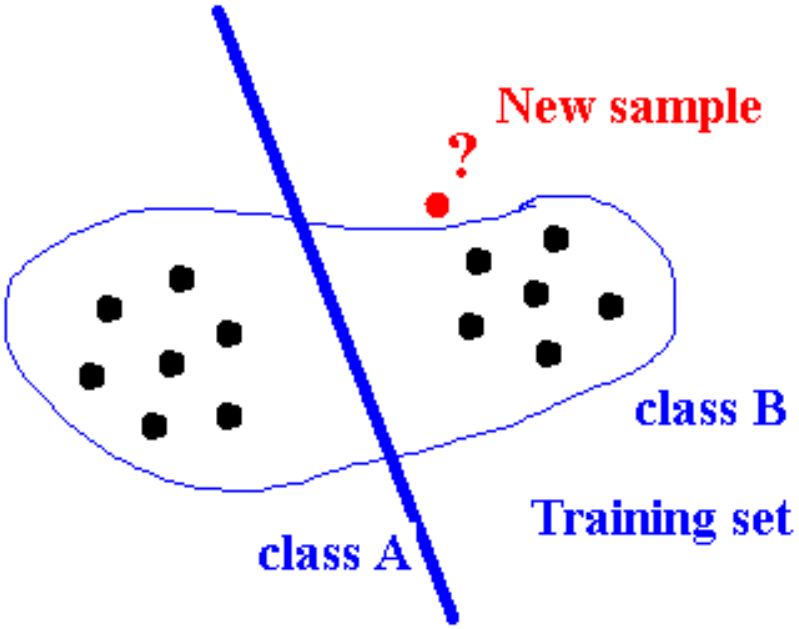
1. Tổng Quan về Phân Loại DL

- Các tình huống
  - Email: *"spam"* hay *"bình thường"*
  - Các giao dịch trực tuyến: *"gian lận"* hay *"thông thường"*
  - Y tế: *"bị bệnh"* hay *"không bị bệnh"*; khối u *"lành tính"* hay *"ác tính"*,...
  - Y ∈{0, 1}: 0: *"negative"* hay 1: *"positive"* classes
  - Y ∈{0, 1, 2, 3}: Dữ liệu thuộc nhiều lớp
  - Mỗi lớp có gắn một nhãn (label): Ví dụ *"spam"* hay *"not spam"*

1. Tổng Quan về Phân Loại DL



- Cho trước tập huấn luyện (training set), dẫn ra mô tả về class A và B
- Cho trước mẫu/đối tượng mới, xác định class mà mẫu đó thuộc về?
- Liệu class đó có thực sự phù hợp/đúng cho mẫu/đối tượng đó?

1. Tổng Quan về Phân Loại DL

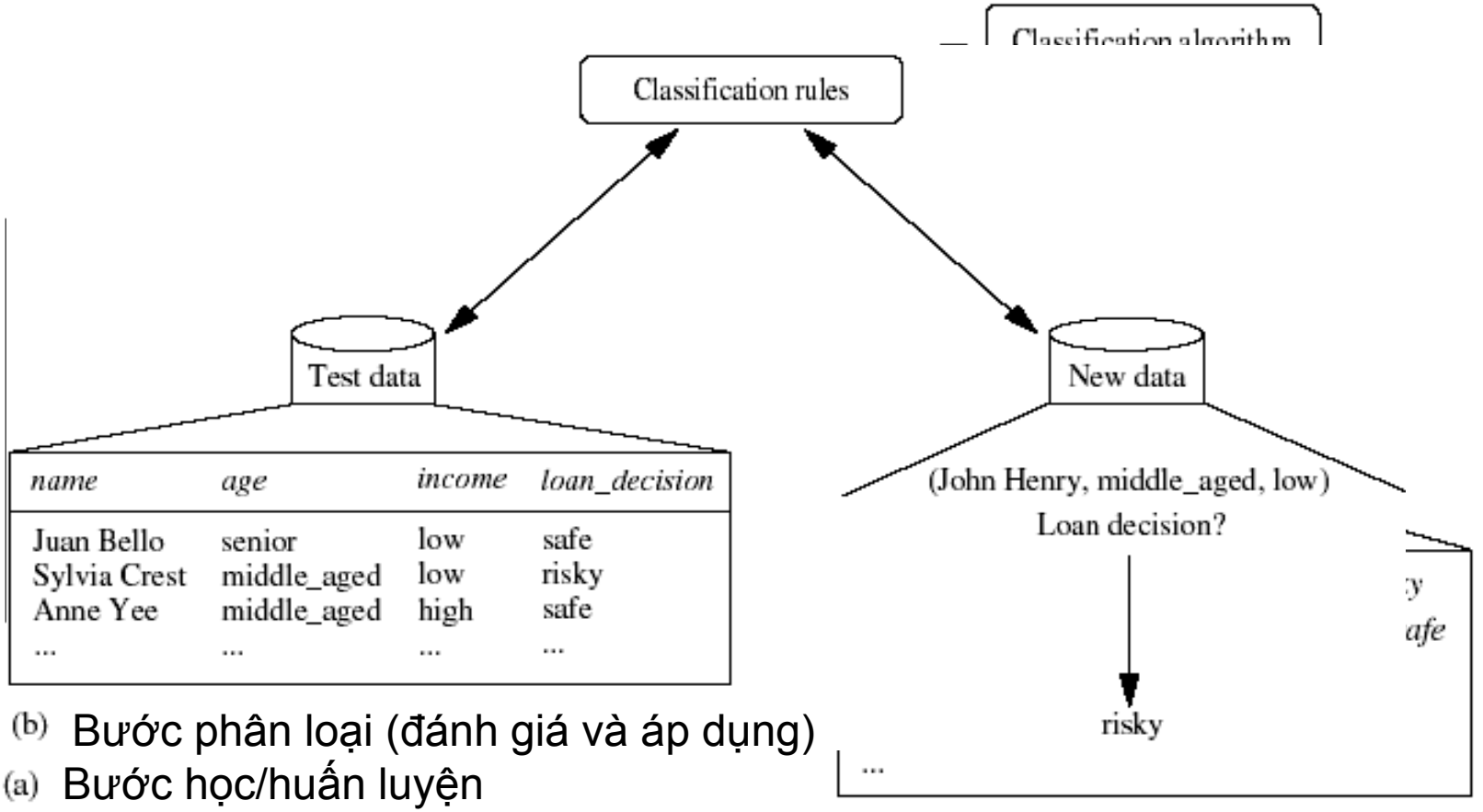
- Phân loại dữ liệu (classification)
  - Dạng phân tích dữ liệu nhằm rút trích các mô hình mô tả các lớp dữ liệu hoặc dự đoán xu hướng dữ liệu
  - Quá trình gồm hai bước:
    - Bước học (huấn luyện): xây dựng bộ phân loại (classifier) bằng việc phân tích (học) tập huấn luyện
    - Bước phân loại (classification): phân loại dữ liệu/đối tượng mới nếu độ chính xác của bộ phân loại được đánh giá là có thể chấp nhận được (acceptable)

y = f (X) với y là nhãn (phần mô tả) của một lớp (class) và X là dữ liệu/đối tượng

- **Bước học**: X trong tập huấn luyện, một trị y được cho trước với X → xác định f

- **Bước phân loại**: đánh giá f với (X', y') và X' <> mọi X trong tập huấn luyện; nếu chấp nhận được thì dùng f để xác định y'' cho X'' (mới)

1. Tổng Quan về Phân Loại DL

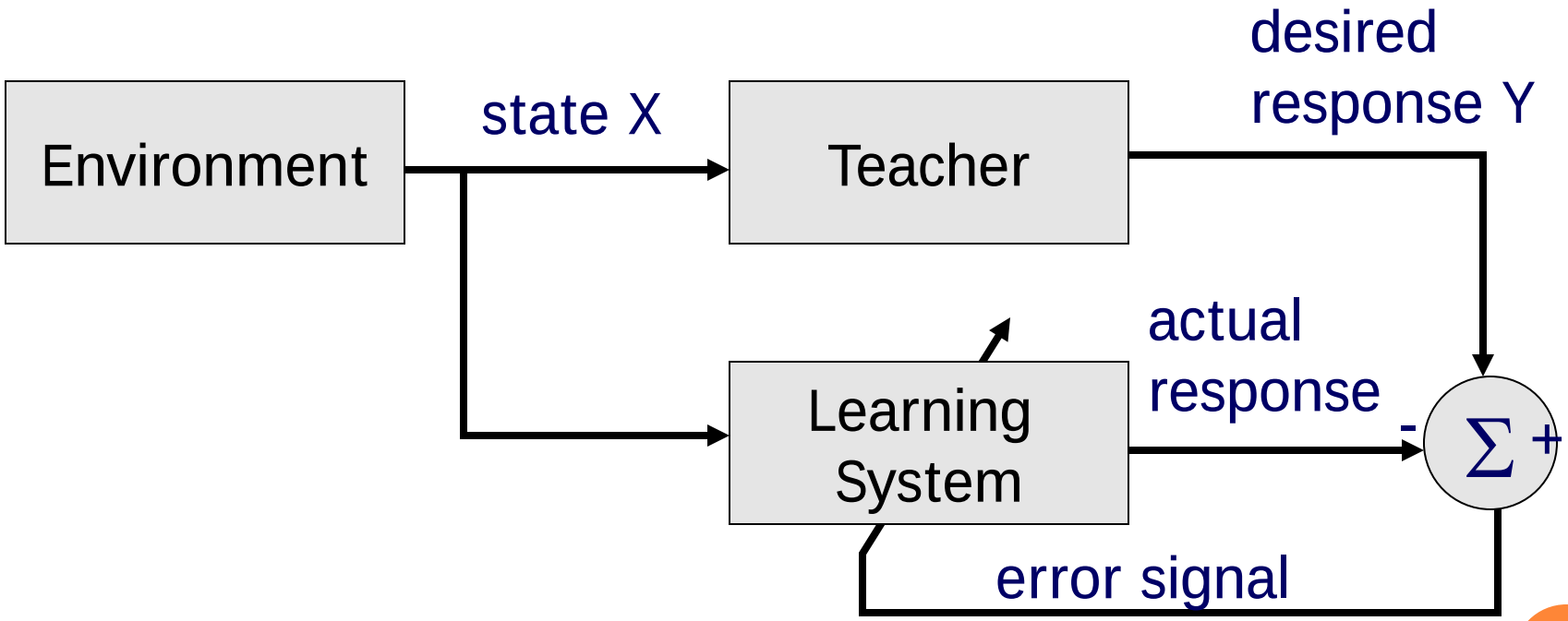


(b) Bước phân loại (đánh giá và áp dụng)

(a) Bước học/huấn luyện

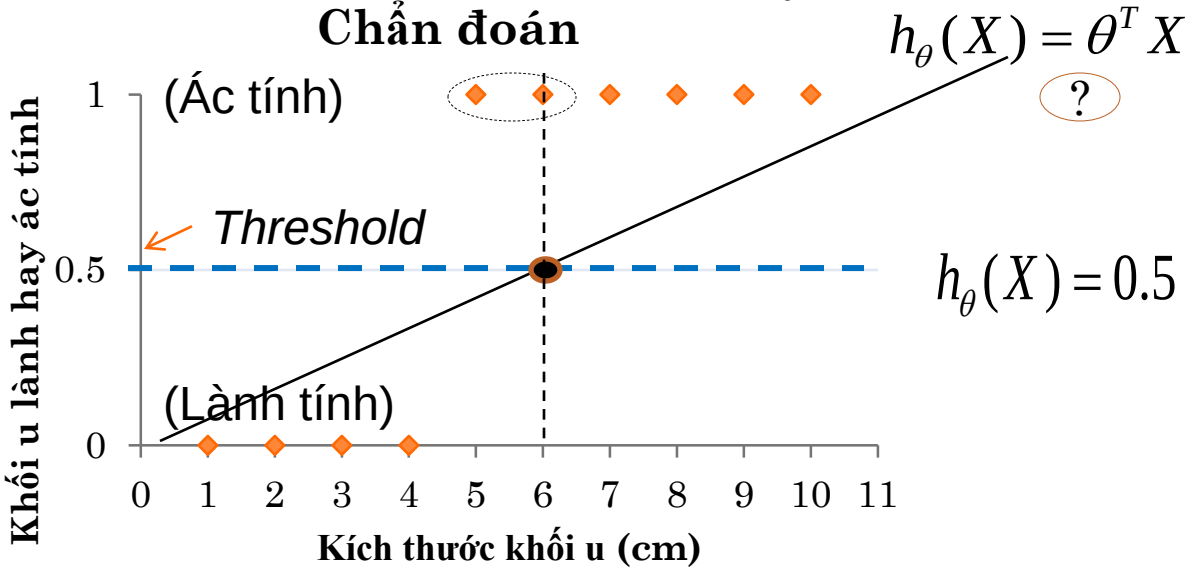
1. Tổng Quan về Phân Loại DL

- Phân loại dữ liệu: Dạng học có giám sát (supervised learning)



1. Tổng Quan về Phân Loại DL

- Phân loại khối u lành hay ác tính dựa vào kích thước



- Nếu $h_\theta(X) \geq 0.5$ thì dự đoán: "Y=1" ngược lại "Y=0"
- Thực tế, $h_\theta(X) > 1$ hoặc $h_\theta(X) < 0$
- *Hồi qui logistic (Logistic regression)* $0 \leq h_\theta(X) \leq 1$

=> *Phân loại (Classification)*

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

1. Tổng Quan về Phân Loại DL

- Các giải thuật phân loại dữ liệu
  - Hồi qui logistic (logistic regression)
  - Cây quyết định (decision tree)
  - Mạng Bayesian
  - Mạng neural nhân tạo (ANN)
  - Phân loại với k phần tử cận gần nhất (k-nearest neighbor)
  - Phân loại với suy diễn dựa trên tình huống (case-based reasoning)
  - Phân loại dựa trên tiến hoá gen (genetic algorithms)
  - Phân loại với lý thuyết tập thô (rough sets)
  - Phân loại với lý thuyết tập mờ (fuzzy sets) …

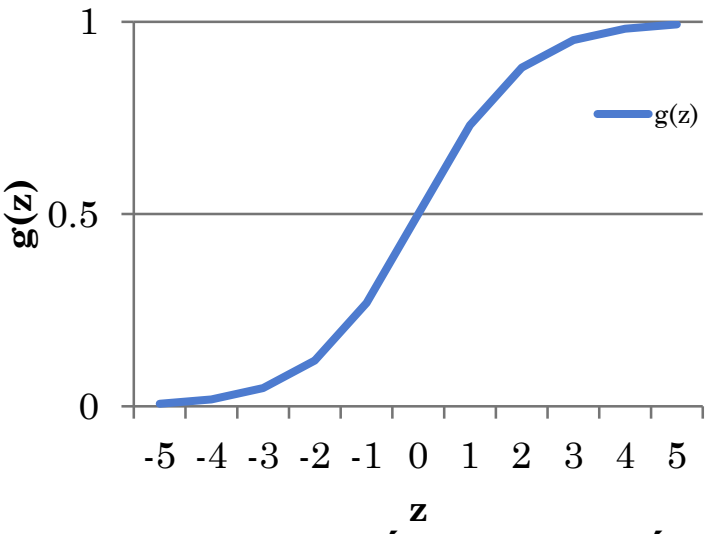
2. PHÂN LOẠI DL VỚI HỒI QUI LOGISTIC

- $h_\theta(X)=\theta^T X$ *(có thể lớn hơn 1* hoặc *nhỏ hơn 0)*
- Cần có $h_\theta(X)$ sao cho $0 \le h_\theta(X) \le 1$
- Mô hình lại: $h_\theta(X)=g(\theta^T X)$ với $g(z)=\dfrac{1}{1+e^{-z}}$

$$h_\theta(X)=\frac{1}{1+e^{-\theta^T X}}$$

- Sigmoid function

hay Logistic function



Liên quan đến bộ tham số θ

2. PHÂN LOẠI DL VỚI HỒI QUI LOGISTIC

- Giải thích giá trị của $h_\theta(X) = \frac{1}{1+e^{-\theta^T X}}$
  - Là xác suất dự đoán rằng "y=1" với input là x
  - Ex., $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ kt.khoi\_u \end{bmatrix}$

  $h_\theta(x)=0.7$

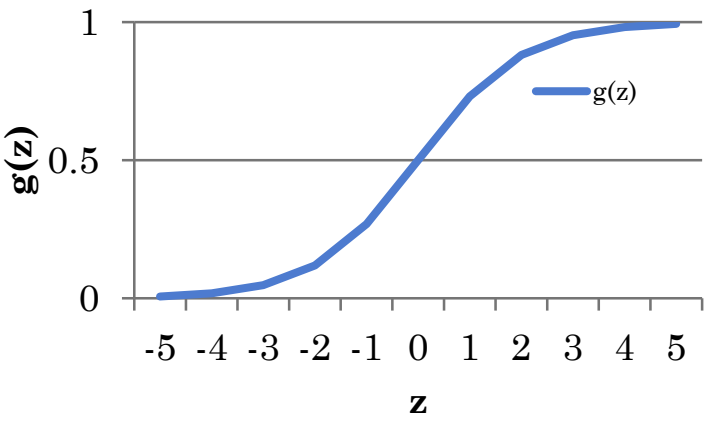
⇒ 70% khối u với k.thước đã cho có thể là ác tính

⇒ $h_\theta(x)=P(y=1|x;\theta)$ *(xác suất y=1, với x cho trước và được thông số hóa bởi θ)*

2. PHÂN LOẠI DL VỚI HỒI QUI LOGISTIC

- Chú ý $h_\theta(X)=g(\theta^T X)$ với $g(z)=\frac{1}{1+e^{-z}}$ hay

$$h_\theta(X)=\frac{1}{1+e^{-\theta^T X}}$$



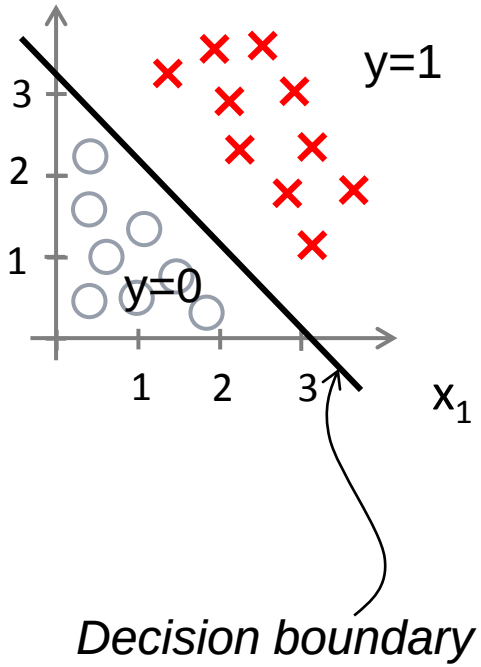
- g(z) ≥ 0.5 khi z ≥ 0
- g(z) <0.5 khi z <0

- dự đoán y=1 khi $h_\theta(X) \geq 0.5$ hay $\theta^T X \geq 0$
- dự đoán y=0 khi $h_\theta(X) < 0.5$ hay $\theta^T X < 0$

2. PHÂN LOẠI DL VỚI HỒI QUI LOGISTIC

- Đường phân lớp (decision boundary)
  - $h_\theta(X)=g(\theta^T X)=g(\theta_0+\theta_1 x_1+\theta_2 x_2)$
  - *Chọn*

$$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$



- Dự đoán "y=1" nếu $\theta^T X \geq 0$

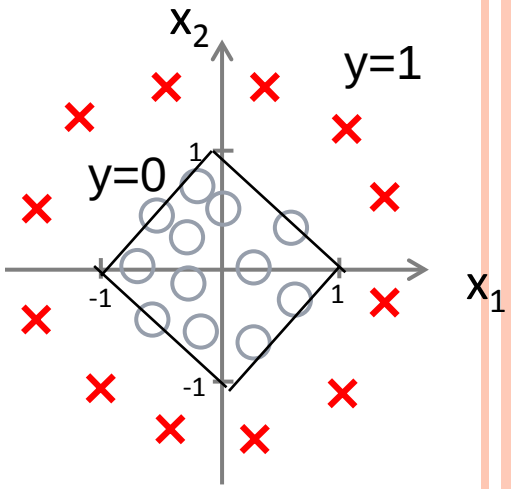
  hay $-3+x_1+x_2 \geq 0$

  $=> x_1+x_2 \geq 3$

2. PHÂN LOẠI DL VỚI HỒI QUI LOGISTIC

- Đường phân lớp (decision boundary)
  - $h_\theta(X)=g(\theta^T X)=g(\theta_0+\theta_1 x_1+\theta_2 x_2+\theta_3 x^2_1+\theta_4 x^2_2)$
  - Dự đoán "y=1" nếu $\theta^T X \geq 0$

    hay $-1+x^2_1+x^2_2 \geq 0$



2. PHÂN LOẠI DL VỚI HỒI QUI LOGISTIC

- Cost function của hồi qui logistic
  - Training set; $\{(x^{(1)},y^{(1)}), (x^{(2)},y^{(2)}),\cdots, (x^{(N)},y^{(N)})\}$
  - N examples

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}; x_0 = 1; y \in \{0,1\}$$

$$h_\theta(X) = \frac{1}{1+e^{-\theta^T X}}$$

  - Làm sao để chọn bộ thông số θ?

2. PHÂN LOẠI DL VỚI HỒI QUI LOGISTIC

- Liên hệ với hồi qui tuyến tính $J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(h_\theta(x^{(i)}) - y^{(i)})^2$

- Trong hồi qui phi tuyến

  $J(\theta)=cost(h_\theta(x),y)$

  $cost(h_\theta(x^{(i)}),y^{(i)}) = ½(h_\theta(x^{(i)})-y^{(i)})^2$

Để đơn giản hóa ta ghi

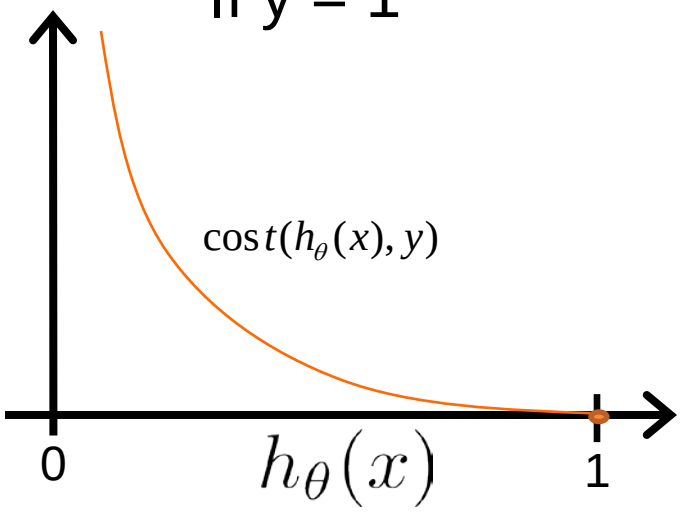
$cost(h_\theta(x),y) = ½(h_\theta(x)-y)^2$

2. PHÂN LOẠI DL VỚI HỒI QUI LOGISTIC

- Cost function của hồi qui logistic

$$\cos t(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) : y = 1 \\ -\log(1 - h_\theta(x)) : y = 0 \end{cases}$$

If y = 1



- Cost = 0 nếu y=1, $h_\theta(x)=1$
- Khi $h_\theta(x)$->0 thì cost -> ∞

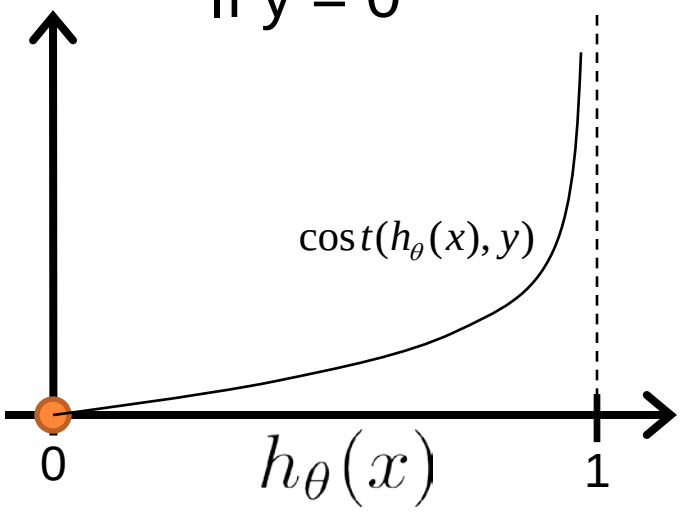
⇒Khi $h_\theta(x)=0$ có nghĩa là ta dự đoán rằng:
P(y=1|x, θ)=0, trong khi đó y=1, do vậy chi phí của giải thuật trong trường hợp này là rất lớn

2. PHÂN LOẠI DL VỚI HỒI QUI LOGISTIC

- Cost function của hồi qui logistic

$$\cos t(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) : y = 1 \\ -\log(1 - h_\theta(x)) : y = 0 \end{cases}$$

If y = 0



- Cost = 0 nếu y=0, $h_\theta(x)=0$
- Khi $h_\theta(x)$->1 thì cost -> ∞

⇒Khi $h_\theta(x)=1$ có nghĩa là ta dự đoán rằng:
P(y=1|x, θ)=1, trong khi đó y=0, do vậy chi phí của giải thuật trong trường hợp này là rất lớn

2. Phân Loại DL với Hồi Qui Logistic

- Đơn giản hàm chi phí và giải thuật gradient descent

$$\cos t(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) : y = 1 \\ -\log(1 - h_\theta(x)) : y = 0 \end{cases}$$

- Do y=0|1, nên hàm chi phí có thể đơn giản hóa như sau

$$cost(h_\theta(x), y) = -y\log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))$$

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} \cos t(h_\theta(x^{(i)}) - y^{(i)}) = -\frac{1}{N}\sum_{i=1}^{N} y^{(i)} \log h_\theta(x^{(i)}) + (1-y)\log(1-h_\theta(x^{(i)}))$$

- Thực hiện tìm $\min_\theta J(\theta)$ ta sẽ tìm được bộ thông số θ (giải thuật gradient descent)
- Để dự doán giá trị của y dựa vào giá trị x (mới đưa vào): $h_\theta(X) = \frac{1}{1 + e^{-\theta^T X}}$

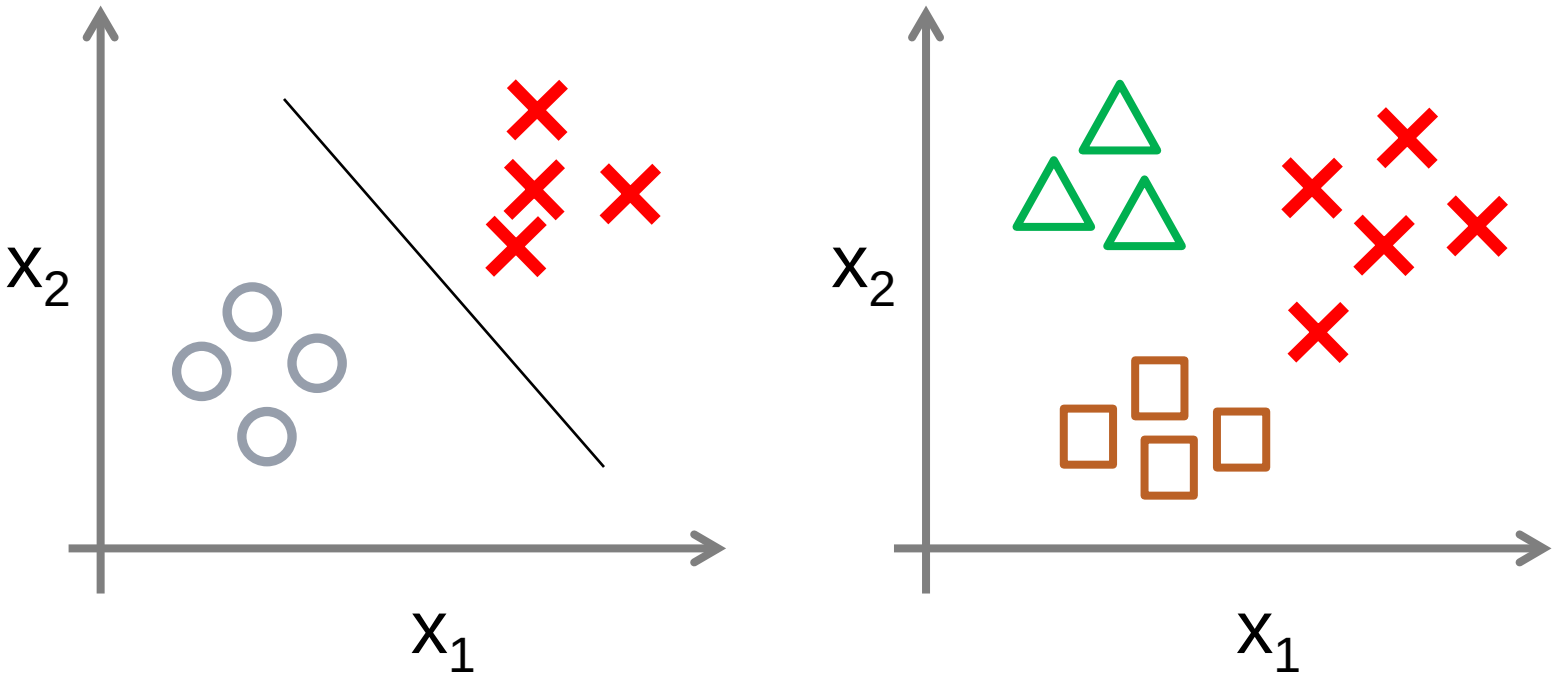
Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

2. Phân Loại DL với Hồi Qui Logistic

- Hồi qui logistic để phân lớp tập dữ liệu có nhiều nhãn (thuộc nhiều lớp):
  - Thư mục email: *"business", "friend", "family", "hobby"* (y={1,2,3,4})
  - Chẩn đoán: "cảm cúm", "sốt siêu vi", "rubella" (y={1,2,3})
  - Dự báo thời tiết: "nắng", "nhiều mây", "mưa" (y={1,2,3})

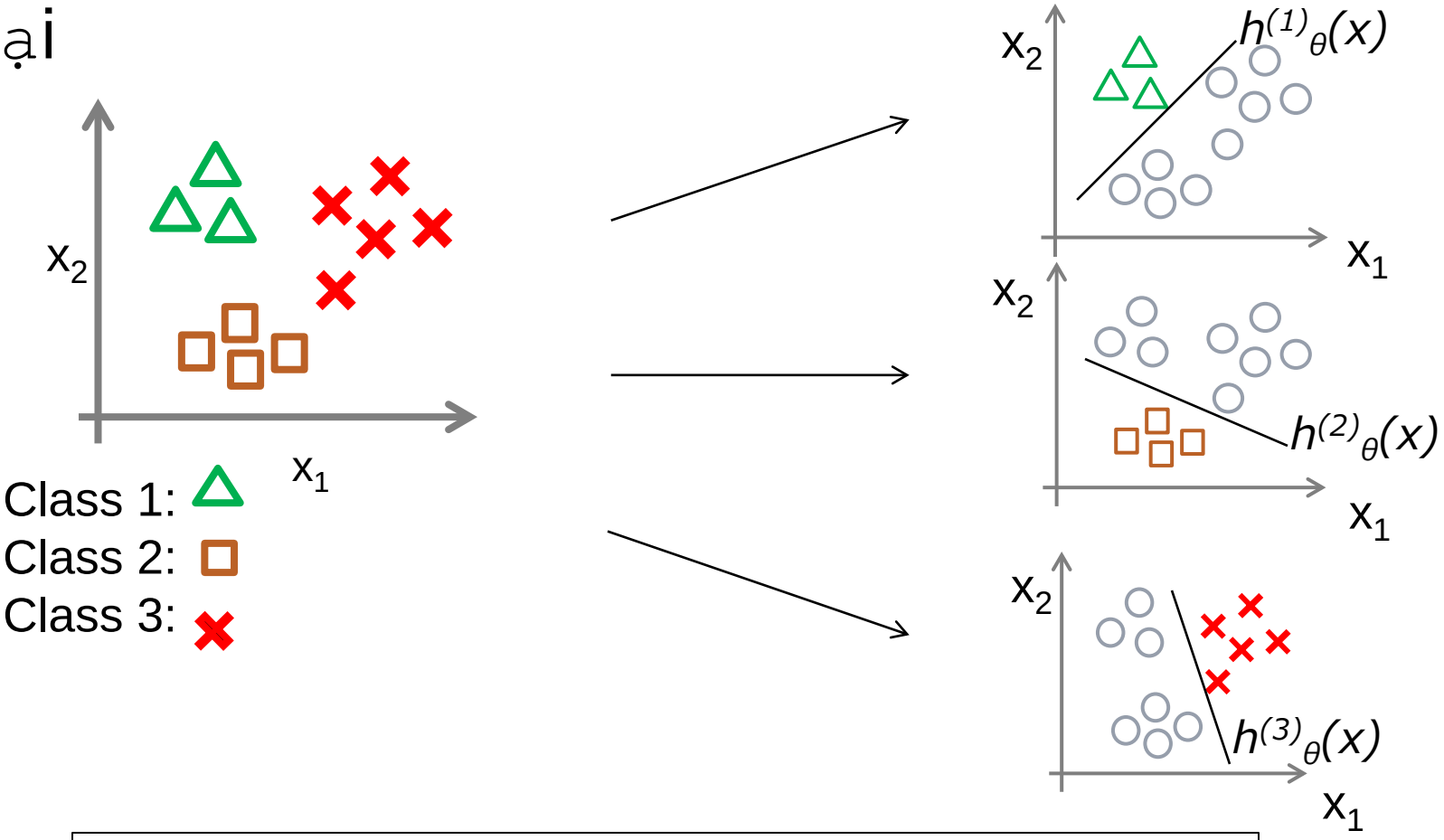
2. Phân Loại DL với Hồi Qui Logistic

- Dữ liệu đa lớp (multi-class)



2. PHÂN LOẠI DL VỚI HỒI QUI LOGISTIC

- Dữ liệu đa lớp (multi-class): một và phần còn lại



$h^{(i)}_\theta(x)=P(y=1|x;\theta)$ với $(i=1,2,..k)$, k là số lớp

2. Phân Loại DL với Hồi Qui Logistic

- Huấn luyện bộ phân lớp hồi qui logistic $h^{(i)}_{\theta}(x)$ cho mỗi lớp i
- Với một phần tử x mới đưa vào, ta dự đoán y bằng cách chọn lớp i sao cho $h^{(i)}_{\theta}(x)$ là lớn nhất

$$h^{(i)}_{\theta}(x)=P(y=1|x;\theta) \quad \text{với } (i=1,2,..k), \text{ k là số lớp}$$

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn