



# CS 412 Intro. to Data Mining

การทำนาย

## Chapter 8. Classification: Basic Concepts

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017








# Chapter 8. Classification: Basic Concepts

---

- ❑ Classification: Basic Concepts 
- ❑ Decision Tree Induction
- ❑ Bayes Classification Methods
- ❑ Linear Classifier
- ❑ Model Evaluation and Selection
- ❑ Techniques to Improve Classification Accuracy: Ensemble Methods
- ❑ Additional Concepts on Classification
- ❑ Summary

แบ่งออกเป็น 2 กลุ่ม คือ supervised กับ unsupervised

# Supervised vs. Unsupervised Learning (1)

❑ Supervised learning (classification) คือ การสร้างโมเดลแบบมีผู้สอน

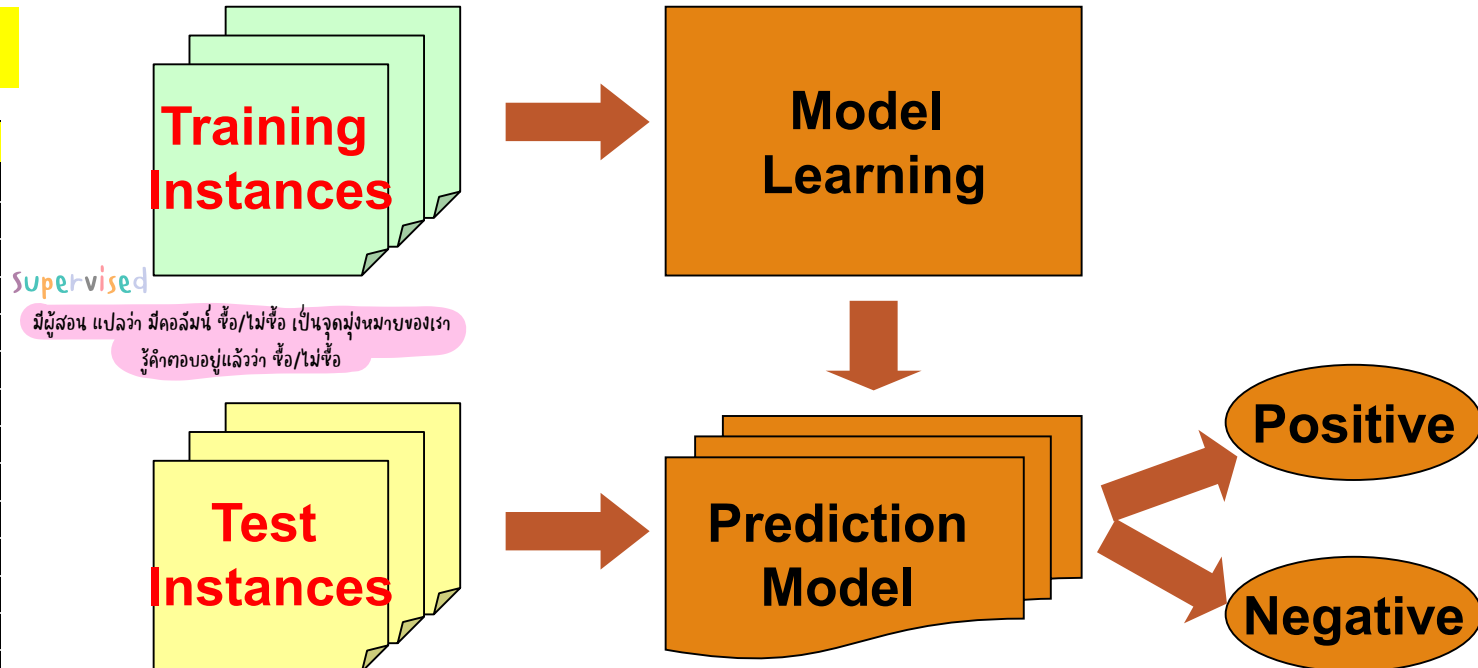
❑ Supervision: The training data such as observations or measurements are accompanied by **labels** indicating the classes which they belong to

❑ New data is classified based on the models built from the training set

เก็บข้อมูลของลูกค้าร้านขายคอมพิวเตอร์

Training Data with class label:

อายุ	รายได้	สถานะ	เครดิตดีแค่ไหน	ซื้อ/ไม่ซื้อ
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





# Supervised vs. Unsupervised Learning (2)

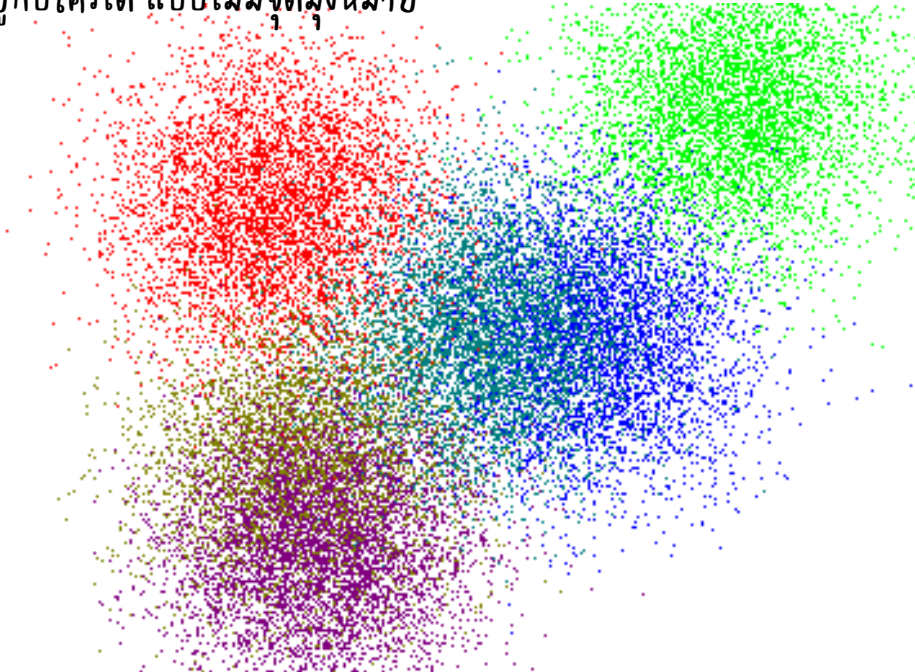
## □ Unsupervised learning (clustering) คือ การสร้างโมเดลแบบไม่มีผู้สอน

□ The class labels of training data are unknown

ไม่มีจุดมุ่งหมายตั้งแต่ต้น

□ Given a set of observations or measurements, establish the possible existence of classes or clusters in the data

จัดกลุ่มว่าใครอยู่กับใครได้ แบบไม่มีจุดมุ่งหมาย



# Prediction Problems: Classification vs. Numeric Prediction

## ❑ Classification สร้างโมเดลเพื่อใช้ฟัเจอร์มาทำนายคำตอบ

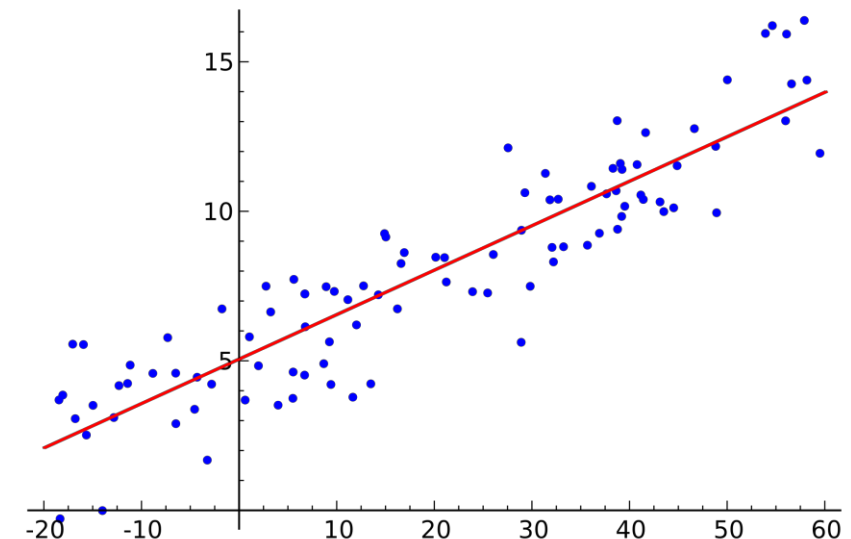
- ❑ Predict categorical class labels (discrete or nominal)
- ❑ Construct a model based on the training set and the **class labels** (the values in a classifying attribute) and use it in classifying new data

## ❑ Numeric prediction

- ❑ Model continuous-valued functions (i.e., predict unknown or missing values)

## ❑ Typical applications of classification

- ❑ Credit/loan approval
- ❑ Medical diagnosis: if a tumor is cancerous or benign
- ❑ Fraud detection: if a transaction is fraudulent
- ❑ Web page categorization: which category it is



# Classification—Model Construction, Validation and Testing

- ❑ **Model construction** เหมือนการ เรียน -> สอบ -> นำไปใช้งาน
  - ❑ Each sample is assumed to belong to a predefined class (shown by the **class label**)
  - ❑ The set of samples used for model construction is **training set**
  - ❑ Model: Represented as decision trees, rules, mathematical formulas, or other forms
- ❑ **Model Validation and Testing:**
  - ❑ **Test:** Estimate accuracy of the model
    - ❑ The known label of test sample is compared with the classified result from the model
    - ❑ *Accuracy*: % of test set samples that are correctly classified by the model
    - ❑ Test set is independent of training set
  - ❑ **Validation:** If *the test set* is used to select or refine models, it is called **validation** (or development) **(test) set**
- ❑ **Model Deployment:** If the accuracy is acceptable, use the model to classify new data

# Chapter 8. Classification: Basic Concepts

---

- ❑ Classification: Basic Concepts
- ❑ Decision Tree Induction
- ❑ Bayes Classification Methods
- ❑ Linear Classifier
- ❑ Model Evaluation and Selection
- ❑ Techniques to Improve Classification Accuracy: Ensemble Methods
- ❑ Additional Concepts on Classification
- ❑ Summary





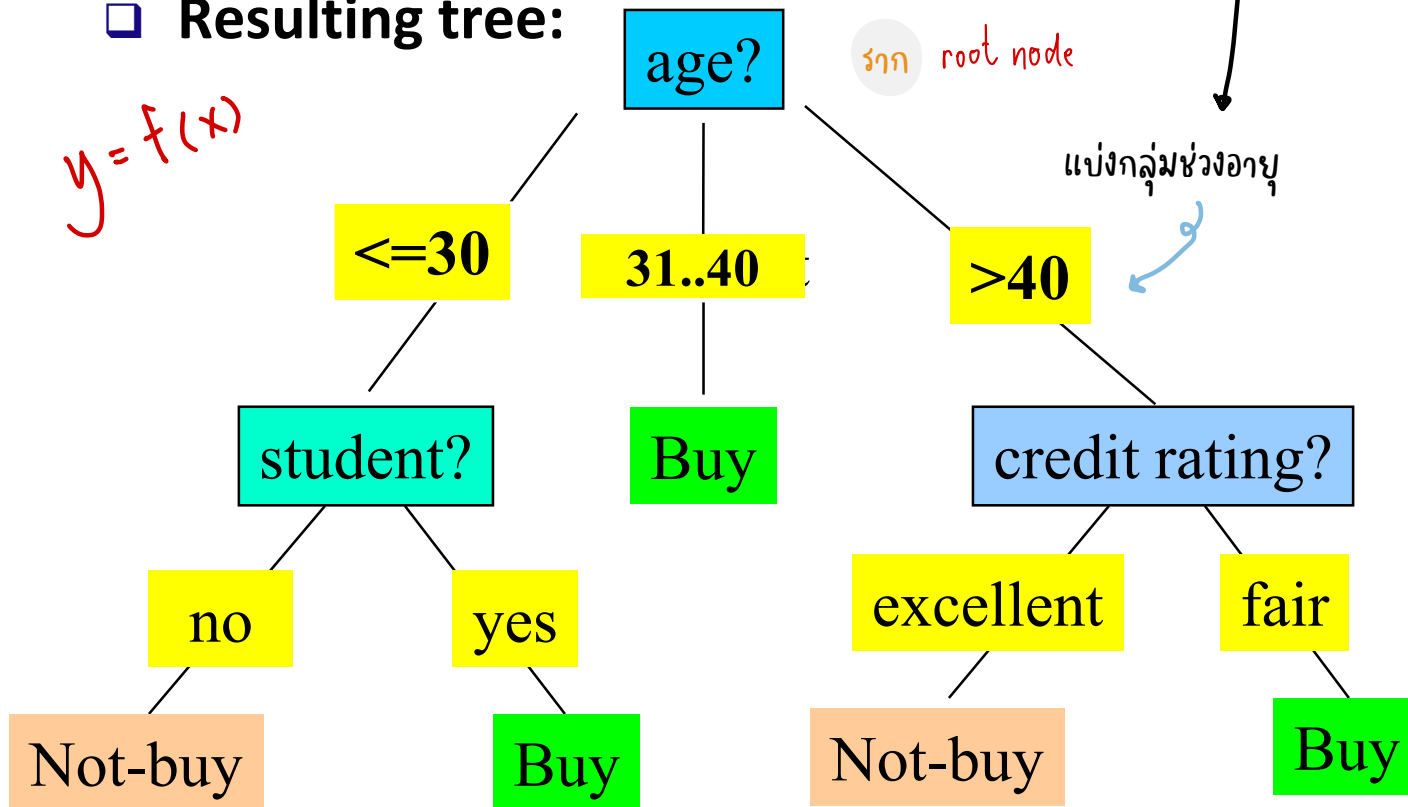
# Decision Tree Induction: An Example

$X$  (Feature)  $y$  (Label)

## Decision tree construction:

- A top-down, recursive, divide-and-conquer process

## Resulting tree:



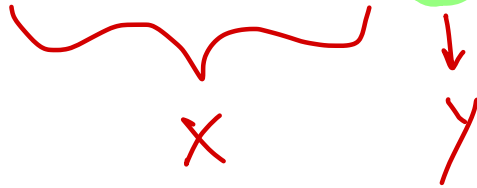
Training data set: Who buys computer?

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

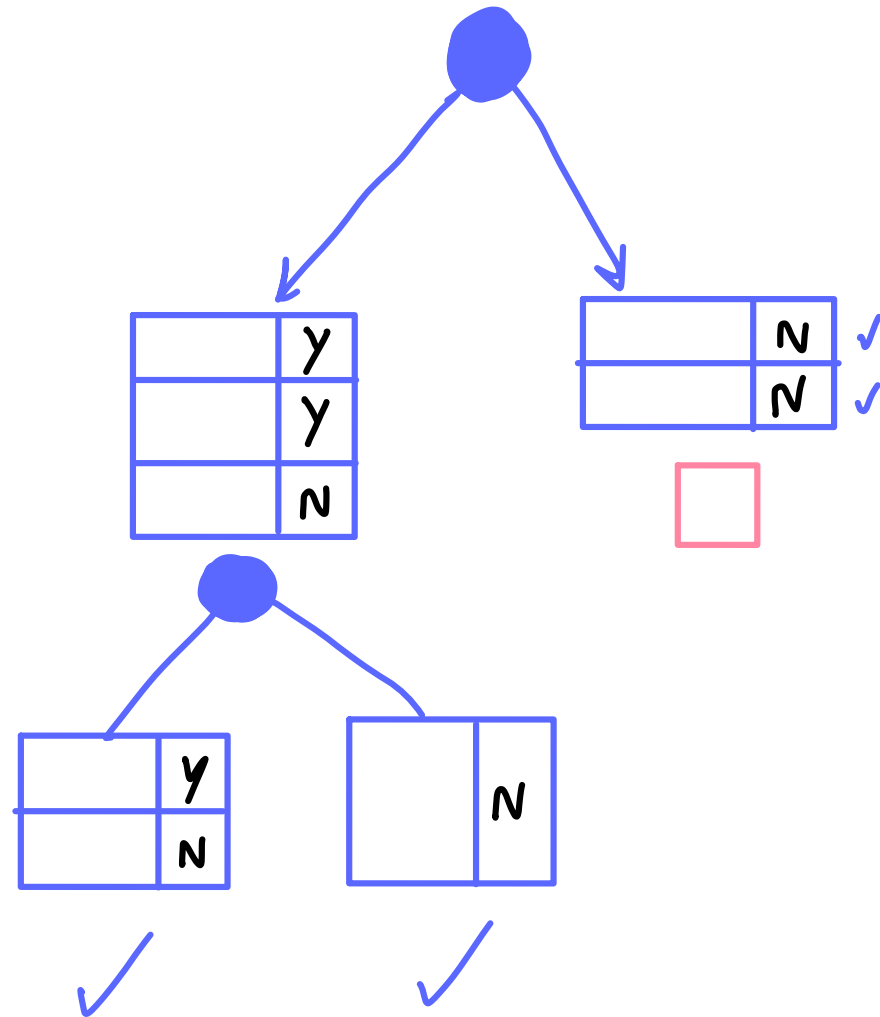
Note: The data set is adapted from "Playing Tennis" example of R. Quinlan

ในการสร้างตารางของ Decision Tree จะเริ่มจากรากไปใบ

				Y
				N
				Y
				N
				N

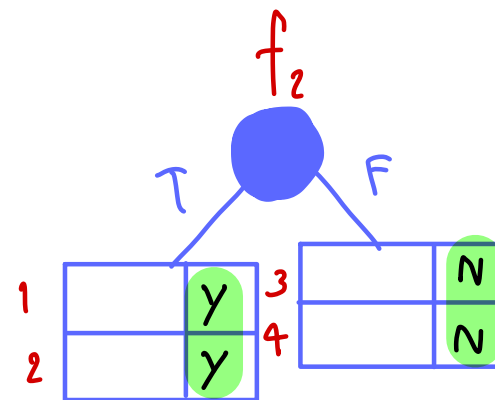
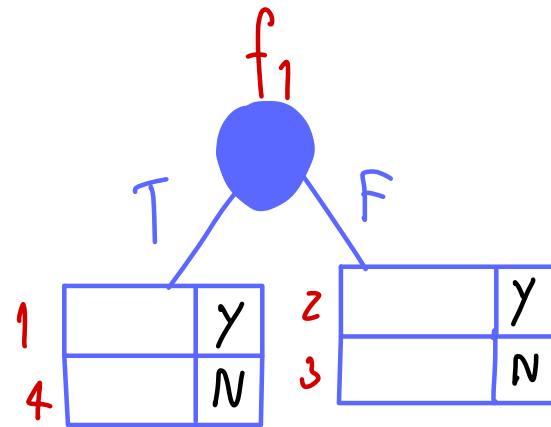


→ เลือก root node



ในการสร้างตารางของ Decision Tree จะเริ่มจากรากไปใบ

	$f_1$	$f_2$	$f_3$	$y$
1	T	T	F	Y
2	F	T	F	Y
3	F	F	F	N
4	T	F	T	N



ถูกจัดกลุ่มแล้ว ไม่ต้องแตกต่อ



# From Entropy to Info Gain: A Brief Review of Entropy

## □ Entropy (Information Theory)

- A measure of uncertainty associated with a random number
- Calculation: For a discrete random variable  $Y$  taking  $m$  distinct values  $\{y_1, y_2, \dots, y_m\}$

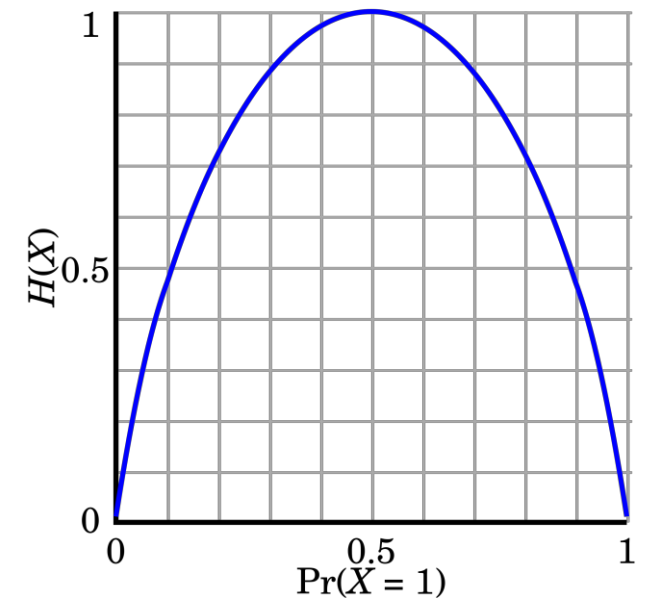
$$H(Y) = - \sum_{i=1}^m p_i \log(p_i) \quad \text{where } p_i = P(Y = y_i)$$

## □ Interpretation

- Higher entropy  $\rightarrow$  higher uncertainty
- Lower entropy  $\rightarrow$  lower uncertainty

## □ Conditional entropy

$$H(Y|X) = \sum_x p(x) H(Y|X = x)$$



**m = 2**

# Information Gain: An Attribute Selection Measure

- ❑ Select the attribute with the highest information gain (used in typical decision tree induction algorithm: ID3/C4.5)
- ❑ Let  $p_i$  be the probability that an arbitrary tuple in  $D$  belongs to class  $C_i$ , estimated by  $|C_{i,D}|/|D|$
- ❑ Expected information (entropy) needed to classify a tuple in  $D$ :

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

- ❑ Information needed (after using  $A$  to split  $D$  into  $v$  partitions) to classify  $D$ :

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

- ❑ Information gained by branching on attribute  $A$  หา node หรือ คำถามที่ดีที่สุด

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



# Example: Attribute Selection with Information Gain

□ Class P: buys\_computer = "yes"

□ Class N: buys\_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) = 0.940$$

age	p <sub>i</sub>	n <sub>i</sub>	I(p <sub>i</sub> , n <sub>i</sub> )
<=30	2	3	0.971
31...40	4	0	0
>40	3	2	0.971

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

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

$\frac{5}{14} I(2,3)$  means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's.

Hence

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Similarly, we can get

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$



11

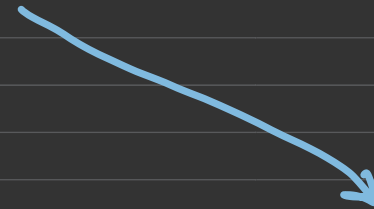
$$G(\text{age}) = 0.246$$

$$G(\text{income}) = 0.029$$

$$G(\text{Student}) = 0.151$$

$$G(\text{Credit}) = 0.048$$


14



age

$< 30$



5


Student

yes

yes

No

No

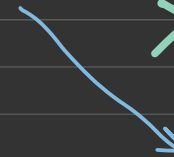
$31 - 40$




4

Yes

$> 40$




5

$\sum x$

Hw

> 30

$$\text{info}(D) = I(2,3) = \overset{y, N}{\boxed{-\frac{2}{5} \log_2 \frac{2}{5}}} \overset{y}{\boxed{-\frac{3}{5} \log_2 \frac{3}{5}}}$$

$$\text{info}_{\text{income}}(D) = \underset{\text{high}}{\frac{2}{5} I(0,2)} + \underset{\text{medium}}{\frac{2}{5} I(1,1)} + \underset{\text{low}}{\frac{1}{5} I(1,0)}$$

$$\text{info}_{\text{student}}(D) = \underset{\text{yes}}{\frac{2}{5} I(2,0)} + \underset{\text{no}}{\frac{3}{5} I(0,3)}$$

$$\text{info}_{\text{credit\_rating}}(D) = \underset{\text{fair}}{\frac{3}{5} I(1,2)} + \underset{\text{excellent}}{\frac{2}{5} I(1,1)}$$

31-40

$$\text{Info}(D) = I(y, N) = \boxed{-\frac{4}{4} \log_2 \frac{4}{4}} - \boxed{\frac{0}{4} \log_2 \frac{0}{4}}$$

$$\text{Info}_{\text{income}}(D) = \frac{2}{4} I_{\text{high}}(1, 1) + \frac{1}{4} I_{\text{medium}}(0, 1) + \frac{1}{4} I_{\text{low}}(1, 0)$$

$$\text{Info}_{\text{student}}(D) = \frac{2}{4} I_{\text{yes}}(2, 0) + \frac{2}{4} I_{\text{no}}(2, 0)$$

$$\text{Info}_{\text{credit\_rating}}(D) = \frac{2}{4} I_{\text{fair}}(2, 0) + \frac{2}{4} I_{\text{excellent}}(2, 0)$$



> 40

$$\text{Info}(D) = I(y, N) = \boxed{-\frac{3}{5} \log_2 \frac{3}{5}} - \frac{2}{5} \log_2 \frac{2}{5}$$

$$\text{Info}_{\text{income}}(D) = \frac{3}{5} I(\underset{\text{high}}{2}, 1) + \frac{2}{5} I(\underset{\text{medium}}{1}, 1) + - I(\underset{\text{low}}{,}, )$$

$$\text{Info}_{\text{student}}(D) = \frac{3}{5} I(\underset{\text{yes}}{2}, 1) + \frac{2}{5} I(\underset{\text{no}}{1}, 1)$$

$$\text{Info}_{\text{credit\_rating}}(D) = \frac{3}{5} I(\underset{\text{fair}}{3}, 0) + \frac{2}{5} I(\underset{\text{excellent}}{0}, 2)$$