



# **CS 412 Intro. to Data Mining**


## **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

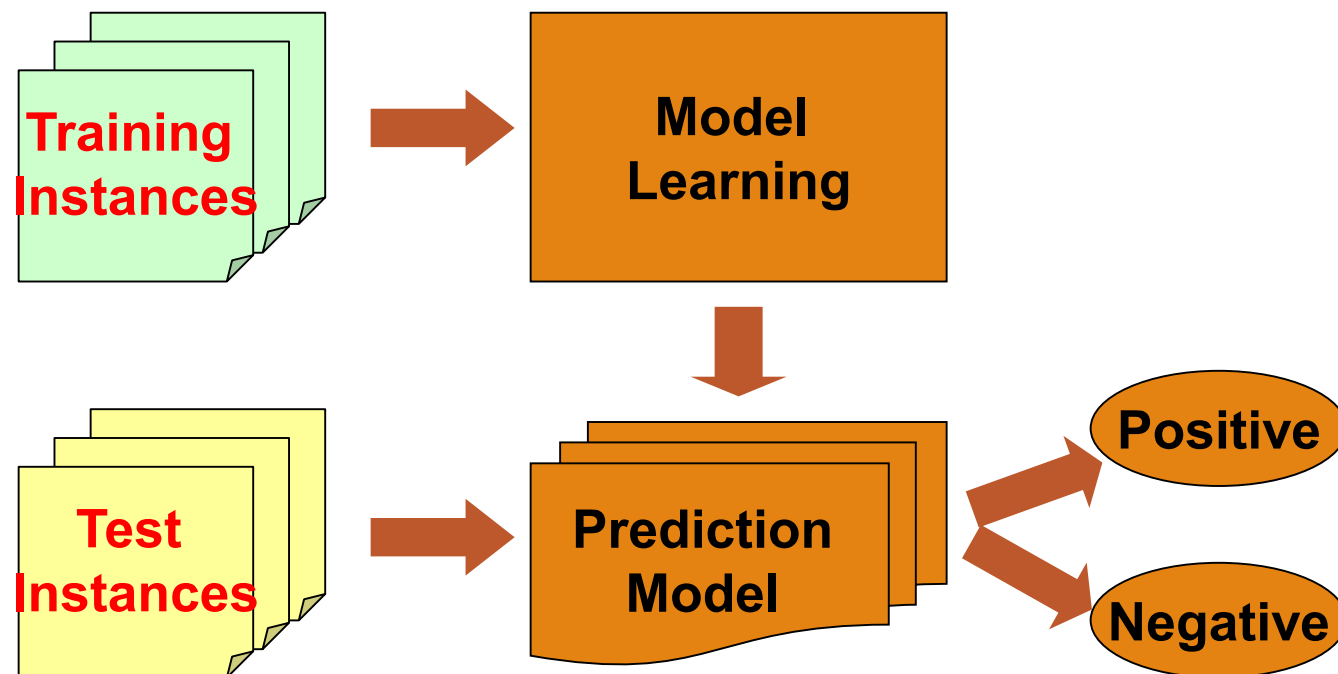
# Supervised vs. Unsupervised Learning (1)

## Supervised learning (classification)

- Supervision: <sup>มีคนสอน</sup> The <sup>มีจุดมุ่งหมายในการเรียน</sup> training data such as <sup>มีการสังเกต</sup> observations or <sup>วัดผล</sup> 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

Training Data with class label:

| 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)

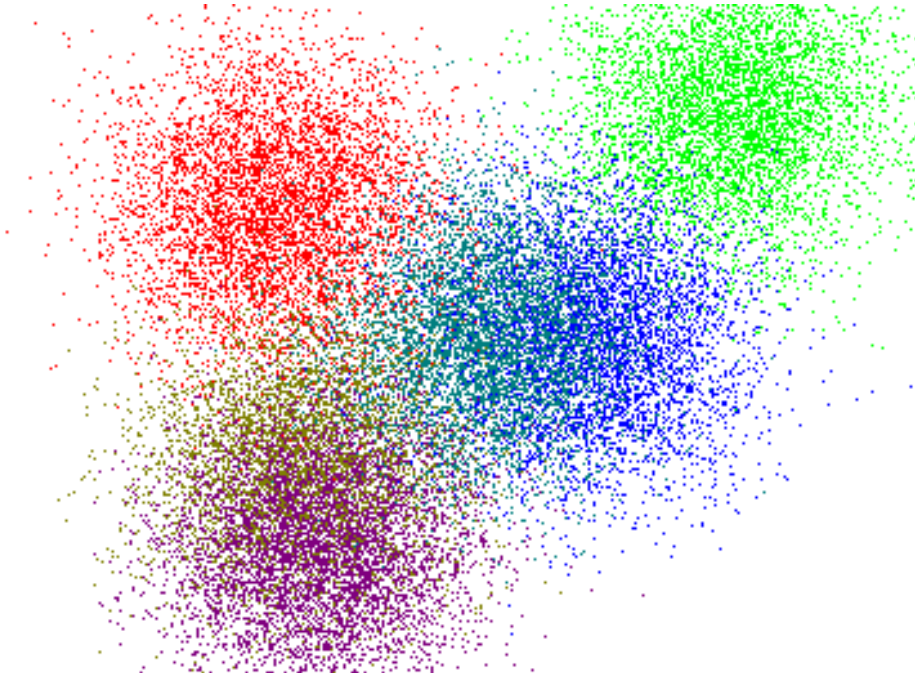
## □ Unsupervised learning (clustering)

- The class labels of training data are unknown
- Given a set of observations or measurements, establish the possible existence of classes or clusters in the data

Data มีแต่ X พิเจอร์ แบ่งกลุ่มเป็น

เหมือนกัน = ใกล้เคียง

ไม่เหมือนกัน = ไกล



# Prediction Problems: Classification vs. Numeric Prediction

## □ Classification ทำนายว่าอยู่ในกลุ่มไหน

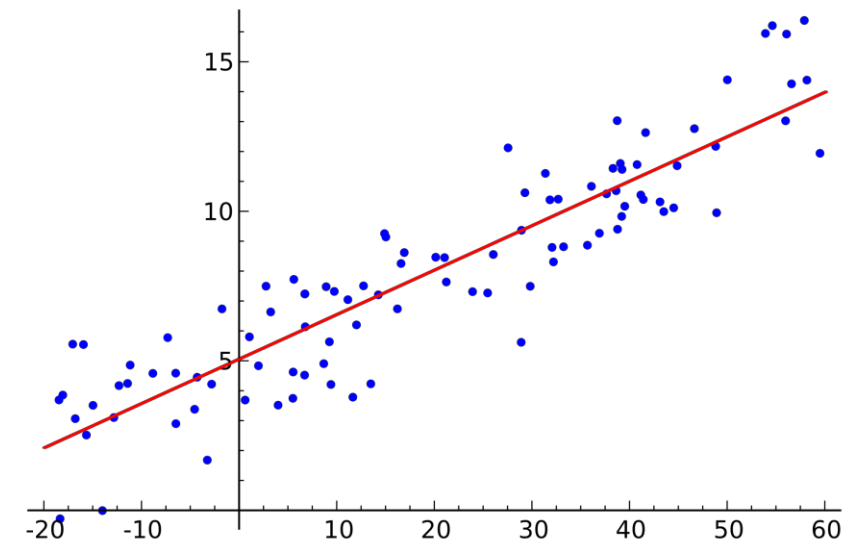
- Predict categorical class labels (discrete or nominal)
- Construct a model based on the training set and the **class labels** (the values in a classifying attribute) and use it in classifying new data

## □ 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**      สร้าง Model = ฝึกผล , ทายคำตอบ = ผ่าน = 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

ท่านไม่ตัดสินใจ

☐ Decision Tree Induction



☐ Bayes Classification Methods

☐ Linear Classifier

☐ Model Evaluation and Selection

☐ Techniques to Improve Classification Accuracy: Ensemble Methods

☐ Additional Concepts on Classification

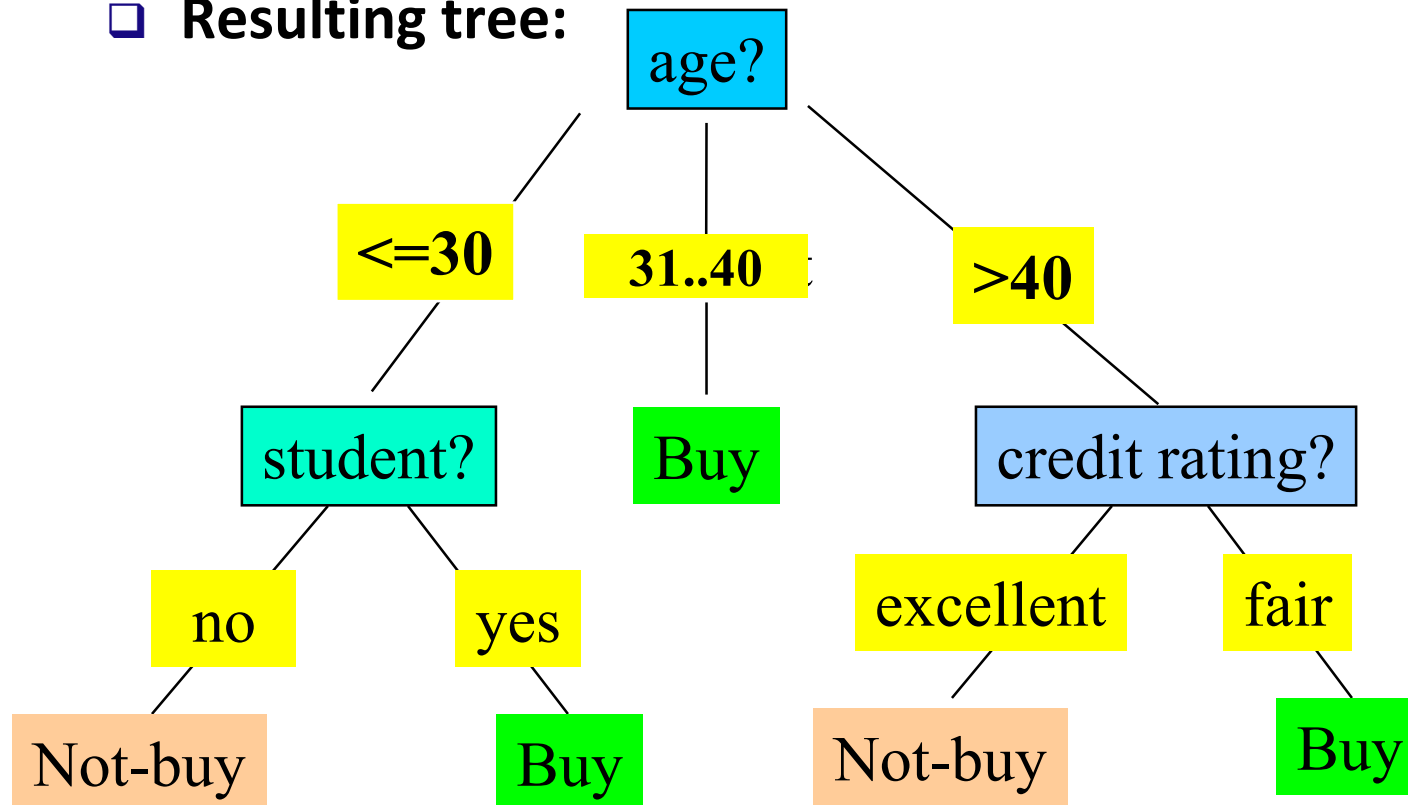
☐ Summary

# Decision Tree Induction: An Example

## 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



# 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  $\{y_1, y_2, \dots, y_m\}$

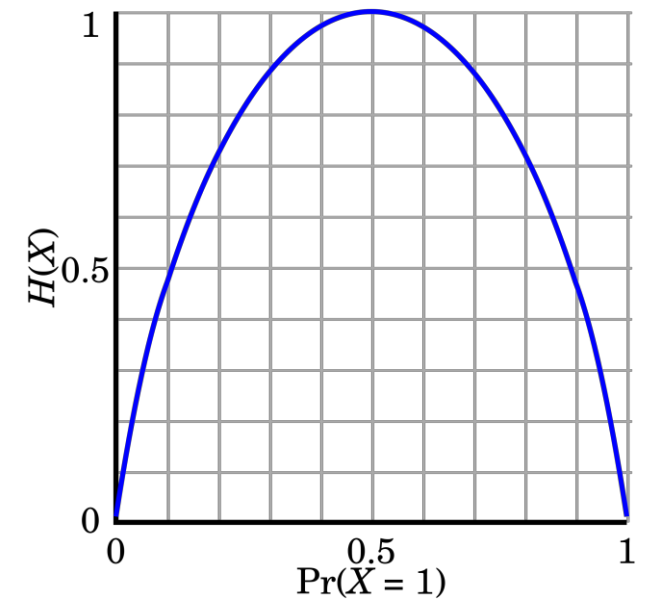
$$H(Y) = - \sum_{i=1}^m p_i \log(p_i) \quad \text{where } p_i = P(Y = y_i)$$

## □ 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  $A$

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

# 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 พิจารณาค่า Gain มากสุด เป็น root node

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

Similarly, we can get

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

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