedge	No
6	Large	red	wedge	No
7	Large	red	pillar	No

Q1 Answer

- **Information at the root**
  - I(3,4)= $-\frac{3}{7}\cdot\log_2\frac{3}{7} - \frac{4}{7}\cdot\log_2\frac{4}{7} = 0.985$
- **Which attribute to select for splitting?**
- **Attribute Size:** Medium (1), Small (2), Large (4)
- Medium: I(1,0)=0
- Small: I(1,1)=1
- Large: I(2,2)=1
- E(Size)= $\frac{1}{7}\cdot 0 + \frac{2}{7}\cdot 1 + \frac{4}{7}\cdot 1 = \frac{6}{7} = 0.857$
- **Gain(Size) = I(3,4) - E(Size) = 0.985-0.857 = 0.128**

Q1 Answer

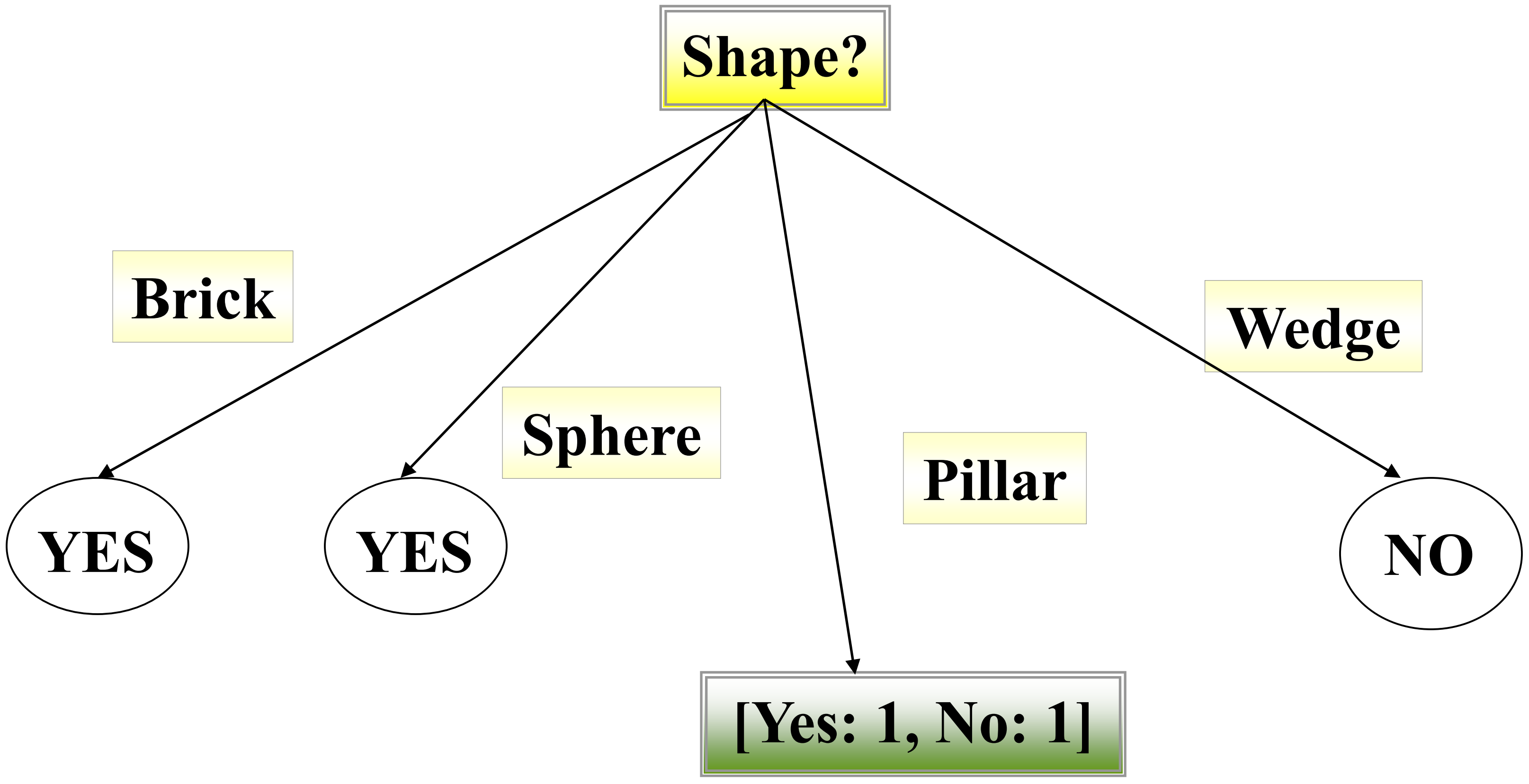
- **Attribute Color:** Blue (1), Green (2), Red (4)
- Blue: I(1,0)=0
- Green: I(2,0) = 0
- Red: I(1,3) $= -\frac{3}{4}\cdot\log_2\frac{3}{4} - \frac{1}{4}\cdot\log_2\frac{1}{4} = 0.811$
- E(Color) $= \frac{1}{7}\cdot 0 + \frac{2}{7}\cdot 0 + \frac{4}{7}\cdot 0.811 = 0.463$
- **Gain(Color)=I(3,4) – E(Color) = 0.985-0.463=0.522**

Q1 Answer

- **Attribute Shape:** Brick (1), Sphere (2), Pillar (2), Wedge (2)
- Brick: I(1,0)=0
- Sphere: I(2,0) = 0
- Pillar: I(1,1) =1
- Wedge: I(2,0) = 0
- E(Shape)= $\frac{1}{7}\cdot 0+\frac{2}{7}\cdot 0+\frac{2}{7}\cdot 1+\frac{2}{7}\cdot 0=0.286$
- **Gain(Shape)=I(3,4) – E(Shape)= 0.985-0.286=0.699**

Therefore, among the three attributes, **Shape** has the maximal information gain!

- Select Shape



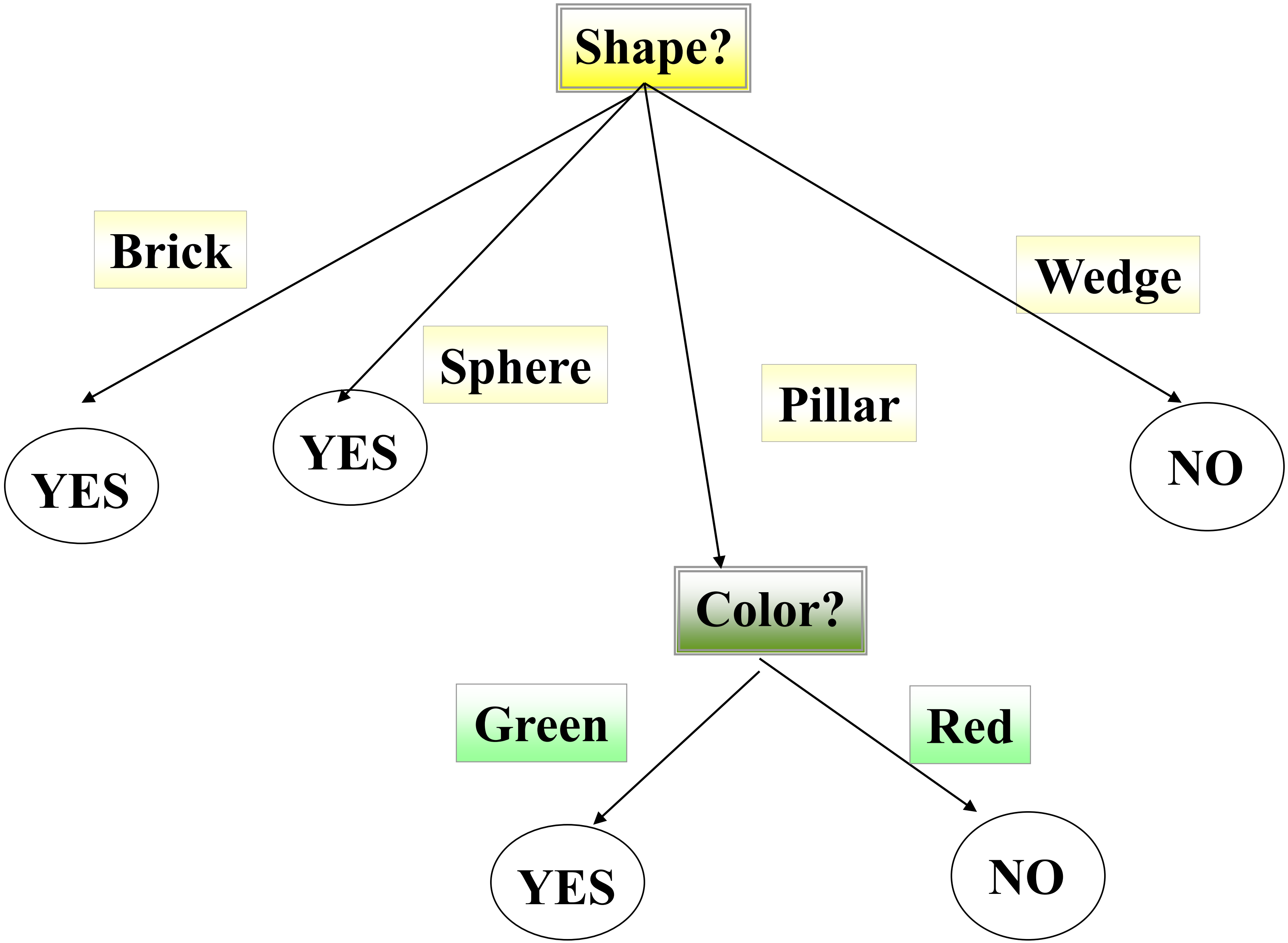
Id	Size	Color	Shape	Can be Used?
3	Large	Green	Pillar	Yes
7	Large	Red	Pillar	No

- **Information at the non-leaf node:** I(1,1)= 1

- **Attribute Size:** Large(2)
- Large: I(1,1)=1
- E(Size) = $\frac{2}{2} \cdot 1 = 1$
- **Gain(Size) = 1-1=0**

- **Attribute Color:** Green(1), Red(1)
- Green: I(1,0)=0
- Red: I(1,0) = 0
- E(Color) = $\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0 = 0$
- **Gain(Color) = 1-0=1**

- Select Color

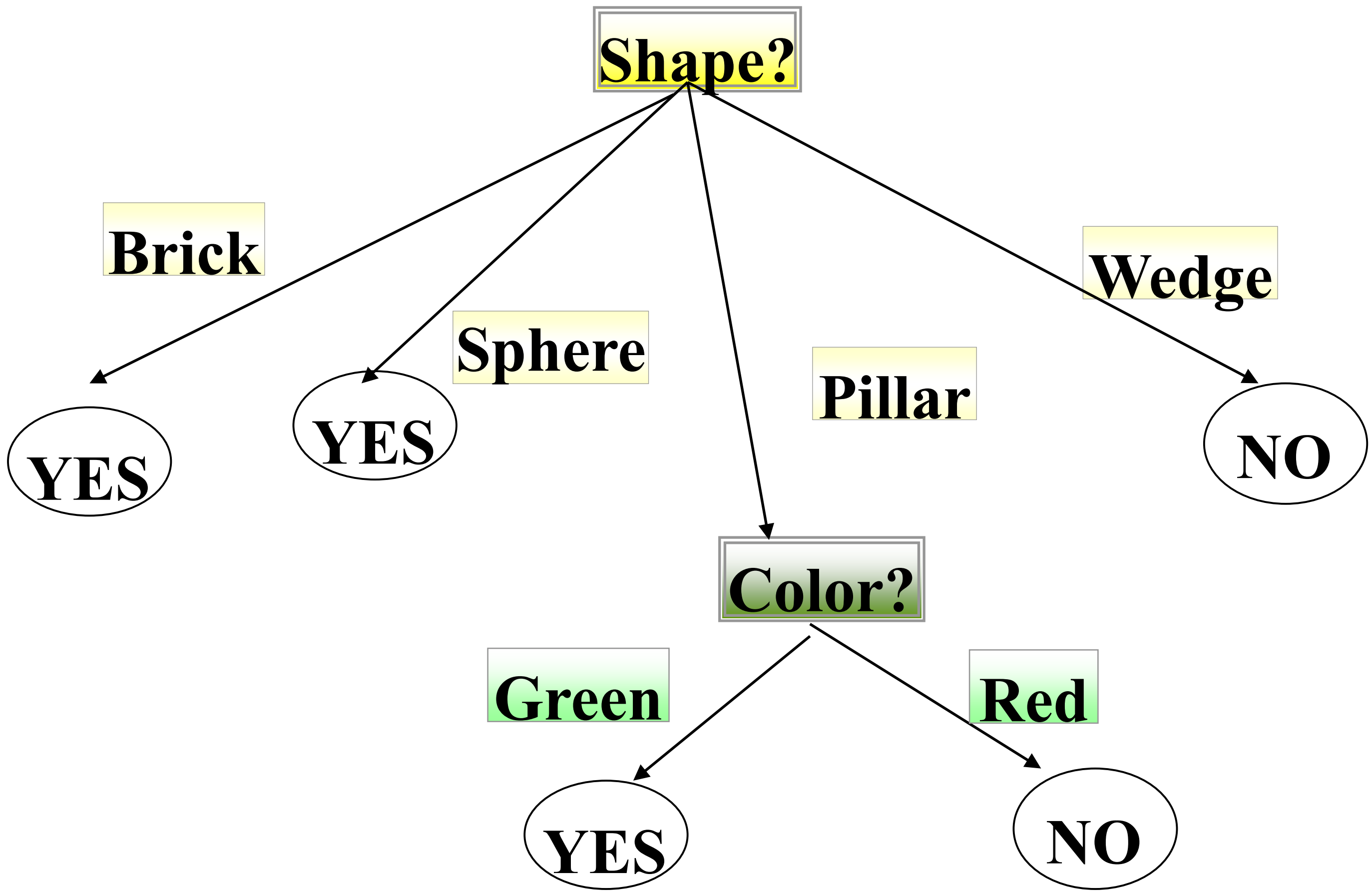


Q1 Answer

- Classify the test example

(Size="small", Color="green", Shape= "pillar")

as ``**YES**" class, i.e., **appropriate** for construction.



Q2

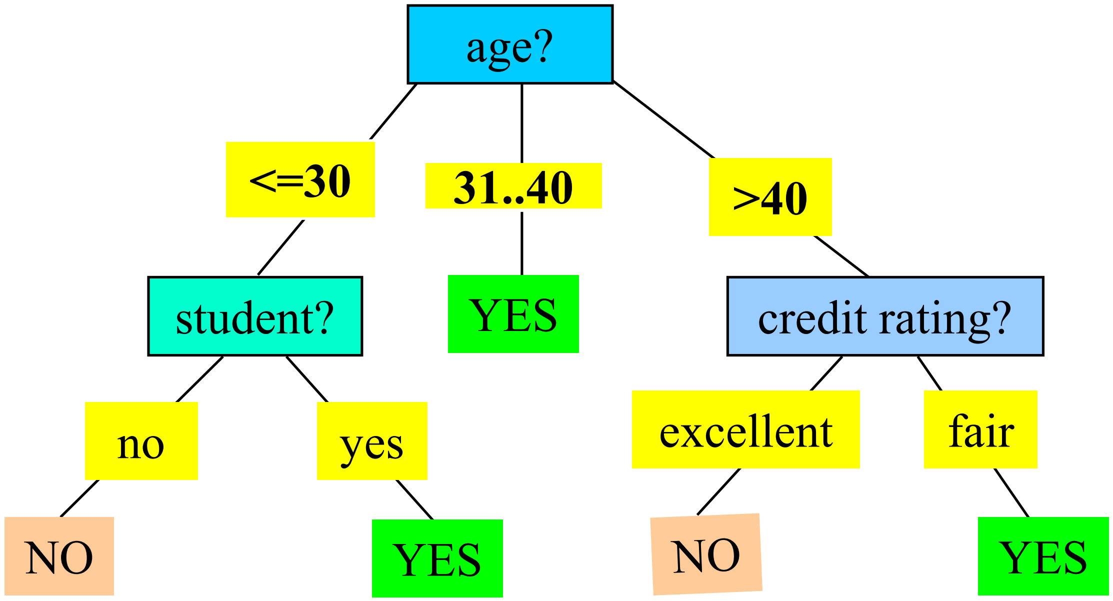
- The tree growth phase in the construction of a tree classifier is computationally expensive and also data-intensive. Briefly describe why this is so.

Q2 Answer

- Because of recursive splitting, the data is scanned again and again. This makes it computationally and memory-wise expensive.

Q3

- Extract rules from the decision given below.



Q3 Answer:

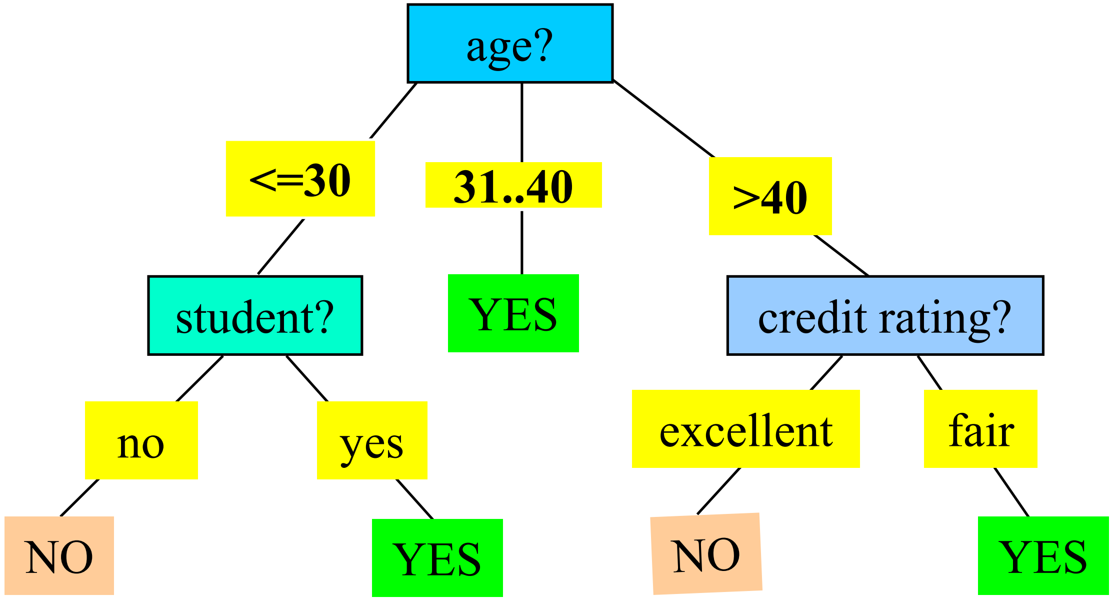
IF *age* <=30 AND *student* = *no* THEN *buys_computer* = *NO*

IF age <=30 AND *student* = *yes* THEN *buys_computer* = *YES*

IF 31 <= *age* <= 40 THEN *buys_computer* = *YES*

IF *age* > 40 AND *credit_rating* = *excellent* THEN *buys_computer* = *NO*

IF *age* >40 AND *credit_rating* = *fair* THEN *buys_computer* = *YES*



Q4 CBA-RG

The CBA-RG Algorithm

```
F_1 = {large 1-ruleitems};
CAR_1 = genRules(F_1);
prCAR_1 = pruneRules(CAR_1);
for (k = 2; F_{k-1} ≠ ∅; k++) do
    C_k = candidateGen(F_{k-1});
    for each data case d ∈ D do
        C_d = ruleSubset(C_k, d);
        for each candidate c ∈ C_d do
            c.condsupCount++;
            if d.class = c.class then c.rulesupCount++
        end
    end
    F_k = {c ∈ C_k | c.rulesupCount ≥ minsup};
    CAR_k = genRules(F_k);
    prCAR_k = pruneRules(CAR_k);
end
CARs = ∪_k CAR_k;
prCARs = ∪_k prCAR_k;
```

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

Minimum Class supports:
$\boldsymbol{minsup}$: **15%**
$\boldsymbol{minconf}$: **60%**

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

$\boxed{F_1 = \{\text{large 1-ruleitems}\};}$
$CAR_1 = \text{genRules}(F_1);$
$prCAR_1 = \text{pruneRules}(CAR_1);$

- Enumerate all possible attribute + class
- Check $minsup > 15\%$

$$ruleitem = <(condset, condsupCount), (y, rulesupCount)>$$

Enumerations	$<(\{A,e\},4),((C,y),3)>$, $\cancel{<(\{A,e\},4),((C,n),1)>}$, $<(\{A,g\},5),((C,n),3)>$, $<(\{A,g\},5),((C,y),2)>$, $\cancel{<(\{A,f\},1),((C,n),1)>}$, $<(\{B,p\},3),((C,y),2)>$, $\cancel{<(\{B,p\},3),((C,n),1)>}$, $<(\{B,q\},5),((C,y),3)>$, $<(\{B,q\},5),((C,n),2)>$, $<(\{B,w\},2),((C,n),2)>$
F_1	$<(\{A,e\},4),((C,y),3)>$, $<(\{A,g\},5),((C,n),3)>$, $<(\{A,g\},5),((C,y),2)>$, $<(\{B,p\},3),((C,y),2)>$, $<(\{B,q\},5),((C,y),3)>$, $<(\{B,q\},5),((C,n),2)>$, $<(\{B,w\},2),((C,n),2)>$

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

F_1 = {large 1-ruleitems};
$\boxed{CAR_1 = \text{genRules}(F_1);}$
$prCAR_1 = \text{pruneRules}(CAR_1);$

- Check $minconf > 60\%$

Enumerations	$<(\{A,e\},4),((C,y),3)>$, $\cancel{<(\{A,e\},4),((C,n),1)>}$, $<(\{A,g\},5),((C,n),3)>$, $<(\{A,g\},5),((C,y),2)>$, $\cancel{<(\{A,f\},1),((C,n),1)>}$, $<(\{B,p\},3),((C,y),2)>$, $\cancel{<(\{B,p\},3),((C,n),1)>}$, $<(\{B,q\},5),((C,y),3)>$, $<(\{B,q\},5),((C,n),2)>$, $<(\{B,w\},2),((C,n),2)>$
F_1	$<(\{A,e\},4),((C,y),3)>$, $<(\{A,g\},5),((C,n),3)>$, $\cancel{<(\{A,g\},5),((C,y),2)>}$, $<(\{B,p\},3),((C,y),2)>$, $<(\{B,q\},5),((C,y),3)>$, $\cancel{<(\{B,q\},5),((C,n),2)>}$, $<(\{B,w\},2),((C,n),2)>$

CAR_1	$<(\{A,e\},4),((C,y),3)>$, $<(\{A,g\},5),((C,n),3)>$, $<(\{B,p\},3),((C,y),2)>$, $<(\{B,q\},5),((C,y),3)>$, $<(\{B,w\},2),((C,n),2)>$

Example: CBA-RG

- Dataset:

$F_1 = \{\text{large 1-ruleitems}\};$
$CAR_1 = \text{genRules}(F_1);$
$\boxed{prCAR_1 = \text{pruneRules}(CAR_1);}$

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

- Pruning is optional.

Enumerations	$<(\{A,e\},4),((C,y),3)>, \cancel{<(\{A,e\},4),((C,n),1)>},$ $<(\{A,g\},5),((C,n),3)>, <(\{A,g\},5),((C,y),2)>,$ $\cancel{<(\{A,f\},1),((C,n),1)>}, <(\{B,p\},3),((C,y),2)>,$ $\cancel{<(\{B,p\},3),((C,n),1)>}, <(\{B,q\},5),((C,y),3)>,$ $<(\{B,q\},5),((C,n),2)>, <(\{B,w\},2),((C,n),2)>$
F_1	$<(\{A,e\},4),((C,y),3)>, <(\{A,g\},5),((C,n),3)>,$ $\cancel{<(\{A,g\},5),((C,y),2)>}, <(\{B,p\},3),((C,y),2)>,$ $<(\{B,q\},5),((C,y),3)>, \cancel{<(\{B,q\},5),((C,n),2)>},$ $<(\{B,w\},2),((C,n),2)>$

CAR_1	$<(\{A,e\},4),((C,y),3)>, <(\{A,g\},5),((C,n),3)>,$ $<(\{B,p\},3),((C,y),2)>, <(\{B,q\},5),((C,y),3)>,$ $<(\{B,w\},2),((C,n),2)>$
$prCAR_1$	$<(\{A,e\},4),((C,y),3)>, <(\{A,g\},5),((C,n),3)>,$ $<(\{B,p\},3),((C,y),2)>, <(\{B,q\},5),((C,y),3)>,$ $<(\{B,w\},2),((C,n),2)>$

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

- Algorithm Line 4-12
- Enumerate all possible pairs from F_1
- Compare each pair with dataset

F_1	$<(\{A,e\},4),((C,y),3)>$ $<(\{A,g\},5),((C,n),3)>$, $<(\{A,g\},5),((C,y),2)>$, $<(\{B,p\},3),((C,y),2)>$, $<(\{B,q\},5),((C,y),3)>$, $<(\{B,q\},5),((C,n),2)>$, $<(\{B,w\},2),((C,n),2)>$
Enumerations	$<\{(A,e),(B,p)\},(C,y)>$,$<\{(A,e),(B,q)\},(C,y)>$, $<\{(A,g),(B,p)\},(C,y)>$,$<\{(A,g),(B,q)\},(C,y)>$, $<\{(A,g),(B,q)\},(C,n)>$,$<\{(A,g),(B,w)\},(C,n)>$

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

- Algorithm Line 4-12
- Enumerate all possible pairs from F_1
- Compare each pair with dataset

F_1	$<(\{A,e\},4),\big((C,y),3\big)>, <(\{A,g\},5),\big((C,n),3\big)>,$ $<(\{A,g\},5),\big((C,y),2\big)>, <(\{B,p\},3),\big((C,y),2\big)>,$ $<(\{B,q\},5),\big((C,y),3\big)>, <(\{B,q\},5),\big((C,n),2\big)>,$ $<(\{B,w\},2),\big((C,n),2\big)>$
Enumerations	$<\{(A,e),(B,p)\},(C,y)>, <\{(A,e),(B,q)\},(C,y)>,$ $<\{(A,g),(B,p)\},(C,y)>, <\{(A,g),(B,q)\},(C,y)>,$ $<\{(A,g),(B,q)\},(C,n)>, <\{(A,g),(B,w)\},(C,n)>$
C_2	$<(\{(A,e),(B,p)\},3),\big((C,y),2\big)>,$ $<(\{(A,e),(B,q)\},1),\big((C,y),1\big)>,$ $<(\{(A,g),(B,p)\},0),\big((C,y),0\big)>,$ $<(\{(A,g),(B,q)\},3),\big((C,y),2\big)>,$ $<(\{(A,g),(B,q)\},3),\big((C,n),1\big)>,$ $<(\{(A,g),(B,w)\},2),\big((C,n),2\big)>$

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

$F_k = \{c \in C_k \mid c.\text{rulesupCount} \geq minsup\};$
$CAR_k = \text{genRules}(F_k);$
$prCAR_k = \text{pruneRules}(CAR_k);$

C_2	$<(\{(A,e),(B,p)\},3),((C,y),2)>,$ $\cancel{<(\{(A,e),(B,q)\},1),((C,y),1)>,}$ $\cancel{<(\{(A,g),(B,p)\},0),((C,y),0)>,}$ $<(\{(A,g),(B,q)\},3),((C,y),2)>,$ $<(\{(A,g),(B,q)\},3),((C,n),1)>,$ $<(\{(A,g),(B,w)\},2),((C,n),2)>$
F_2	$<(\{(A,e),(B,p)\},3),((C,y),2)>,$ $<(\{(A,g),(B,q)\},3),((C,y),2)>,$ $<(\{(A,g),(B,q)\},3),((C,n),1)>,$ $<(\{(A,g),(B,w)\},2),((C,n),2)>$

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

$F_k = \{c \in C_k \mid c.\text{rulesupCount} \geq minsup\}$;
$\boxed{CAR_k = \text{genRules}(F_k);}$
$prCAR_k = \text{pruneRules}(CAR_k)$;

C_2	$<(\{(A,e),(B,p)\},3),((C,y),2)>,$ $\cancel{<(\{(A,e),(B,q)\},1),((C,y),1)>,}$ $\cancel{<(\{(A,g),(B,p)\},0),((C,y),0)>,}$ $<(\{(A,g),(B,q)\},3),((C,y),2)>,$ $<(\{(A,g),(B,q)\},3),((C,n),1)>,$ $<(\{(A,g),(B,w)\},2),((C,n),2)>$
F_2	$<(\{(A,e),(B,p)\},3),((C,y),2)>,$ $<(\{(A,g),(B,q)\},3),((C,y),2)>,$ $\cancel{<(\{(A,g),(B,q)\},3),((C,n),1)>,}$ $conf = 33\% < 60\%$ $<(\{(A,g),(B,w)\},2),((C,n),2)>$

CAR_2	$<(\{(A,e),(B,p)\},3),((C,y),2)>,$ $<(\{(A,g),(B,q)\},3),((C,y),2)>,$ $<(\{(A,g),(B,w)\},2),((C,n),2)>$

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

$F_k = \{c \in C_k \mid c.\text{rulesupCount} \geq minsup\};$

$CAR_k = \text{genRules}(F_k);$

$prCAR_k = \text{pruneRules}(CAR_k);$

CAR_2	$<(\{(A,e),(B,p)\},3),((C,y),2)>$, $<(\{(A,g),(B,q)\},3),((C,y),2)>$, $<(\{(A,g),(B,w)\},2),((C,n),2)>$

Error rate of r_1: $<(\{(A,e),(B,p)\},3),((C,y),2)>$ = 33%

Error rate of r_1^-: $<(\{A,e\},4),((C,y),3)>$ = 25%

Error rate of r_1 > Error rate of r_1^- $\Rightarrow$ Prune r_1

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

$F_k = \{c \in C_k \mid c.\text{rulesupCount} \geq minsup\}$;
$CAR_k = \text{genRules}(F_k)$;
$\boxed{prCAR_k = \text{pruneRules}(CAR_k);}$

CAR_2	$\cancel{<(\{(A,e),(B,p)\},3),((C,y),2)>,}$ $<(\{(A,g),(B,q)\},3),((C,y),2)>,$ $<(\{(A,g),(B,w)\},2),((C,n),2)>$

Error rate of r_1: $<(\{(A,e),(B,p)\},3),((C,y),2)>$ = 33%

Error rate of r_1^-: $<(\{A,e\},4),((C,y),3)>$ = 25%

Error rate of r_1 > Error rate of r_1^- $\Rightarrow$ Prune r_1

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

$F_k = \{c \in C_k \mid c.\text{rulesupCount} \geq minsup\};$

$CAR_k = \text{genRules}(F_k);$

$\boxed{prCAR_k = \text{pruneRules}(CAR_k);}$

CAR_2	$\cancel{<(\{(A,e),(B,p)\},3),((C,y),2)>,}$ $\boxed{<(\{(A,g),(B,q)\},3),((C,y),2)>,}$ $<(\{(A,g),(B,w)\},2),((C,n),2)>$

Error rate of r_2: $<(\{(A,g),(B,q)\},3),((C,y),2)>$ = 33%

Error rate of r_2^-: $<(\{A,g\},5),((C,y),2)>$ = 60%

Error rate of r_2^-: $<(\{B,q\},5),((C,y),3)>$ = 40%

Error rate of r_2 < Error rates of r_2^- $\Rightarrow$ Keep r_2

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

13 $F_k = \{c \in C_k \mid c.\text{rulesupCount} \geq minsup\}$;
14 $CAR_k = \text{genRules}(F_k)$;
15 $\boxed{prCAR_k = \text{pruneRules}(CAR_k);}$

CAR_2	~~<({(A,e),(B,p)},3),((C,y),2)>,~~ <({(A,g),(B,q)},3),((C,y),2)>, <({(A,g),(B,w)},2),((C,n),2)>

Error rate of r_3: <({(A,g),(B,w)},2),((C,n),2)> = 0%

Error rate of r_3^-: <({A,g},5),((C,n),2)> = 40%

Error rate of r_3^-: <({B,w},2),((C,n),2)> = 0%

Error rate of r_3 >= Error rates of r_3^- $\Rightarrow$ Prune r_3

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

```
F_k = {c ∈ C_k | c.rulesupCount ≥ minsup};
CAR_k = genRules(F_k);
prCAR_k = pruneRules(CAR_k);
```

CAR_2	$\cancel{<(\{(A,e),(B,p)\},3),((C,y),2)>,}$ $<(\{(A,g),(B,q)\},3),((C,y),2)>,$ $\cancel{<(\{(A,g),(B,w)\},2),((C,n),2)>}$

Error rate of r_3: $<(\{(A,g),(B,w)\},2),\big((C,n),2\big)>$ = 0%

Error rate of r_3^-: $<(\{A,g\},5),\big((C,n),2\big)>$ = 40%

Error rate of r_3^-: $<(\{B,w\},2),\big((C,n),2\big)>$ = 0%

Error rate of r_3 >= Error rates of r_3^- $\Rightarrow$ Prune r_3

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

$F_k = \{c \in C_k \mid c.\text{rulesupCount} \geq minsup\}$;
$CAR_k = \text{genRules}(F_k)$;
$\boxed{prCAR_k = \text{pruneRules}(CAR_k);}$

CAR_2	$\cancel{<(\{(A,e),(B,p)\},3),((C,y),2)>,}$ $<(\{(A,g),(B,q)\},3),((C,y),2)>,$ $\cancel{<(\{(A,g),(B,w)\},2),((C,n),2)>}$
$prCAR_2$	$<(\{(A,g),(B,q)\},3),((C,y),2)>$

Example: CBA-RG

- Dataset:

Attribute A	Attribute B	Class C
e	p	y
e	p	y
e	q	y
g	q	y
g	q	y
g	q	n
g	w	n
g	w	n
e	p	n
f	q	n

17 $CARs = \bigcup_k CAR_k$;
18 $prCARs = \bigcup_k prCAR_k$;

CAR_1	$<(\{A,e\},4),\big((C,y),3\big)>, <(\{A,g\},5),\big((C,n),3\big)>, <(\{B,p\},3),\big((C,y),2\big)>, <(\{B,q\},5),\big((C,y),3\big)>, <(\{B,w\},2),\big((C,n),2\big)>$
$prCAR_1$	$<(\{A,e\},4),\big((C,y),3\big)>, <(\{A,g\},5),\big((C,n),3\big)>, <(\{B,p\},3),\big((C,y),2\big)>, <(\{B,q\},5),\big((C,y),3\big)>, <(\{B,w\},2),\big((C,n),2\big)>$

CAR_2	$<(\{(A,e),(B,p)\},3),\big((C,y),2\big)>, <(\{(A,g),(B,q)\},3),\big((C,y),2\big)>, <(\{(A,g),(B,w)\},2),\big((C,n),2\big)>$
$prCAR_2$	$<(\{(A,g),(B,q)\},3),\big((C,y),2\big)>$

$CARs$	$CAR_1 \cup CAR_2$
$prCARs$	$prCAR_1 \cup prCAR_2$

Q5: CBA

- Dataset:

Attribute A	Attribute B	Class C
e	p	$??$
g	q	$??$
g	m	$??$
k	p	$??$

$$\begin{cases} r_5: & B = w \rightarrow n \\ r_1: & A = e \rightarrow y \\ r_6: & A = g, B = q \rightarrow y \\ R7: & A = g \rightarrow n \end{cases}$$

& Default Class **n**

Attribute A	Attribute B	Class C
e	p	$y\ (r1)$
g	q	$y(r6)$
g	m	$n(r7)$
k	p	$n(default)$