Machine Learning Basics

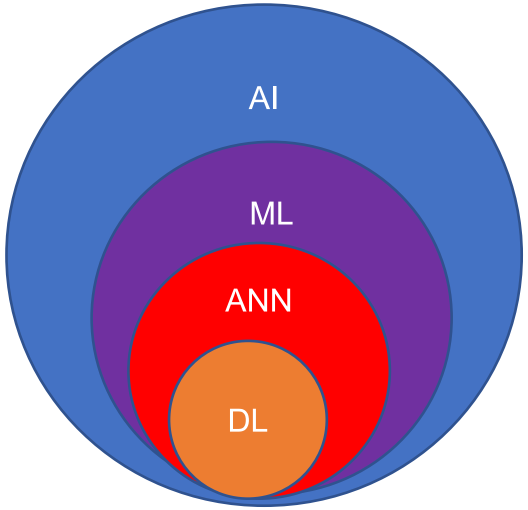
Jen-Ing Hwang
Department of Computer Science and
Information Engineering
Fu Jen Catholic University

Outline

1. Introduction
2. Dataset (資料集合、Database)
3. Types of Learning
4. The Learning Problem
5. Learning Models
6. Machine Learning Stages (階段：1. 訓練 2. 測試)

Introduction

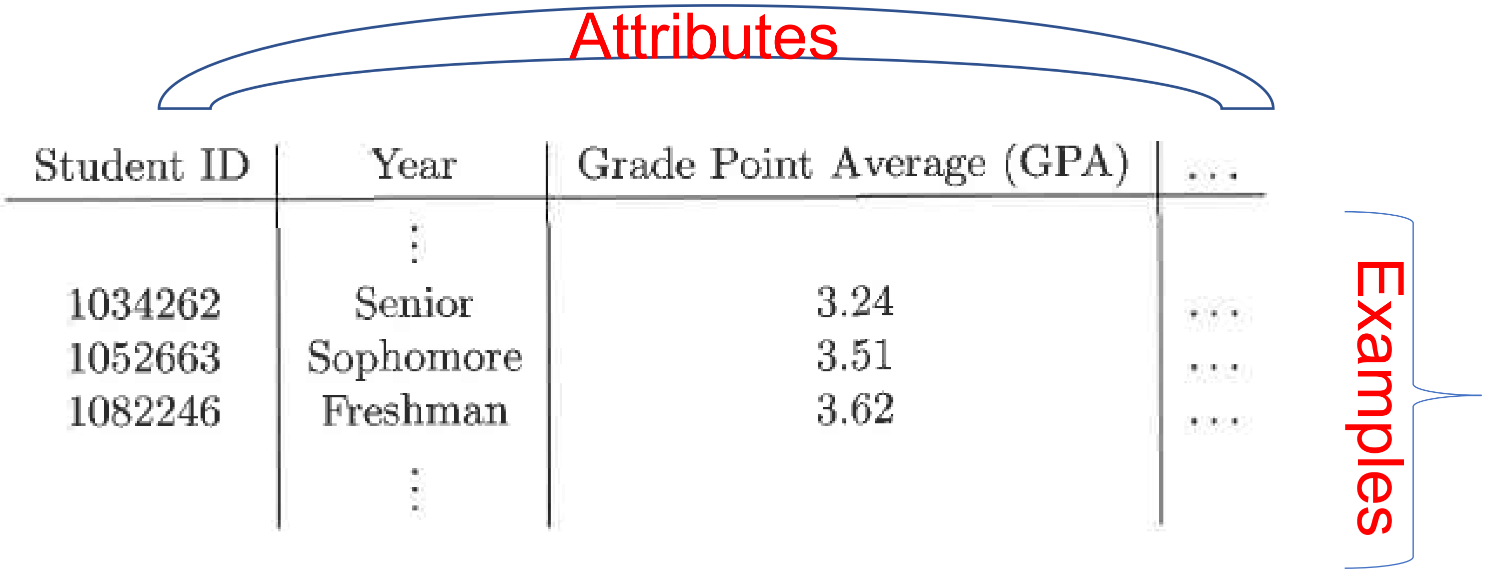
- Machine learning focuses on constructing learning algorithms that can learn from data to acquire knowledge.
- Machine learning是AI的子集合
- 給他data, 從中學習知識和概念
- Deep Learning is part of Machine Learning, in which we use artificial neural networks models.
- Machine learning 讓電腦變聰明的方法(用演算法)



AI (Artificial Intelligence)
ML (Machine Learning)
ANN (Artificial Neural Networks)
DL (Deep Learning)

Dataset

- Dataset: a collection of data examples and their attributes
- 行(Column): 同一個屬性 來描述一個物件的某個特徵
- 列(Object)：代表一個物件、一個example



Attributes

Student ID	Year	Grade Point Average (GPA)	. . .
	⋮		
1034262	Senior	3.24	. . .
1052663	Sophomore	3.51	. . .
1082246	Freshman	3.62	. . .
	⋮		

Examples

Figure adapted from Table 2.1 in Ref. [1]

Attribute and Example

- Attribute (= feature, variable, characteristic, dimension , field)
  - a characteristic that explains the example
  - 可以被分為input and output (label, target)
- Example(= instance, object, sample, entity , record)
  - described by a group of related attributes

Dataset I: Numeric(有連續性) Input Attributes

例如溫度是 32 度

2 Numeric Input Attributes

output labels

Example (不是屬性)	Price (US Dollar)	Engine Power (Horsepower)	Family Cars
1	30,000	150	Yes
2	23,000	120	Yes
3	45,000	200	No
4	34,000	140	Yes
5	12,000	70	No

5 Examples

Dataset II: Nominal(名義) Input Attributes

例如溫度可能有三種，hot, cool 跟 normal

6 Nominal Input Attributes

output labels

Example	Price	Engine Power	Maintenance	Persons	Trunk Size	Safety	Family Cars
1	Medium	Moderate	Low	4&more	Big	High	Yes
2	Medium	Moderate	High	4&more	Big	High	Yes
3	High	Powerful	High	4&more	Big	Low	No
4	Medium	Moderate	Low	4&more	Small	Low	Yes

4 Examples

Types of Attributes

- Nominal values (名目式)
  - Nominal = 用文字來表示它的值
  - related to names, states or symbols
  - a finite(有限的) number of states
  - also referred to as symbolic values, categorical values, or discretized values
  - Examples: ID, name, and color
- Numeric values (數值) <- 比較適合Deep learning
  - a measurable quantity
  - represented in integer or real values
  - an infinite(無限的) number of values
  - also referred to as continuous values(連續的屬性值)
  - Example: length, weight, and temperature

量化是最簡單可以去呈現資料的方法，深度學習的 Input 都是數值

Types of Learning (1/2)

- Supervised Learning (我給你答案 我監督你學習) ->已成熟
  - 給的資料一定有Input和Output (會有對應的輸出結果->好像你有給它標準答案)
    - EX: 什麼樣的價格和Engine Power 會是Family Car
  - Training data with output labels (Output可以是數值也可以是名目)
    - Classification(分類): output labels are **categories**, e.g. “disease” and “no disease” -> Output是名目式的
    - Regression(迴歸): output labels are numeric values, e.g. height -> Output是數值式的
- Unsupervised Learning(你沒有標準答案 沒有Output)
  - Training data without output labels
    - Clustering(分群): grouping similar instances
- Reinforcement Learning(第一時間沒有Input 但之後會有獎賞或處罰)
  - Training algorithms receive no supervised output labels but use delayed reward(下贏) and punishment(下輸) to learn best actions -> ex:電腦下棋
- 分類 V.S. 分群
  - 分類有Output ->已經知道答案 來做分類
  - 分群沒有Output ->我沒有很明確的定義 但我希望你可以幫我判斷哪些人可以在一起
    - ex: Poker分群 (依顏色分兩群 or 依花色分四群 or 依人的圖樣跟數字分成兩群)

Types of Learning (2/2)

像不像(ex:身高體重)，用XY散佈圖的距離來判斷

Supervised Learning

Data: (x, y)
x is data, y is label

Goal: Learn function to map
$x \to y$

Apple example:

This thing is an apple.

Unsupervised Learning

Data: x
x is data, no labels!

Goal: Learn underlying structure

Apple example:

This thing is like the other thing.

Reinforcement Learning

Data: state-action pairs

Goal: Maximize future rewards over many time steps

Apple example:

Eat this thing because it will keep you alive.

學習對應關係
我有蘋果跟橘子
你沒有告訴我橘子和蘋果的規則
但你告訴我什麼是橘子和蘋果
有三個屬性 形狀 顏色 重量
蘋果和橘子的形狀和重量很像
能決定差異的可能是顏色
你沒有事先跟電腦說
只是把資料餵給它
他要自己想辦法找出這個規則

The Learning Problem

- Components of Learning
  - Input
  - Output
  - Target function
  - Data
  - Hypothesis

Example of Machine Learning

- Credit card Approval

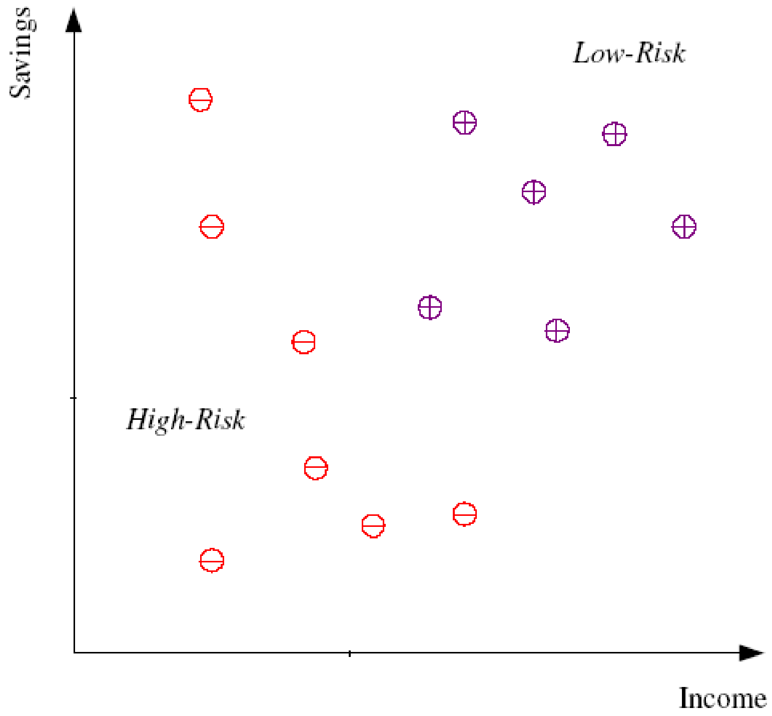
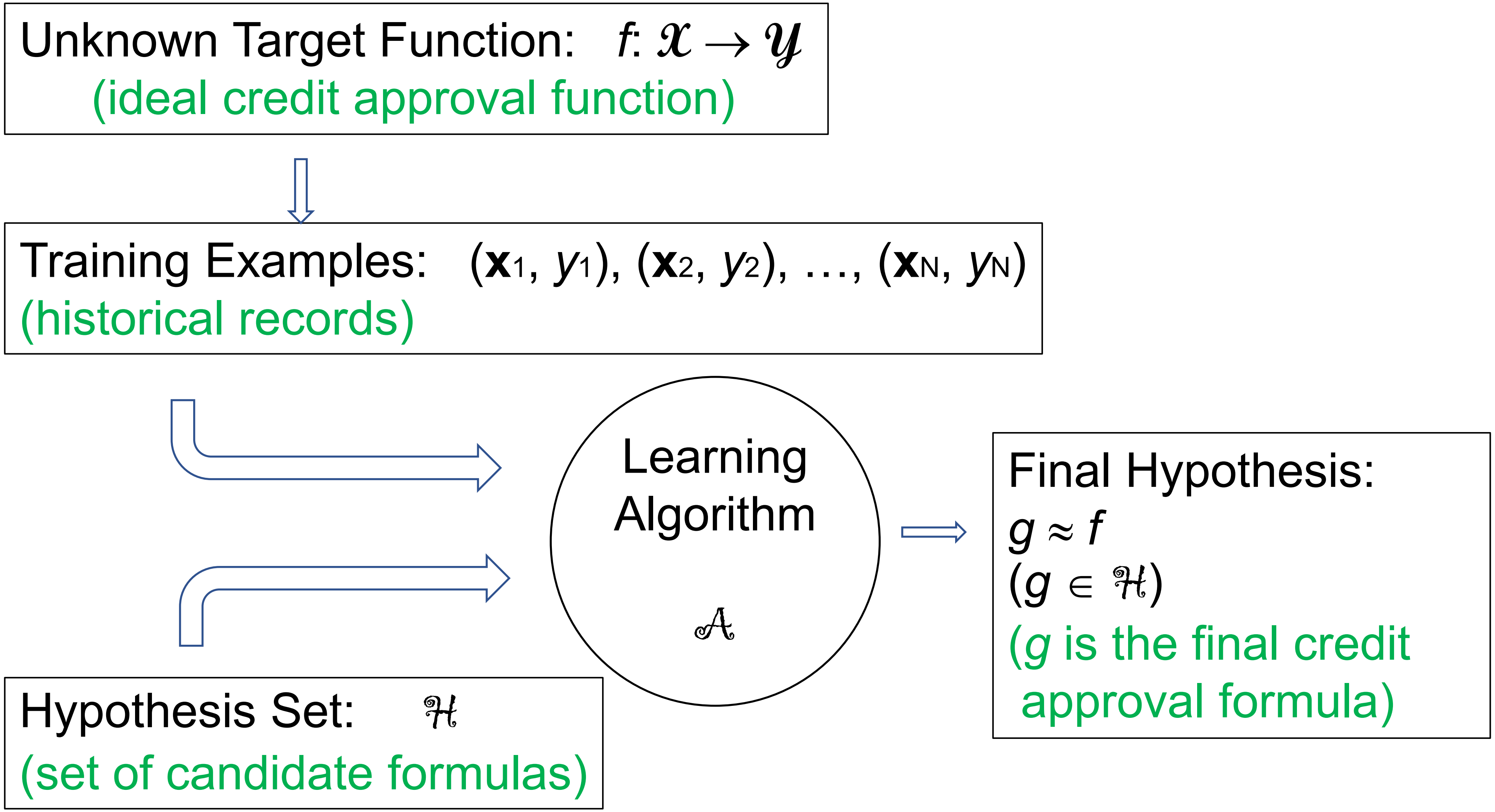


Figure adapted from Fig 1.1 in Ref. [2]

Components of Learning

- Input: $\mathbf{x} = \begin{bmatrix} \text{Income} \\ \text{Savings} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ (customer application)
- Output: y (good/bad customer?)
- Target function: $f: \mathcal{X} \rightarrow \mathcal{Y}$

  (ideal credit approval function)
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$ (historical records)
- Hypothesis: $g: \mathcal{X} \rightarrow \mathcal{Y}$ (formula to be used)



Unknown Target Function: $f: \mathcal{X} \to \mathcal{Y}$
(ideal credit approval function)

Training Examples: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$
(historical records)

Learning Algorithm
$\mathcal{A}$

Hypothesis Set: $\mathcal{H}$
(set of candidate formulas)

Final Hypothesis:
$g \approx f$
$(g \in \mathcal{H})$
(g is the final credit approval formula)

Figure adapted from Fig 1.2 in Ref. [3]

Learning Models

Two solution components of the learning problem:

- The hypothesis set $\mathcal{H}$; $g \in \mathcal{H}$
- The learning algorithm $\mathcal{A}$

$\Rightarrow$ Together, they are referred to as the learning model.

Model I: Linear Model 找到直線

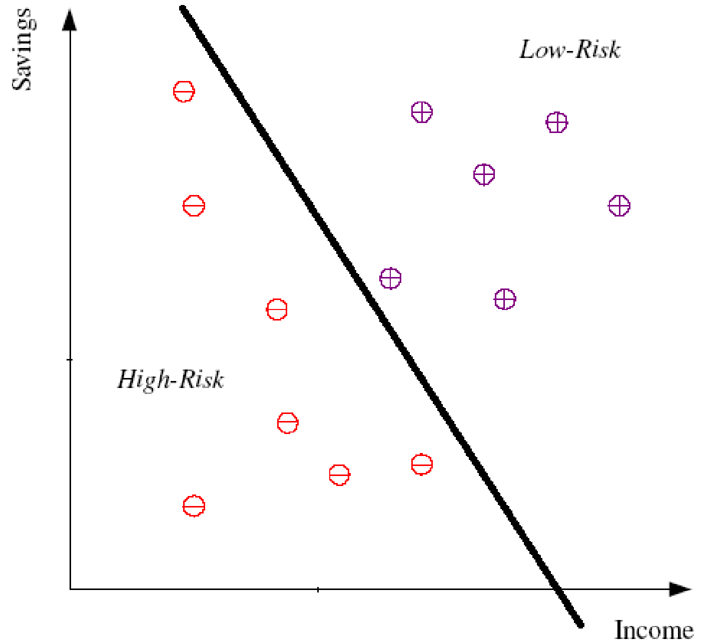
- The hypothesis set $\mathcal{H}$; $g \in \mathcal{H}$

  $\mathcal{H}$ = A set of linear functions

- The learning algorithm $\mathcal{A}$

  $\mathcal{A}$ = Perceptron Learning Algorithm (PLA)

⇒ Together, they are referred to as a linear model.



Model II: Rule-Based Model 找到規則

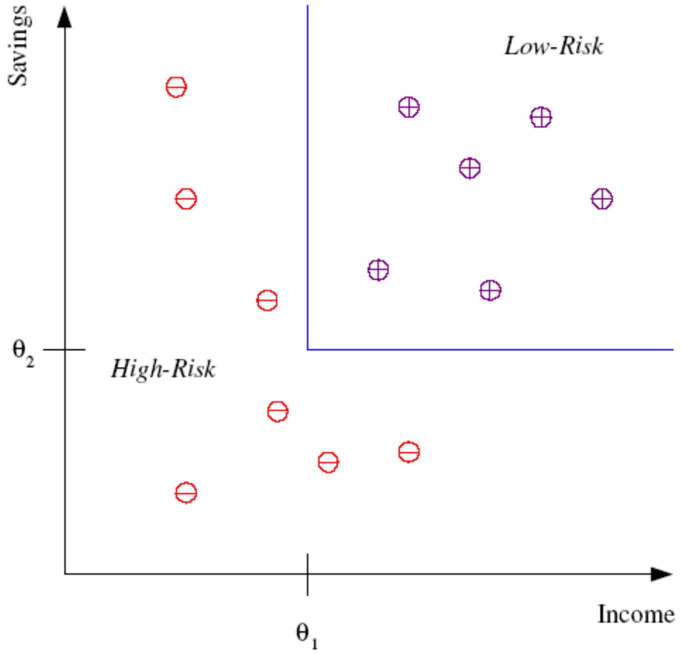
- The hypothesis set $\mathcal{H}$; $g \in \mathcal{H}$

  $\mathcal{H}$ = A set of rules

- The learning algorithm $\mathcal{A}$

  $\mathcal{A}$ = Decision tree learning (e.g., ID3)

⇒ Together, they are referred to as a rule-based model.



Model III: Instance-Based Model

不做事 要找到新的案例出來，才找類似的案例，把Data儲存起來，將來有案例才開始動

- The hypothesis set $\mathcal{H}$; $g \in \mathcal{H}$

  $\mathcal{H}$ = An implicit hypothesis set

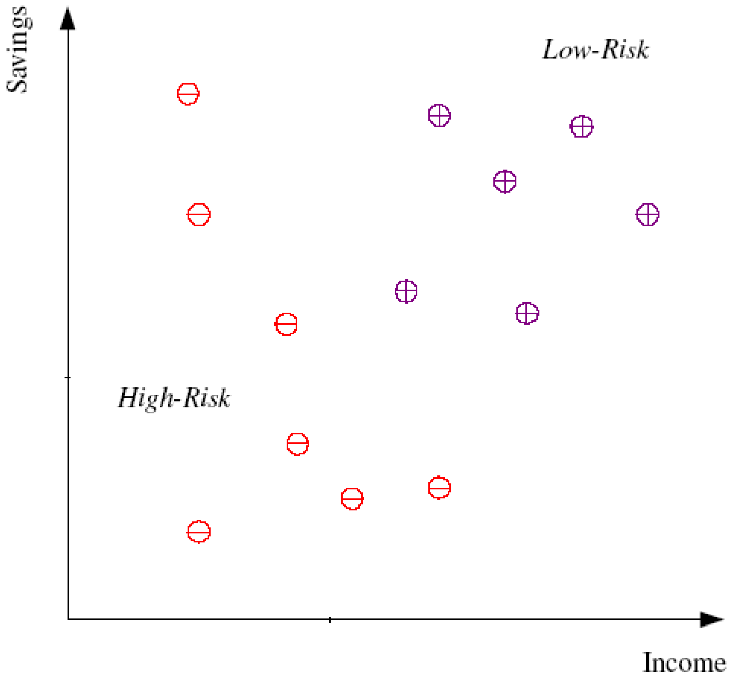
- The learning algorithm $\mathcal{A}$

  $\mathcal{A}$ = *K*-Nearest neighbor (*K*NN)(e.g., 1NN)

  看要看以前的幾個案例ex: 可以找兩個案例，可能之前有兩個情節跟這個案例很像，一個判二十年，一個判十八年，基於這兩個案例，那我就判十九年

  Note that KNN is a lazy learning approach that stores data and waits for the query before generalizing.

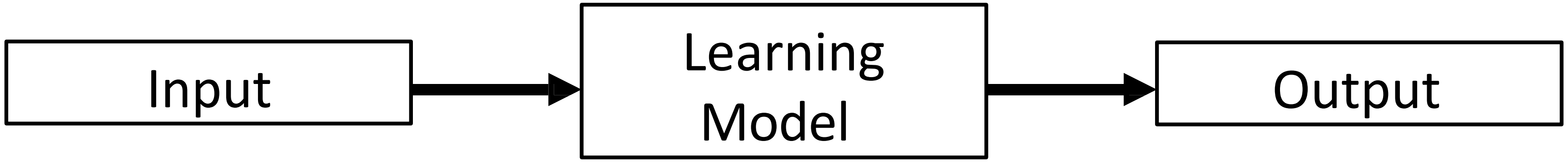
⇒ Together, they are referred to as a Instance-based model. ->基於以前的範例



基於以前的結果來判斷
Ex:類似的案例以前法官判二十年，這個案例就大約判二十年

Machine Learning Stages

Training Stage: Search a hypothesis to fit observed data.

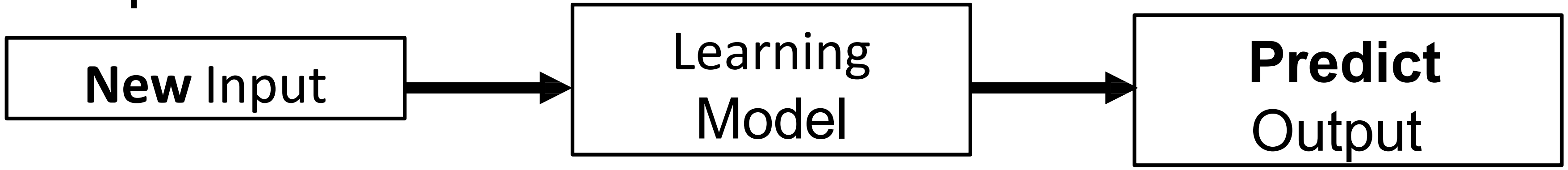


Input

Learning Model

Output

Testing Stage: Predict the output labels for unseen examples.



New Input

Learning Model

Predict Output

Model I: Linear Model

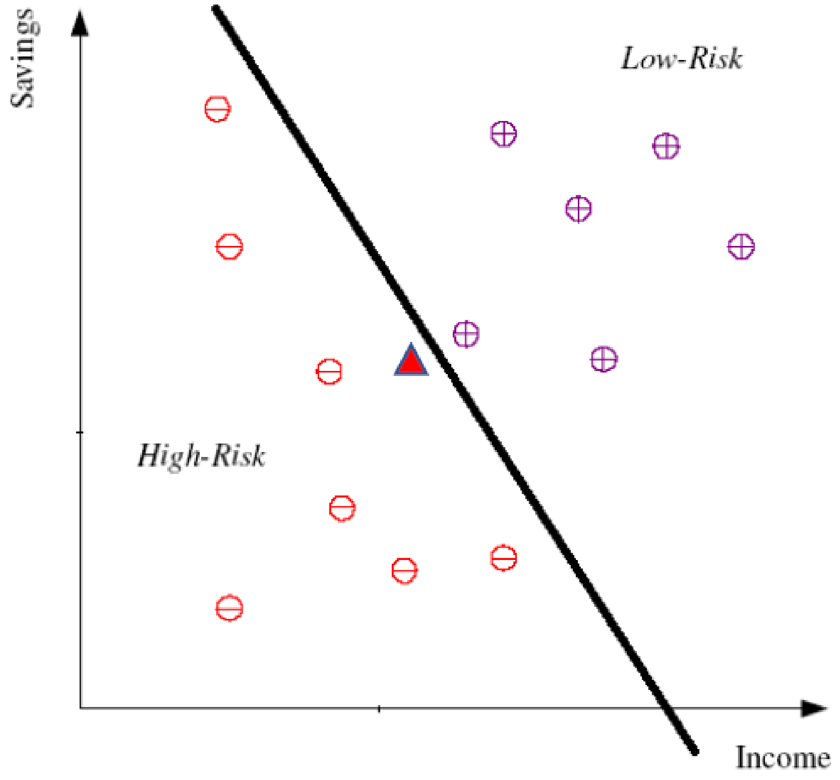
Training Stage:

Find an equation w0 +w1x1 + w2x2 = 0 to separate positive (⊕) and negative (⊖) examples.

Testing Stage:

Predict the new example (▲) as a negative (⊖) example.

不同的假說 不同的演算法
預測的答案會不一樣
Ex: 評審不一樣 得出來的結果也就
不一樣



Model II: Rule-Based Model

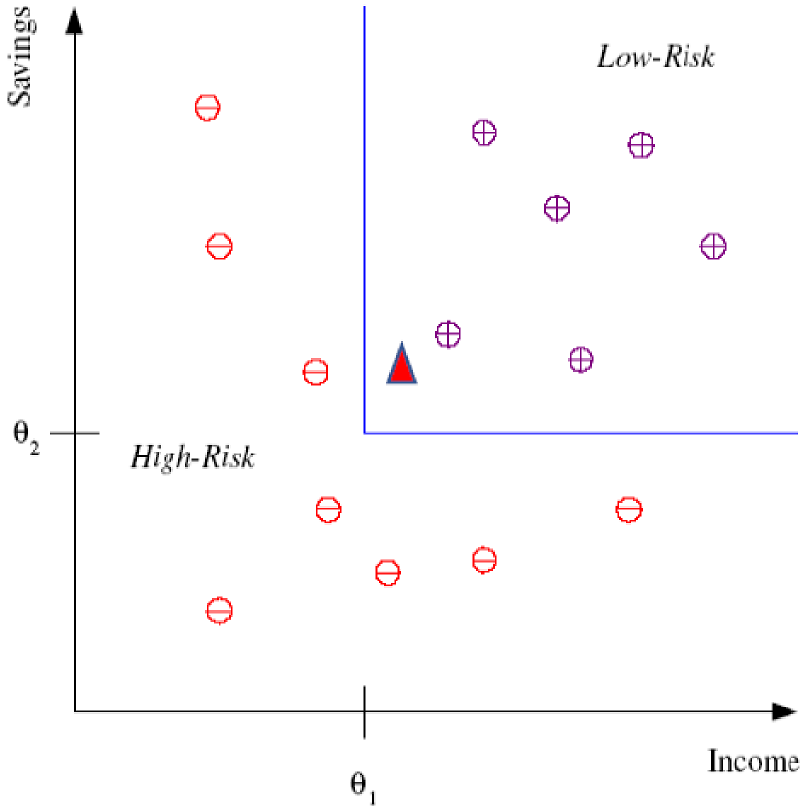
Training Stage:

Find a rule:

IF x1 > θ_1 AND x2 > θ_2
THEN positive (⊕) ELSE negative (⊖)

Testing Stage:

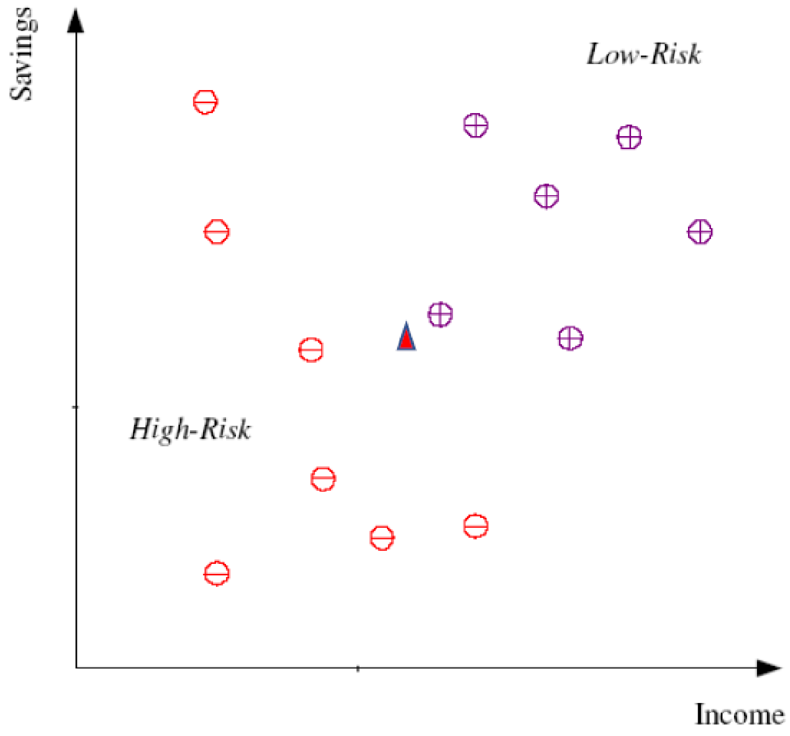
Predict the new example (▲) as a positive (⊕) example.



Model III: Instance-Based Model

Training Stage:

KNN is a lazy learning approach that stores data and waits for the query before generalizing.



Testing Stage:

Predict the new example (▲) as a positive (⊕) example if we apply 1NN.

Summary

- Machine learning is a branch of artificial intelligence focusing on optimizing a performance criterion using example data.
- Dataset: a collection of data examples and their attributes
- Types of learning:
  - Supervised Learning, Unsupervised Learning, and Reinforcement Learning
- The learning problem: Components of learning
  - Input, Output, Target function, Data, and Hypothesis
- The learning model:
  - The hypothesis set and the algorithm
  - There are numerous machine learning models.
- Machine learning stages are training and testing.

Kahoot

(T) 1. Deep Learning is part of machine learning, in which we use artificial neural networks models.

(F) 2. Artificial intelligence is a subfield of machine learning.

(F) 3. A nominal attribute assumes values that are integer or real.

(T) 4. A dataset is a collection of data examples and their attributes.

(F) 5. An unsupervised learning algorithm learns from a training dataset with both input attributes and output levels.

(T) 6. The machine learning stages are training and testing.

7. The output attributes of classification are:

(a) nominal values

(b) numeric values

(c) nominal or numeric values

(d) nominal and numeric values

8. Three types of machine learning are:

(a) deep learning, moderate learning, and shallow learning

(b) human learning, animal learning, and robotics learning

(c) supervised learning, unsupervised learning, and reinforcement learning

(d) visual learners, reading learners, and writing learners

9. Supervised learning can be divided into:

- (a) classification and clustering
- (b) classification and regression
- (c) clustering and regression
- (d) classification, clustering, and regression

10. What is the goal of clustering?

- (a) Group similar instances
- (b) Predict numerical values for testing examples
- (c) Reduce the number of features
- (d) None of the above

References

1. Introduction to Data Mining, 2nd edition, Pang-Ning Tan, Pearson, 2018.
2. Introduction to Machine Learning 3rd Edition, Ethem Alpaydin, 2014.
3. Learning from Data, Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin, AMLBook, 2012.