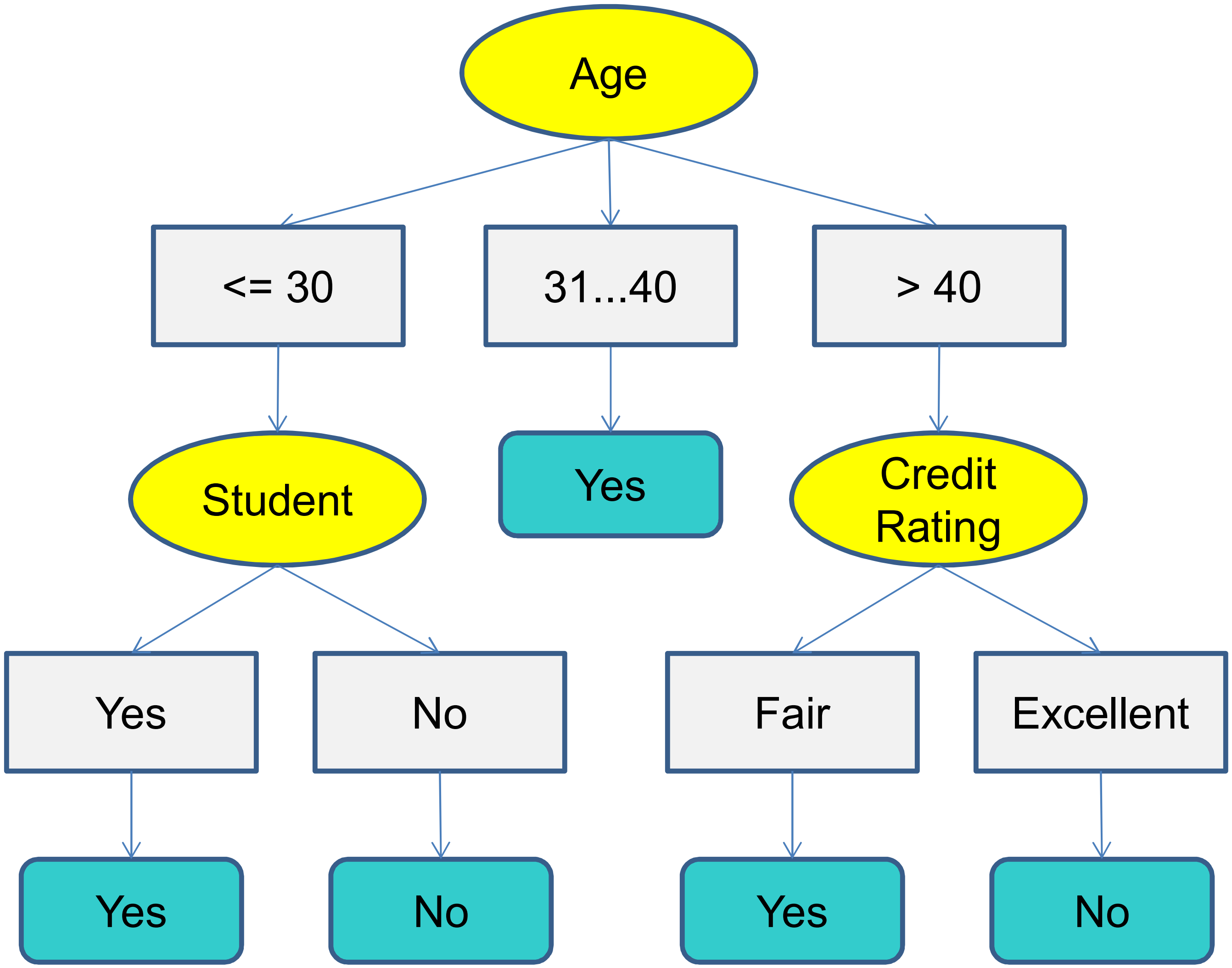
Classification Examples

Quach Dinh Hoang

Dataset

Attributes				*Class*
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

Decision Tree Induction



Entropy - Class

buys_computer
no
no
yes
yes
yes
no
yes
no
yes
yes
yes
yes
yes
no

Sort

buys_computer
no
no
no
no
no
yes
yes
yes
yes
yes
yes
yes
yes
yes

5/14

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

9/14

$$\text{Info(buys\_computer)} = -P(no)^* \log_2(P(no)) - P(yes)^* \log_2(P(yes))$$
$$= -(5/14)\log_2(5/14) - (9/14)\log_2(9/14) = 0.94$$

AVC (Attribute, Value, Class) Table

		buys_computer	
		Yes	No
Age	<=30	2	3
	31…40	4	0
	>40	3	2

		buys_computer	
		Yes	No
Income	low	3	1
	medium	4	2
	high	2	2

		buys_computer	
		Yes	No
Student	no	3	4
	yes	6	1

		buys_computer	
		Yes	No
Credit rating	fair	6	2
	excellent	3	3

Information Gain - Age

		buys_computer		
		Yes	No	
Age	<=30	2	3	**5**
	31…40	4	0	**4**
	>40	3	2	**5**
				14

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

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

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

$Info_{Age}$(buy_computer) = P(<=30)*Info(2,3) + P(31…40)*Info(4,0) + P(>40)*Info(3,2)
= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971 = 0.693

Gain(Age) = Info(buy_computer) - $Info_{Age}$(buy_computer)
= 0.94 - 0.693 = 0.247

Information Gain

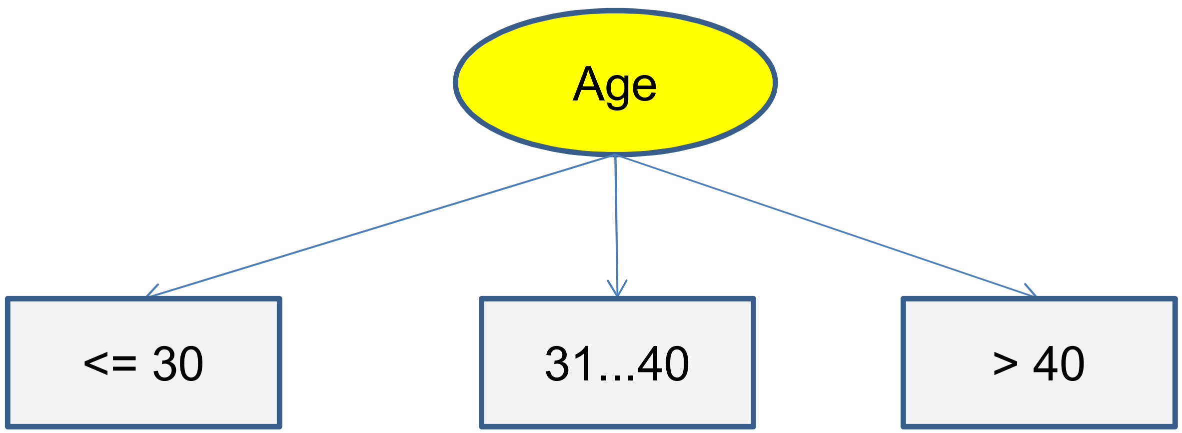
		buys_computer	
		Yes	No
Age	<=30	2	3
	31…40	4	0
	>40	3	2
Gain = 0.247			

		buys_computer	
		Yes	No
Income	low	3	1
	medium	4	2
	high	2	2
Gain = 0.029			

		buys_computer	
		Yes	No
Student	no	3	4
	yes	6	1
Gain = 0.151			

		buys_computer	
		Yes	No
Credit rating	fair	6	2
	excellent	3	3
Gain = 0.048			

Decision Tree – Root Node

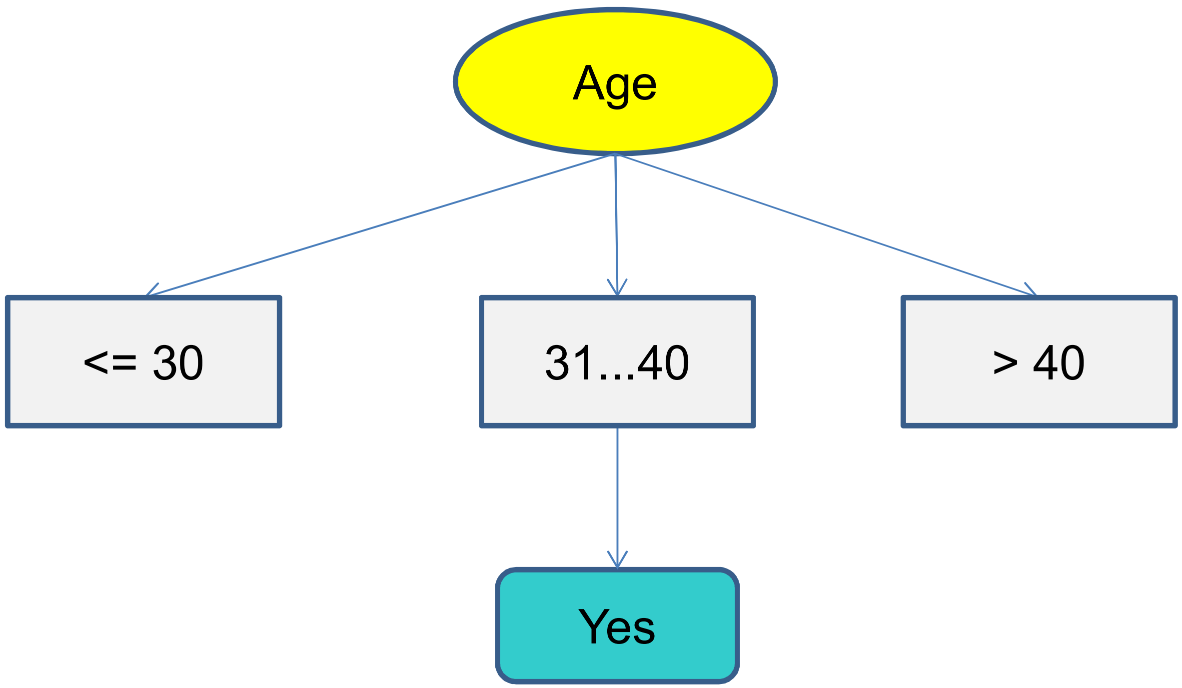


Dataset – Sort by Root Node

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

Age = 31…40

age	income	student	credit_rating	buys_computer
31…40	high	no	fair	yes
31…40	low	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes



Age <= 30

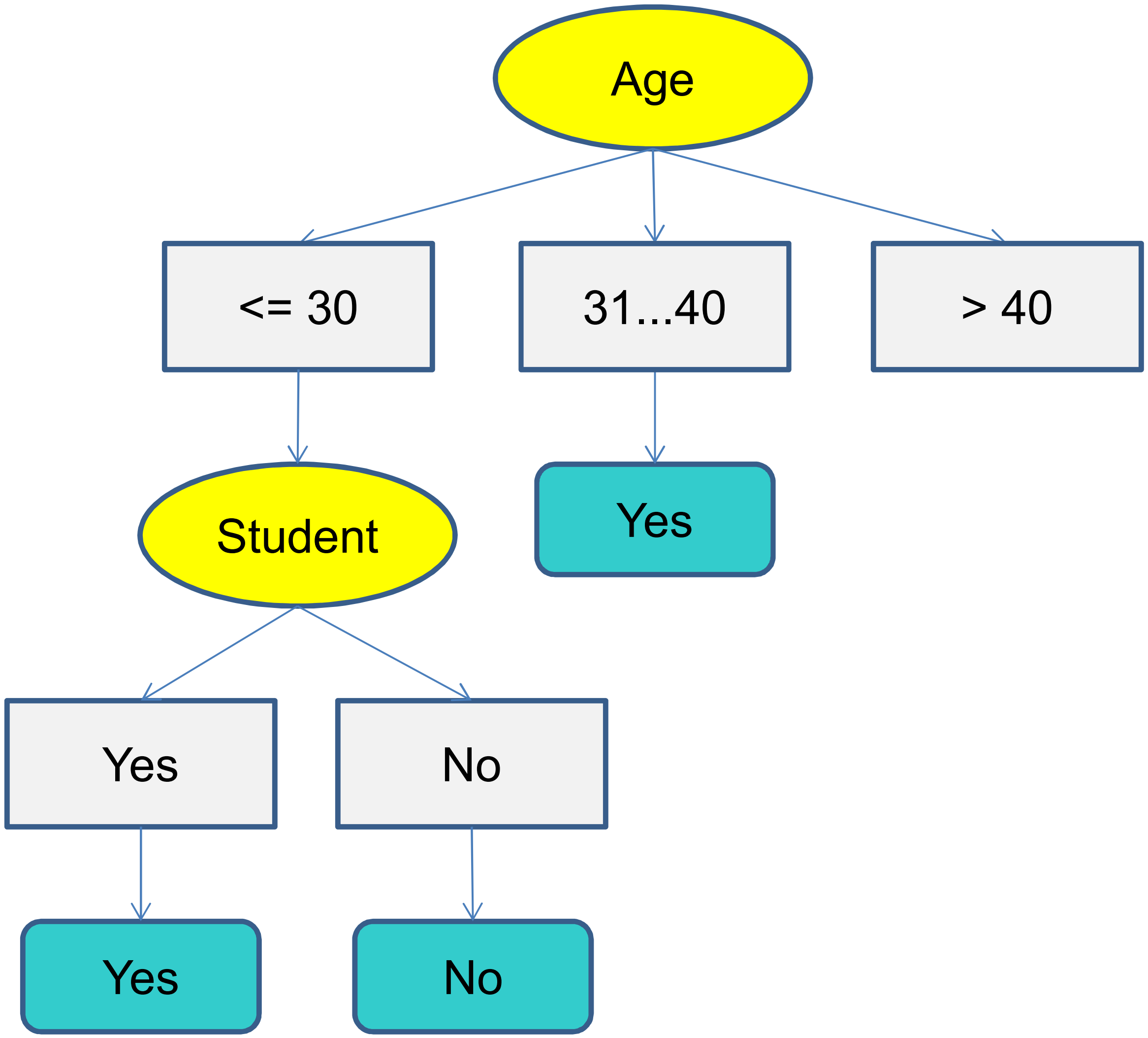
age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

		buys_computer	
		Yes	No
Income	low	1	0
	medium	1	1
	high	0	2
Gain = 0.57			

		buys_computer	
		Yes	No
Student	no	0	3
	yes	2	0
Gain = 0.97			

		buys_computer	
		Yes	No
Credit rating	fair	1	2
	excellent	1	1
Gain = 0.02			

Age <= 30



Age > 40

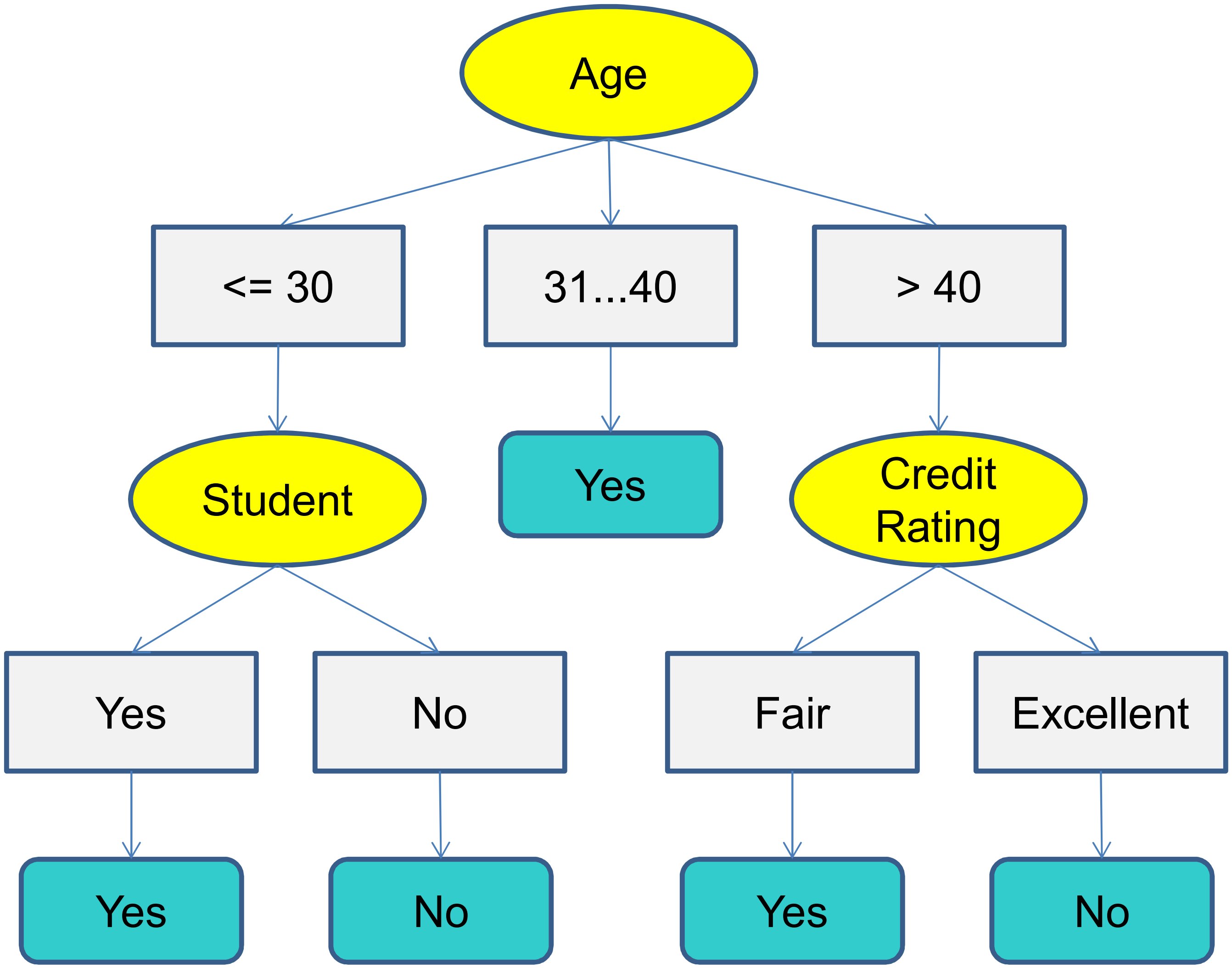
age	income	student	credit_rating	buys_computer
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
>40	medium	yes	fair	yes
>40	medium	no	excellent	no

		buys_computer	
		Yes	No
Income	low	1	1
	medium	2	1
Gain = 0.57			

		buys_computer	
		Yes	No
Student	no	1	1
	yes	2	1
Gain = 0.02			

		buys_computer	
		Yes	No
Credit rating	fair	3	0
	excellent	0	2
Gain = 0.97			

Age > 40



Age
<= 30
31...40
> 40
Student
Yes
Credit Rating
Yes
No
Fair
Excellent
Yes
No
Yes
No

Decision Rules

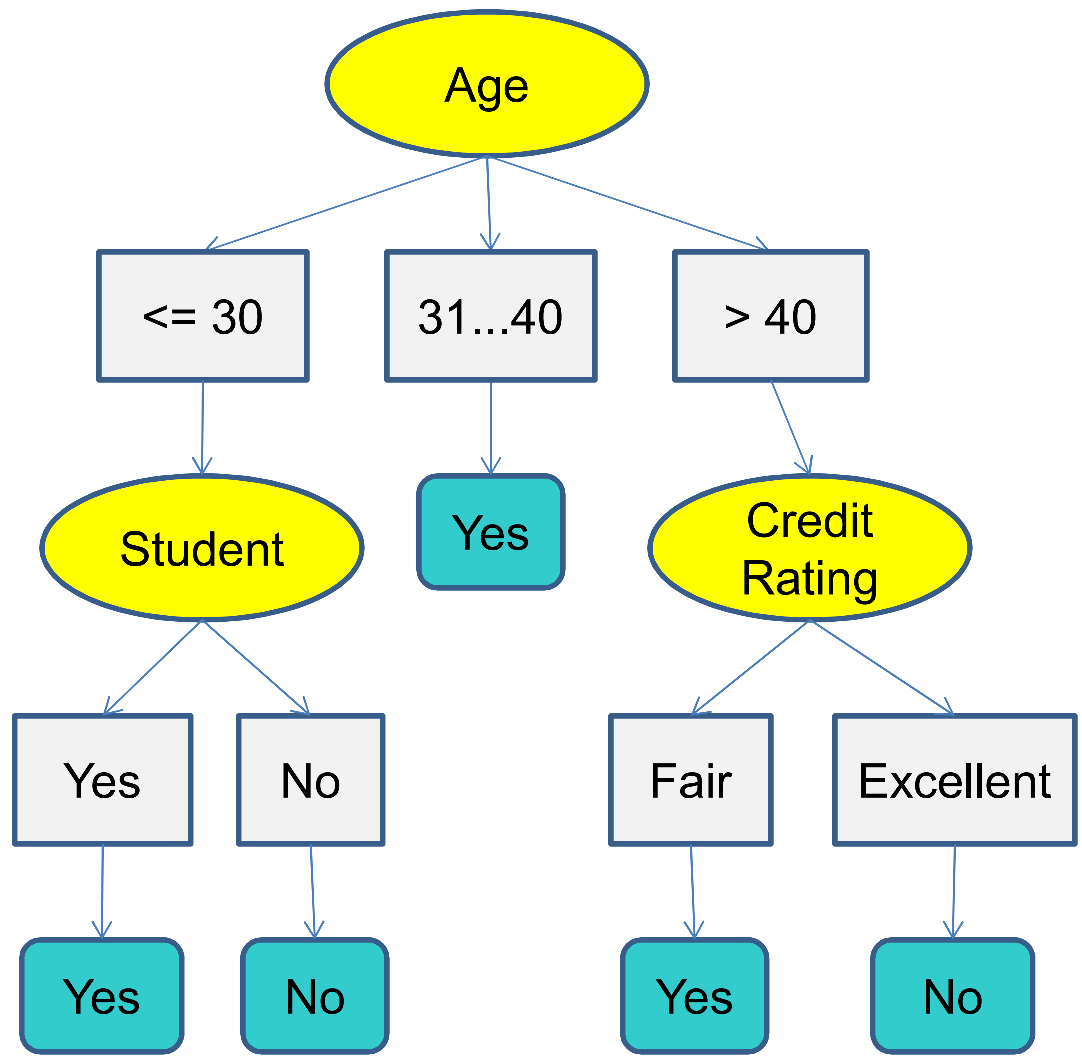
R1: IF (Age <= 30 And Student = Yes) THEN buy_computer = Yes

R2: IF (Age <= 30 And Student = No) THEN buy_computer = No

R3: IF (Age = 31…40) THEN buy_computer = Yes

R4: IF (Age > 40 And CreditRating = Fair) THEN buy_computer = Yes

R5: IF (Age <= 30 And CreditRating = Excellent) THEN buy_computer = Yes



Naive Bayesian Classifier – NBC

- Bayes' theorem $$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Since P(X) is constant for all classes, we only need maximize $$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

- Assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

AVC (Attribute, Value, Class) Table

		buys_computer	
		Yes	No
Age	<=30	2	3
	31…40	4	0
	>40	3	2

		buys_computer	
		Yes	No
Income	low	3	1
	medium	4	2
	high	2	2

		buys_computer	
		Yes	No
Student	no	3	4
	yes	6	1

		buys_computer	
		Yes	No
Credit rating	fair	6	2
	excellent	3	3

Likelihood Tables

		buys_computer	
		Yes	No
Age	<=30	2/9	3/5
	31…40	4/9	0
	>40	3/9	2/5

		buys_computer	
		Yes	No
Income	low	3/9	1/5
	medium	4/9	2/5
	high	2/9	2/5

		buys_computer	
		Yes	No
Student	no	3/9	4/5
	yes	6/9	1/5

		buys_computer	
		Yes	No
Credit rating	fair	6/9	2/5
	excellent	3/9	3/5

P(Student = yes | buy_computer = yes) = 6/9

NBC – Prediction

- X = (age <= 30 , income = medium, student = yes, credit_rating = fair)
- **P(X|C$_i$)**

  P(X | buys_computer = yes) =

  P(age <=30 | buys_computer = yes) *
  P(income = medium | buys_computer = yes) *
  P(student = yes | buys_computer = yes) *
  P(credit_rating = fair | buys_computer = yes)

  P(X | buys_computer = no) =

  P(age <=30 | buys_computer = no) *
  P(income = medium | buys_computer = no) *
  P(student = yes | buys_computer = no) *
  P(credit_rating = fair | buys_computer = no)

NBC – Prediction

- X = (age <= 30 , income = medium, student = yes, credit_rating = fair)
- **$P(C_i)$**

  P(buys_computer = yes) = 9/14

  P(buys_computer = no) = 5/14
- **$P(X|C_i) * P(C_i)$**

  P(X | buys_computer = yes) * P(buys_computer = yes)

  P(X | buys_computer = no) * P(buys_computer = no)