Hồi qui tuyến tính

Ngô Minh Nhựt

Bộ môn Công nghệ Tri thức

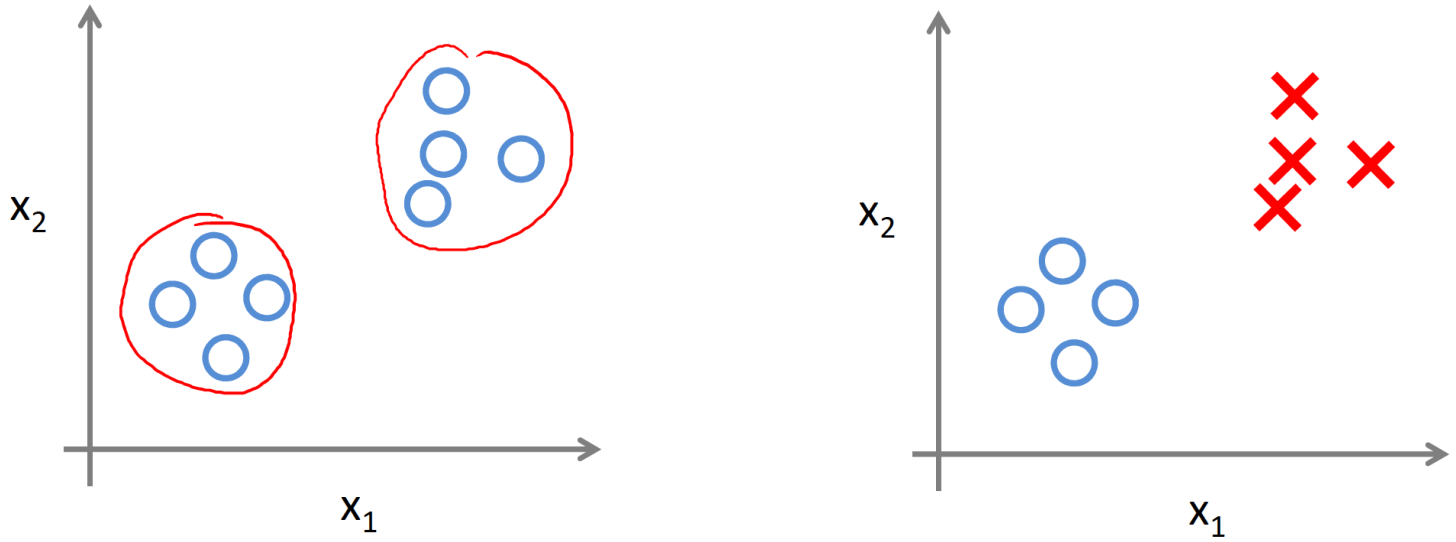
2021

Hồi qui tuyến tính đơn biến

Phần 1

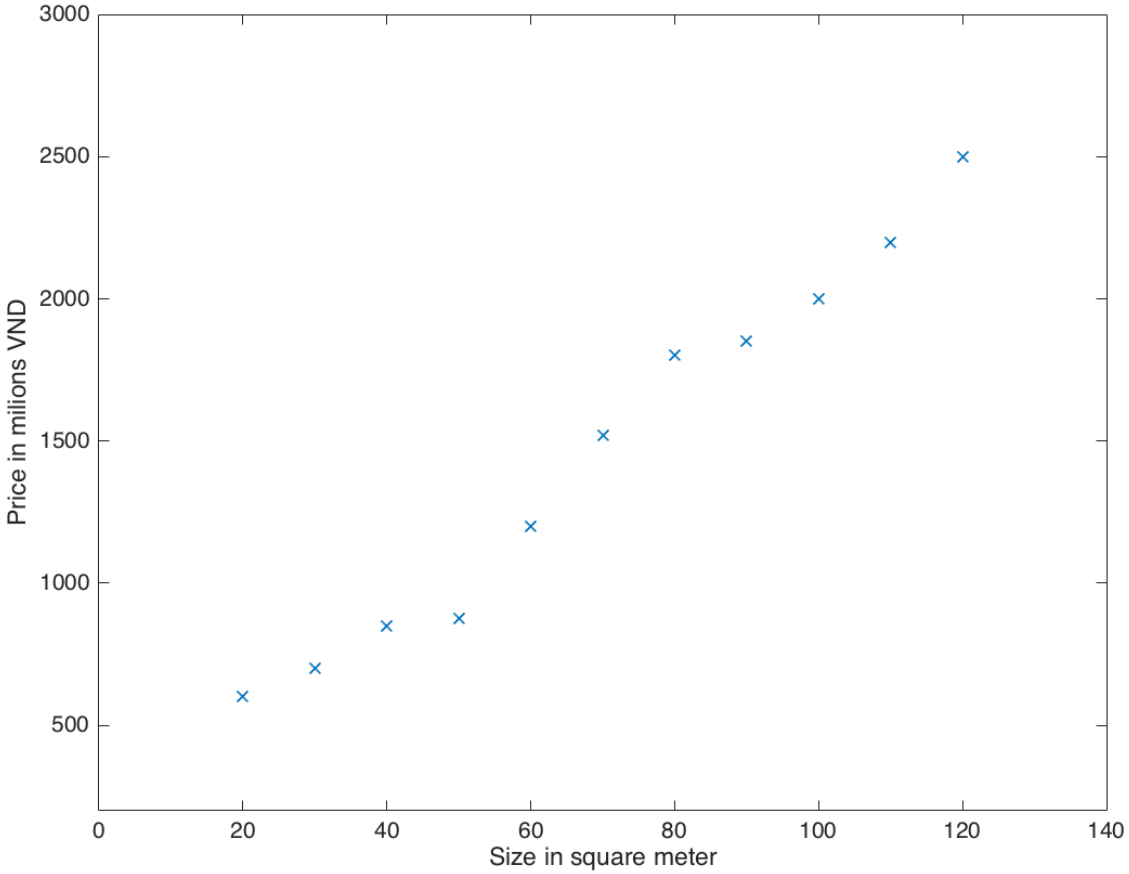
Máy học

- Mục tiêu của máy học là tìm ra cấu trúc của dữ liệu quan sát hoặc mối quan hệ bên trong chúng



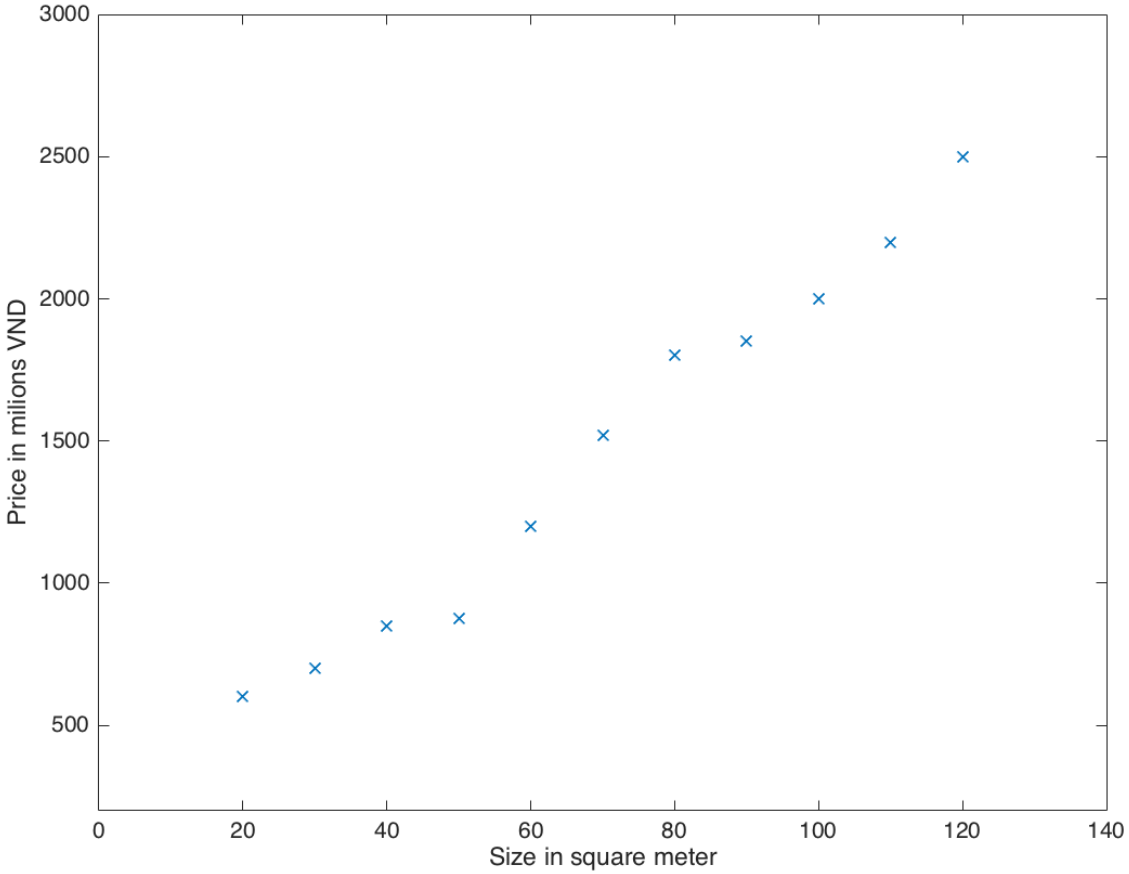
Source: Andrew Ng

Học giám sát



- Học: được cung cấp input và output tương ứng
- Suy luận: cho biết output của input mới

Hồi quy tuyến tính



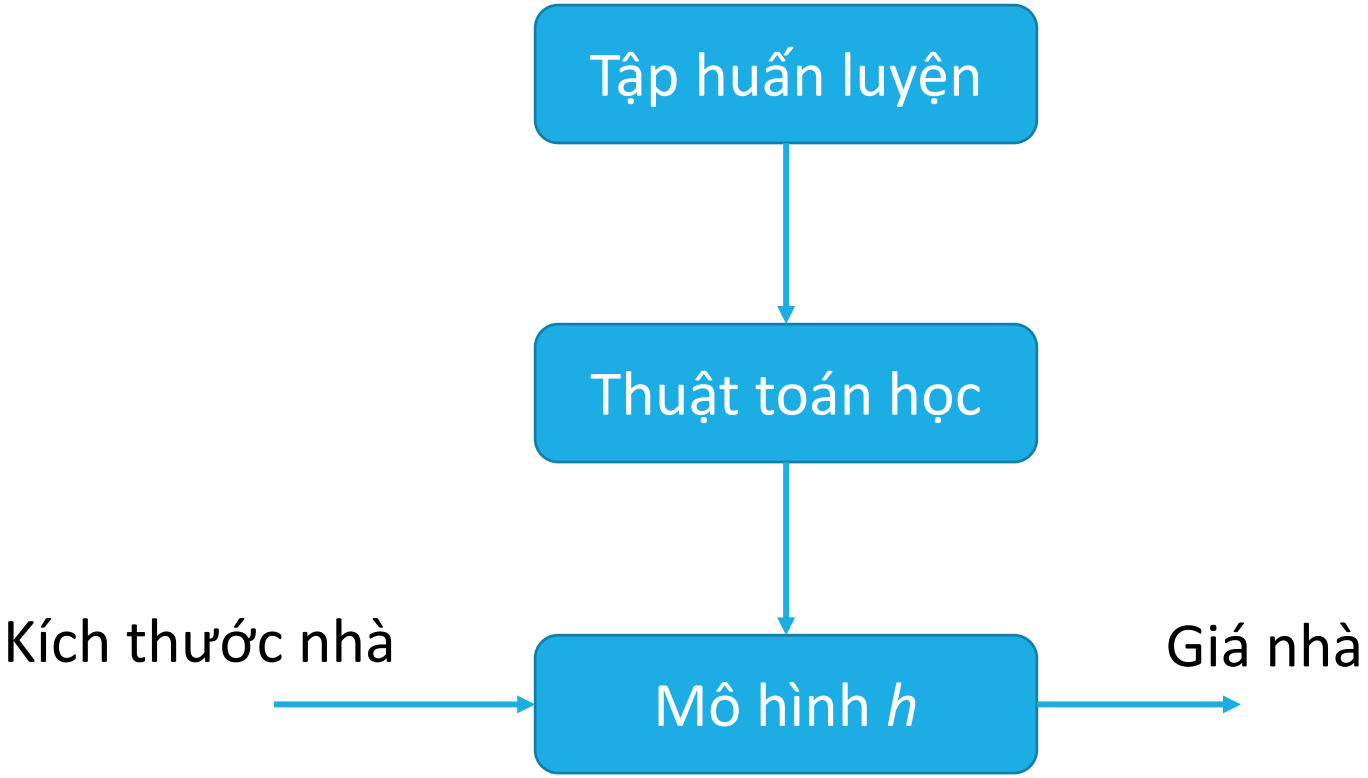
- Học: được cung cấp input và output tương ứng
- Suy luận: cho biết output của input mới
- **Output: giá trị thực liên tục**

Tập huấn luyện

Giá nhà ở
theo kích thước

x: Size (m2)	y: Price (millions VND)
20	600
50	876
80	1800
100	2000
…	…

m: số lượng mẫu
x: input
y: output/nhãn
(x, y): mẫu huấn luyện
$(x^{(i)}, y^{(i)})$: mẫu huấn luyện thứ i



- h: ánh xạ kích thước nhà sang giá
- Hồi qui tuyến tính: ánh xạ tuyến tính
- Trong bài giảng này: hồi qui tuyến tính đơn biến

Hypothesis

- Tập huấn luyện: $\left(x^{(i)}, y^{(i)}\right),\ i = 1, 2, \ldots m$
- Hypothesis : $h(x) = \theta_0 + \theta_1 x$
  - x: input
  - h(x): output
- Mục tiêu: xác định θ_0 và θ_1 để mô hình $\mathrm{h(x)}$ khớp với dữ liệu huấn luyện nhất
  - Cho x, xác định $h(x)$, sao cho $h(x)$ gần y nhất
    - x: input
    - h(x): output ước lượng
    - y: output thực tế

Hypothesis

Giá nhà ở
theo kích thước

x: Size (m2)	y: Price (millions VND)
20	600
50	876
80	1800
100	2000
…	…

Hypothesis : $h(x) = \theta_0 + \theta_1 x$

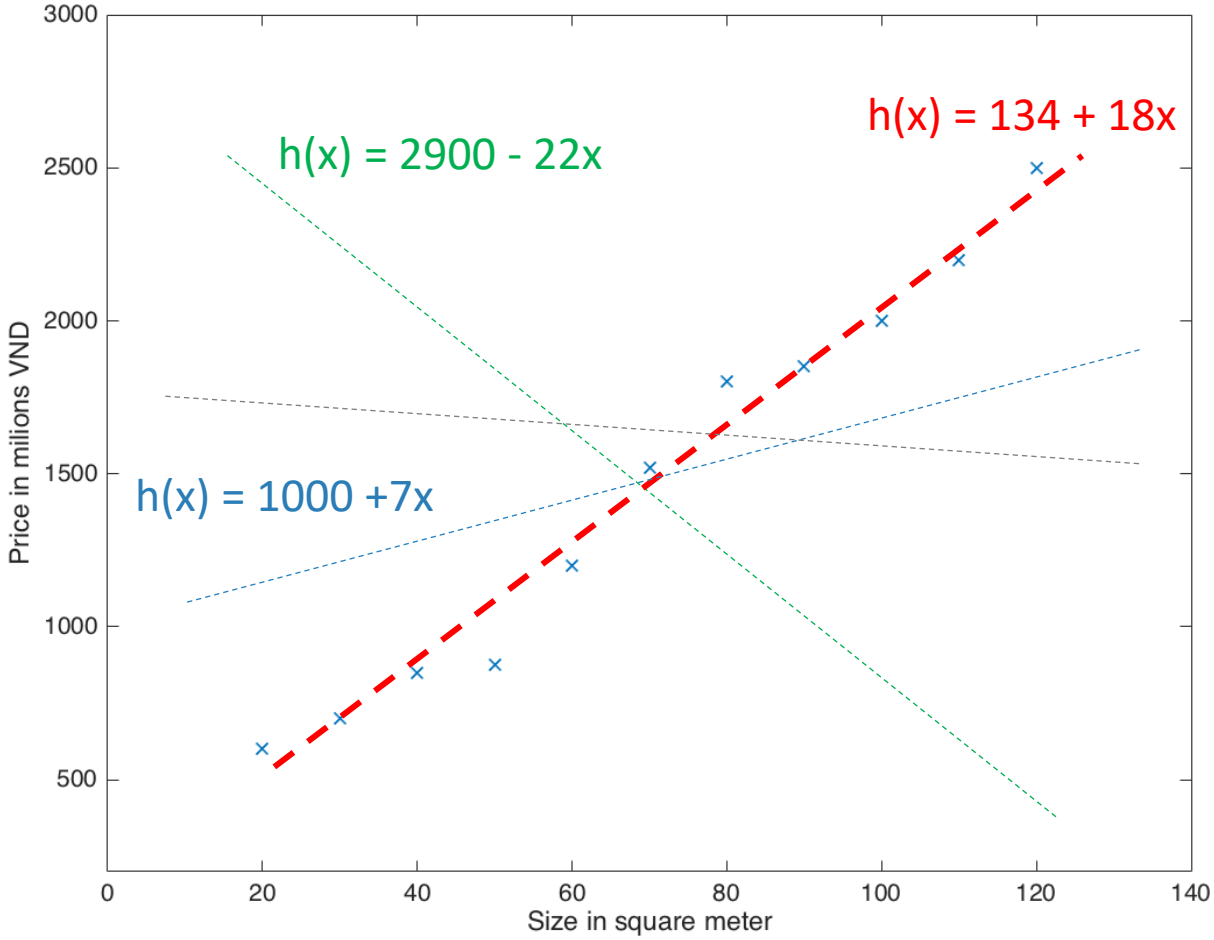
x: input

h(x): output

θ_0, θ_1: tham số

Quá trình học: tìm ra θ_0, θ_1 từ tập huấn luyện

Hypothesis


h(x) = 2900 - 22x
h(x) = 134 + 18x
h(x) = 1000 +7x
Price in milions VND
Size in square meter
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140

Hàm chi phí (cost function)

- Lỗi của mô hình ứng với 1 input
  - $\frac{1}{2}(h(x)-y)^2 = \frac{1}{2}(\theta_0+\theta_1 x-y)^2$
- Hàm chi phí
  - $J(\theta_0,\theta_1) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h\left(x^{(i)}\right)-y^{(i)}\right)^2$
  - $J(\theta_0,\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(\theta_0+\theta_1 x^{(i)}-y^{(i)}\right)^2$
- Mục tiêu:
  - Xác định θ_0,θ_1 sao cho $J(\theta_0,\theta_1)$ đạt cực tiểu

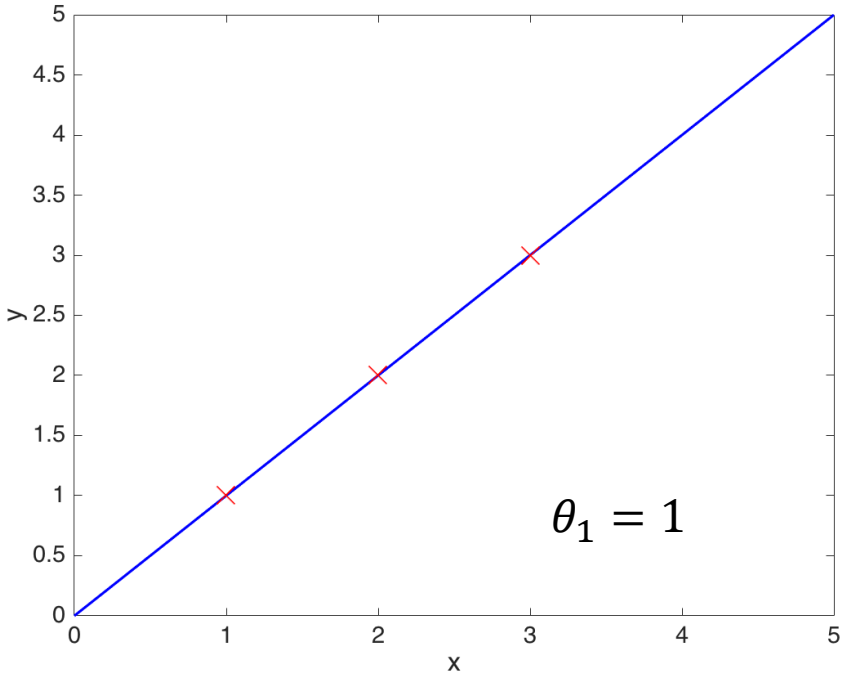
Hàm chi phí

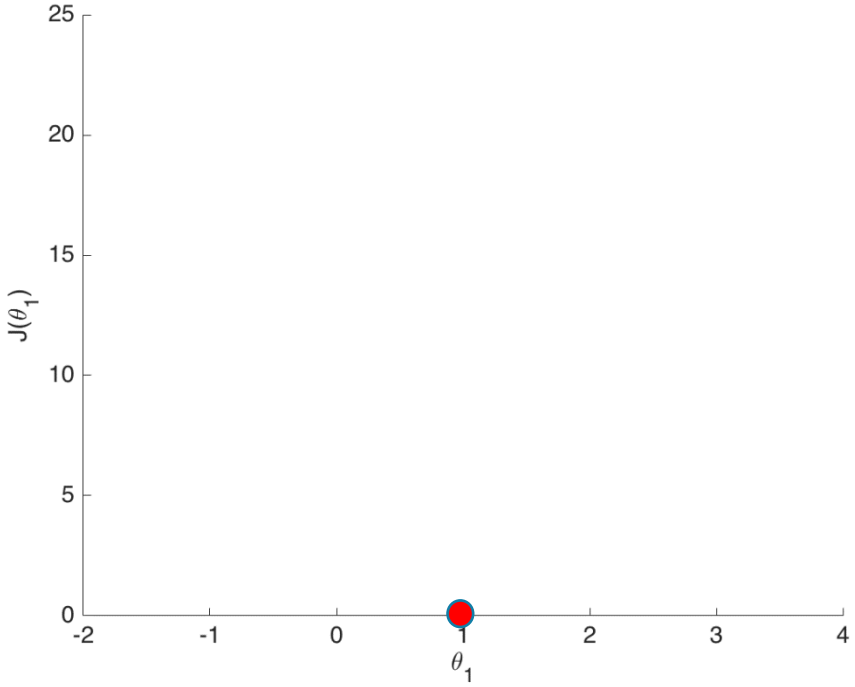
- Hypothesis : $h_\theta(x) = \theta_0 + \theta_1 x$
- Tham số: θ_0, θ_1
- Hàm chí phí:

$$\mathrm{J}(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)^2$$

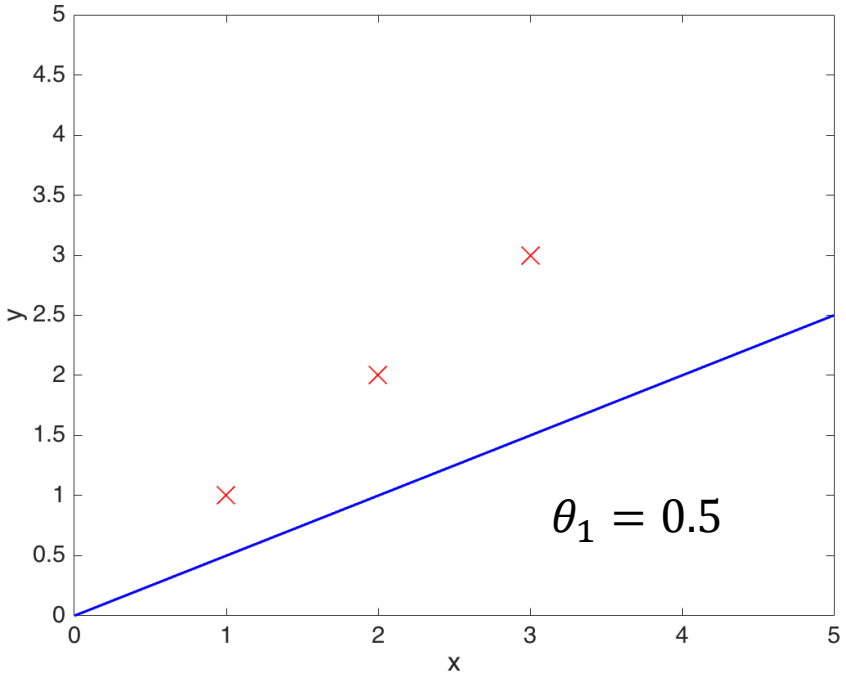
- Mục tiêu: tìm giá trị θ_0, θ_1 sao cho $\mathrm{J}(\theta_0, \theta_1)$ đạt cực tiểu
  - $Minimize_{\theta_0, \theta_1} \mathrm{J}(\theta_0, \theta_1)$
- Để minh họa: $\theta_0 = 0$, $h_\theta(x) = \theta_1 x$
  - $\mathrm{J}(\theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)^2$

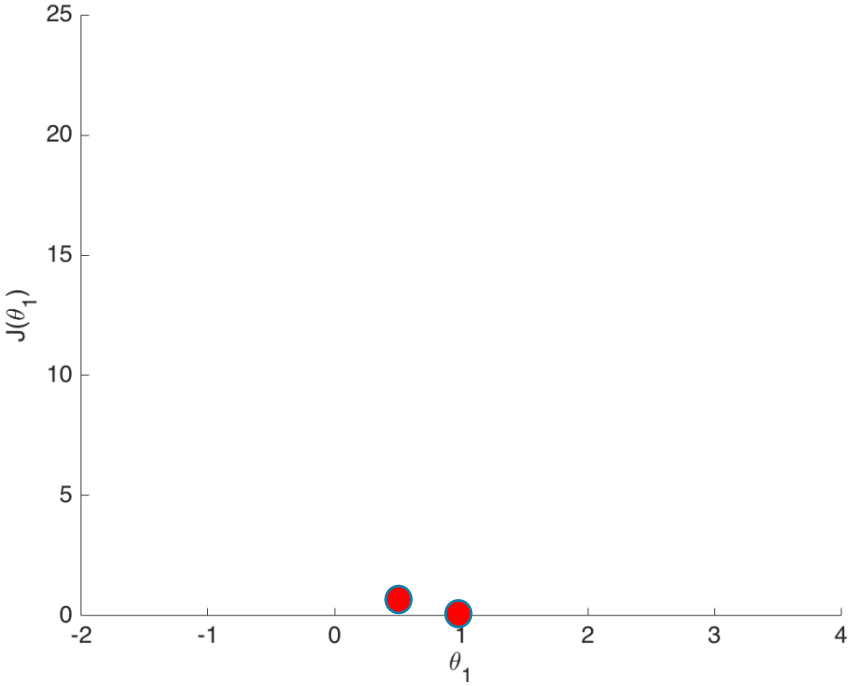
Hàm chi phí



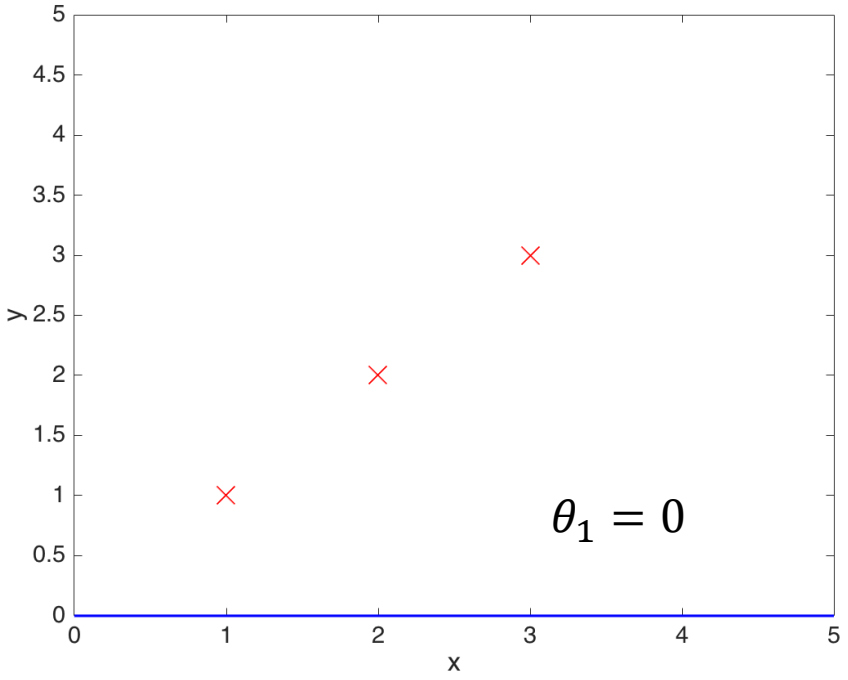


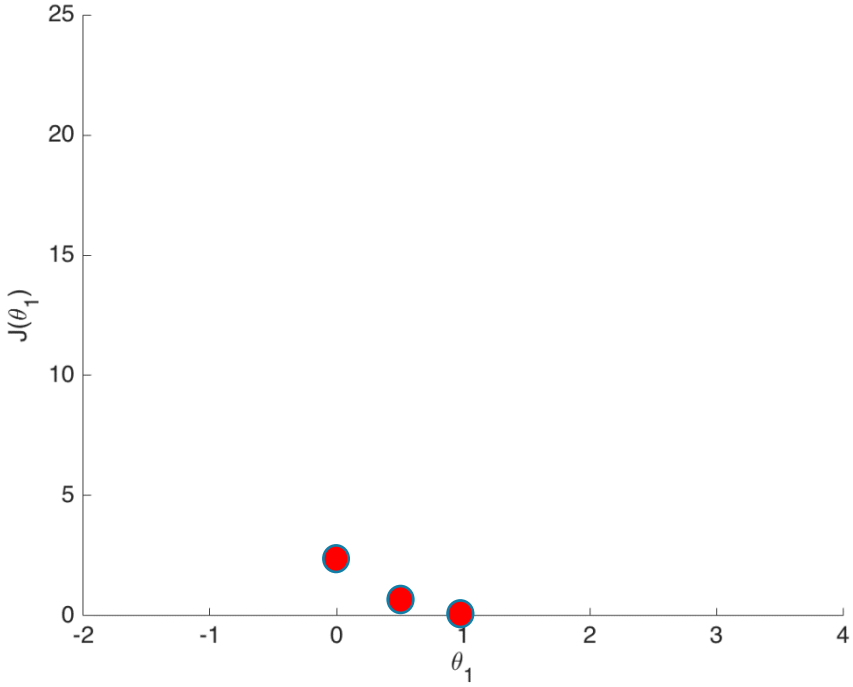
Hàm chi phí


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
y
0
1
2
3
4
5
x
$\theta_1 = 0.5$

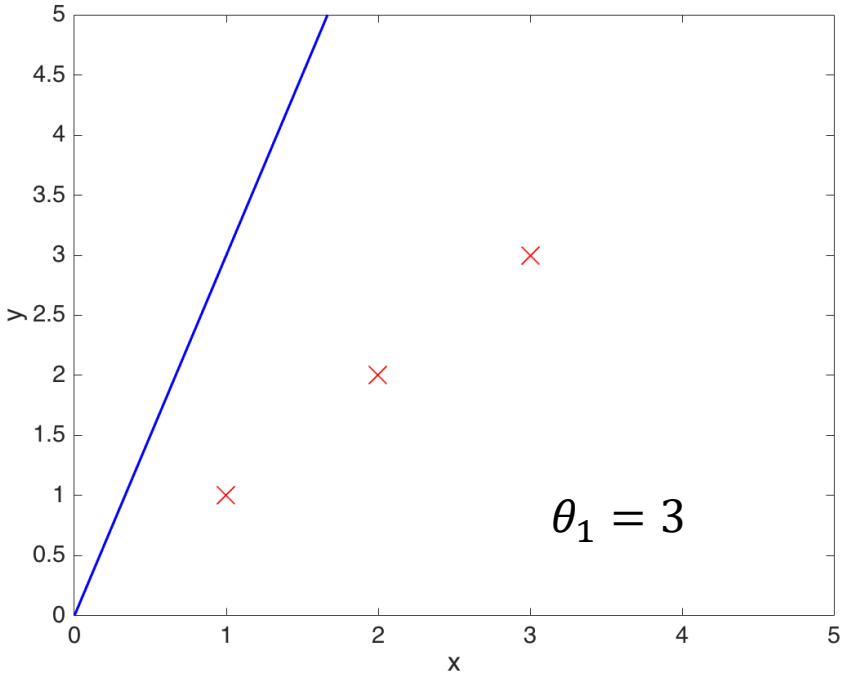

25
20
15
10
5
0
$J(\theta_1)$
-2
-1
0
1
2
3
4
θ_1

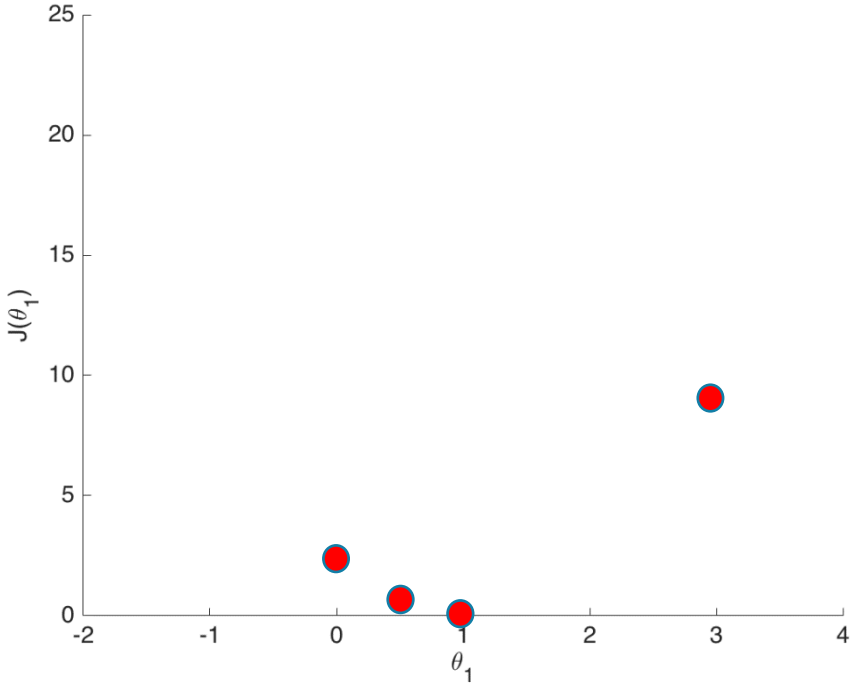
Hàm chi phí


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
y
0
1
2
3
4
5
x
$\theta_1 = 0$

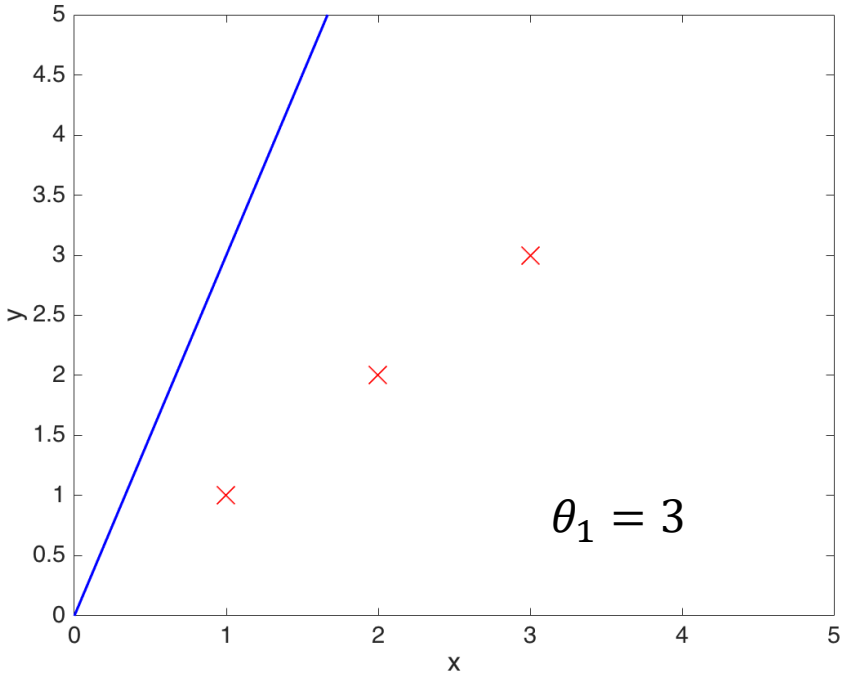

25
20
15
10
5
0
$J(\theta_1)$
-2
-1
0
1
2
3
4
θ_1

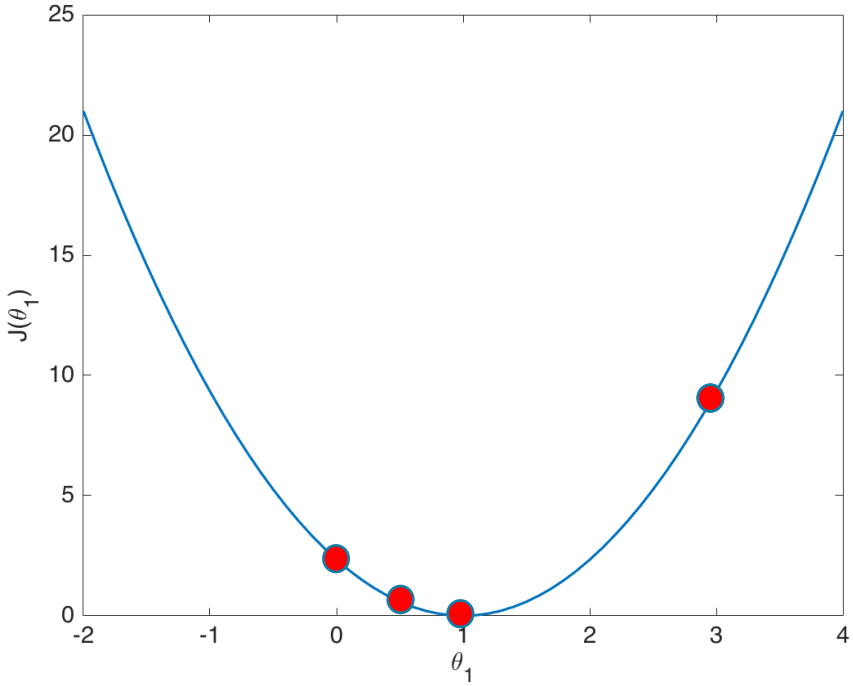
Hàm chi phí


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
y
0
1
2
3
4
5
x
$\theta_1 = 3$

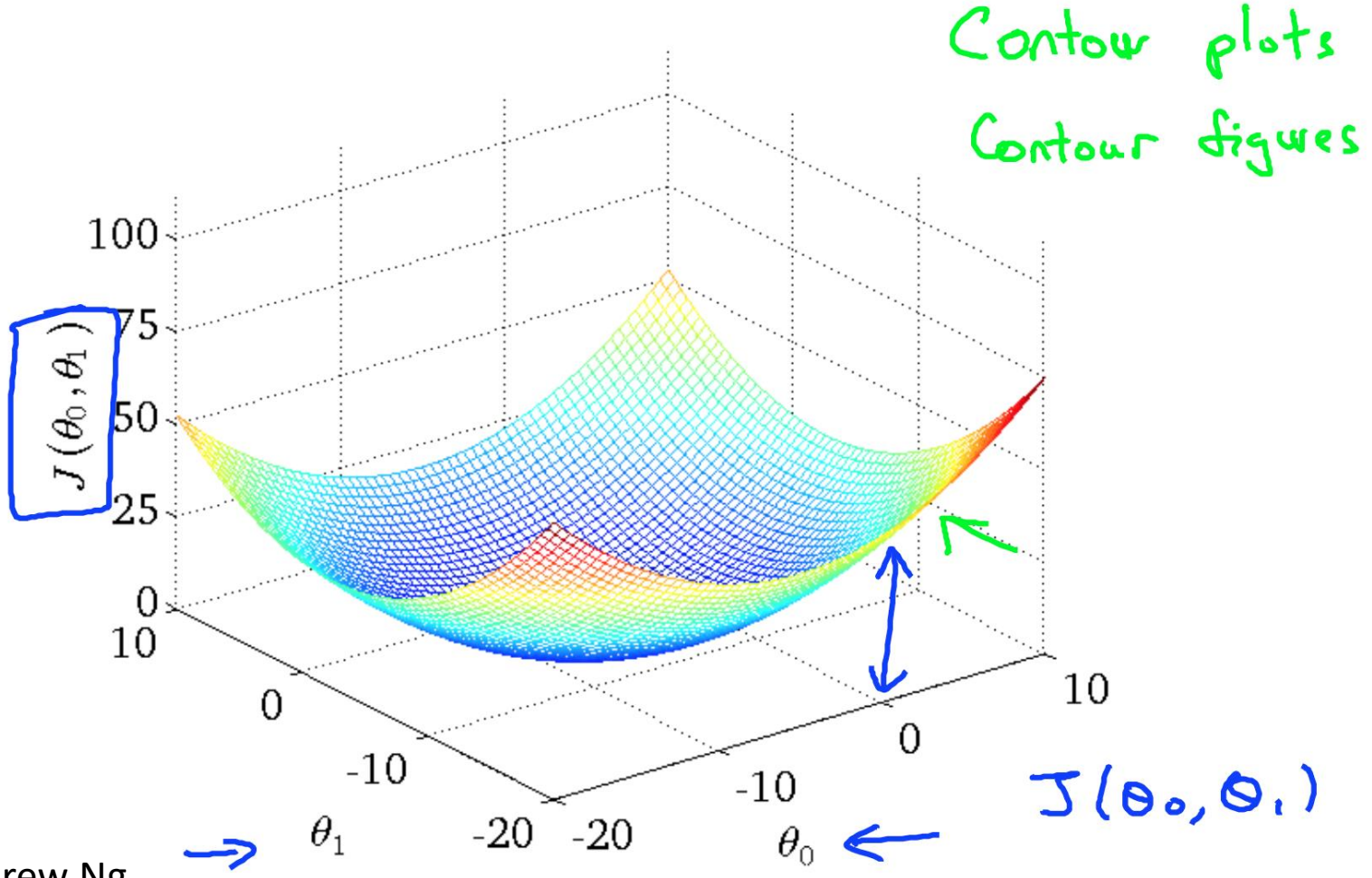

25
20
15
10
5
0
$J(\theta_1)$
-2
-1
0
1
2
3
4
θ_1

Hàm chi phí





Hàm chi phí



Source: Andrew Ng

Thuật toán hạ dốc (gradient descent)

- Hàm chi phí
  - $J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h\left(x^{(i)}\right) - y^{(i)}\right)^2$
- Mục tiêu:
  - Xác định θ_0, θ_1 sao cho $J(\theta_0, \theta_1)$ đạt cực tiểu

Thuật toán hạ dốc (gradient descent)

- Hàm chi phí
  - $J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h\left(x^{(i)}\right) - y^{(i)}\right)^2$
- Mục tiêu:
  - Xác định θ_0, θ_1 sao cho $J(\theta_0, \theta_1)$ đạt cực tiểu
- Đi tìm θ_0, θ_1
  - Bắt đầu từ 1 điểm θ_0, θ_1nào đó (ví dụ: $\theta_0 = 0, \theta_1 = 0$)
  - Thay đổi giá trị của θ_0, θ_1 cho đến khi $J(\theta_0, \theta_1)$ đạt cực tiểu
- Thay đổi θ_0, θ_1 như thế nào?

Thuật toán hạ dốc

- Đạo hàm riêng phần:
  - $\frac{dJ}{d\theta_0} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})$
  - $\frac{dJ}{d\theta_1} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)}$

Thuật toán hạ dốc

- Đạo hàm riêng phần:
  - $\frac{dJ}{d\theta_0} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})$
  - $\frac{dJ}{d\theta_1} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)}$
- Vector gradient:
  - $\left(\frac{dJ}{d\theta_0}, \frac{dJ}{d\theta_1}\right)$

Thuật toán hạ dốc

- Đạo hàm riêng phần:
  - $\frac{dJ}{d\theta_0} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})$
  - $\frac{dJ}{d\theta_1} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)}$
- Vector gradient:
  - $\left(\frac{dJ}{d\theta_0}, \frac{dJ}{d\theta_1}\right)$
- Hướng ngược với vector gradient làm cho hàm số giảm dần

Thuật toán hạ dốc

Lặp cho đến khi hội tụ

{

$$\theta_0 = \theta_0 - \alpha \frac{dJ}{d\theta_0}$$

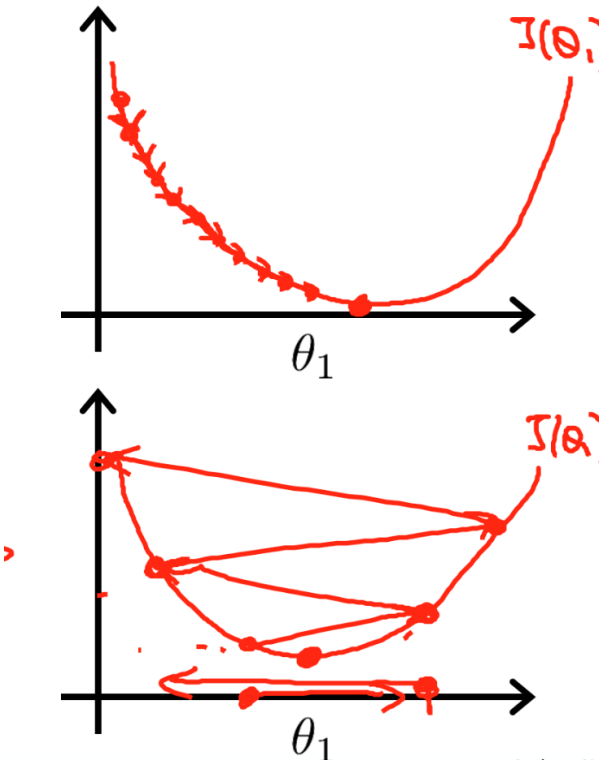
$$\theta_1 = \theta_1 - \alpha \frac{dJ}{d\theta_1}$$

}

- α là hệ số học
- Các bước thực hiện
  - Bước 1: tính vector gradient
  - Bước 2: cập nhật các thành phần của vector θ
  - Bước 3: tính lại chi phí và thông báo

Hệ số học

- Nếu α quá nhỏ thuật toán hội tụ chậm
- Nếu α quá lớn, thuận toán có thể không hội tụ



Source: Andrew Ng

Hồi qui tuyến tính đa biến

Phần 2

Tập huấn luyện

x: Size (m2)	y: Price (millions VND)
20	600
50	876
80	1800
100	2000
…	…

Hypothesis : $h_\theta(x) = \theta_0 + \theta_1 x$

Tập huấn luyện

x_1: Size (m^2)	x_2: Age (year)	x_3: Number of floor (m)	y: Price (millions VND)
20	5	1	600
20	8	3	876
80	10	3	1800
70	7	5	2000
…	…	…	…

$(x^{(i)}, y^{(i)})$: mẫu huấn luyện thứ i

$x^{(i)}$: input của mẫu thứ i

$x^{(i)}_j$: đặc trưng j của mẫu thứ i

n: số lượng đặc trưng

Hypothesis

- Đơn biến

$$h_\theta(x) = \theta_0 + \theta_1 x$$

giá trị đơn

- Đa biến

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

vector

Hypothesis

- Đa biến

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

$$x \in \mathbb{R}^n, \qquad \theta \in \mathbb{R}^{n+1}$$

Đặt $x_0 = 1$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}, \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \cdots \\ \theta_n \end{bmatrix}$$

$$\rightarrow h_\theta(x) = \theta^T x, \qquad \theta, x \in \mathbb{R}^{n+1}$$

Hàm chi phí

- Hypothesis
  - $h_\theta(x) = \theta^T x$
- Các tham số:
  - $\theta = [\theta_0, \theta_1, \theta_2, \dots \theta_n]^T$
- Hàm chi phí
  - $\mathrm{J}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(\theta^T x^{(i)} - y^{(i)}\right)^2$

Thuật toán hạ dốc

- Vector gradient:
  - $\frac{dJ}{d\theta_j} = \frac{1}{m}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})\, x_j^{(i)}$
  - $j = 0, 1, 2, \ldots, n$
- Repeat until convergence

$$\{$$

$$\theta_j = \theta_j - \alpha \frac{dJ}{d\theta_j}$$

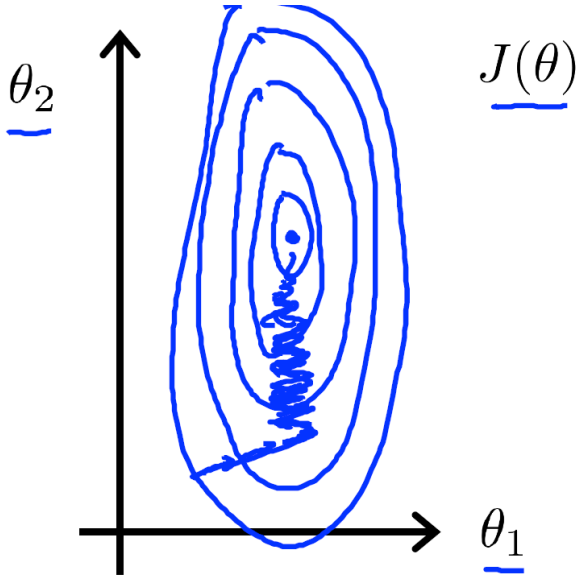
$$\}$$

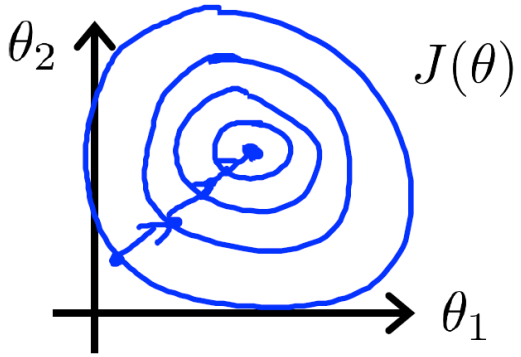
Chuẩn hóa đặc trưng

x_1: Size (m^2)	x_2: Age (year)	x_3: Number of floor (m)	y: Price (millions VND)
20	5	1	600
20	8	3	876
80	10	3	1800
70	7	5	2000
…	…	…	…

Chuẩn hóa đặc trưng

Hiệu chỉnh giá trị của các đặc trưng về cùng vùng biên độ





x_1 = 20, 30, …, 100
x_2 = 1, 2, …, 10

$$x_1 = \frac{x_1}{100}$$

$$x_2 = \frac{x_2}{10}$$

Source: Andrew Ng

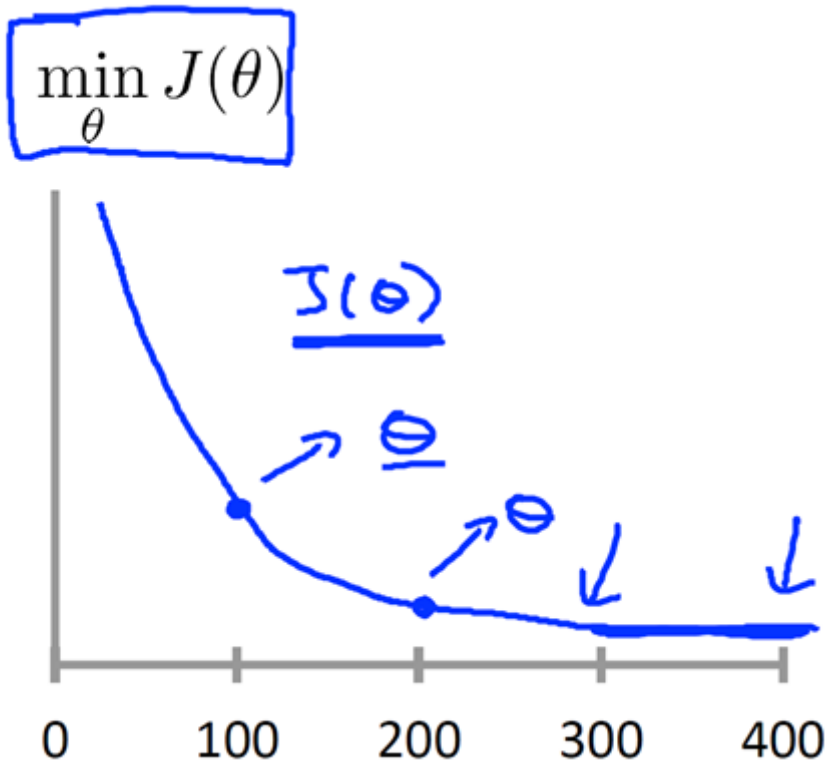
Chuẩn hóa đặc trưng

- Chuẩn hóa theo giá trị trung bình

$$x_j = \frac{x_j - \mu_j}{s_j}$$

  - μ_j: giá trị trung bình
  - s_j: độ lệch chuẩn
- Sau khi chuẩn hóa: $-1 \leq x_j \leq 1$

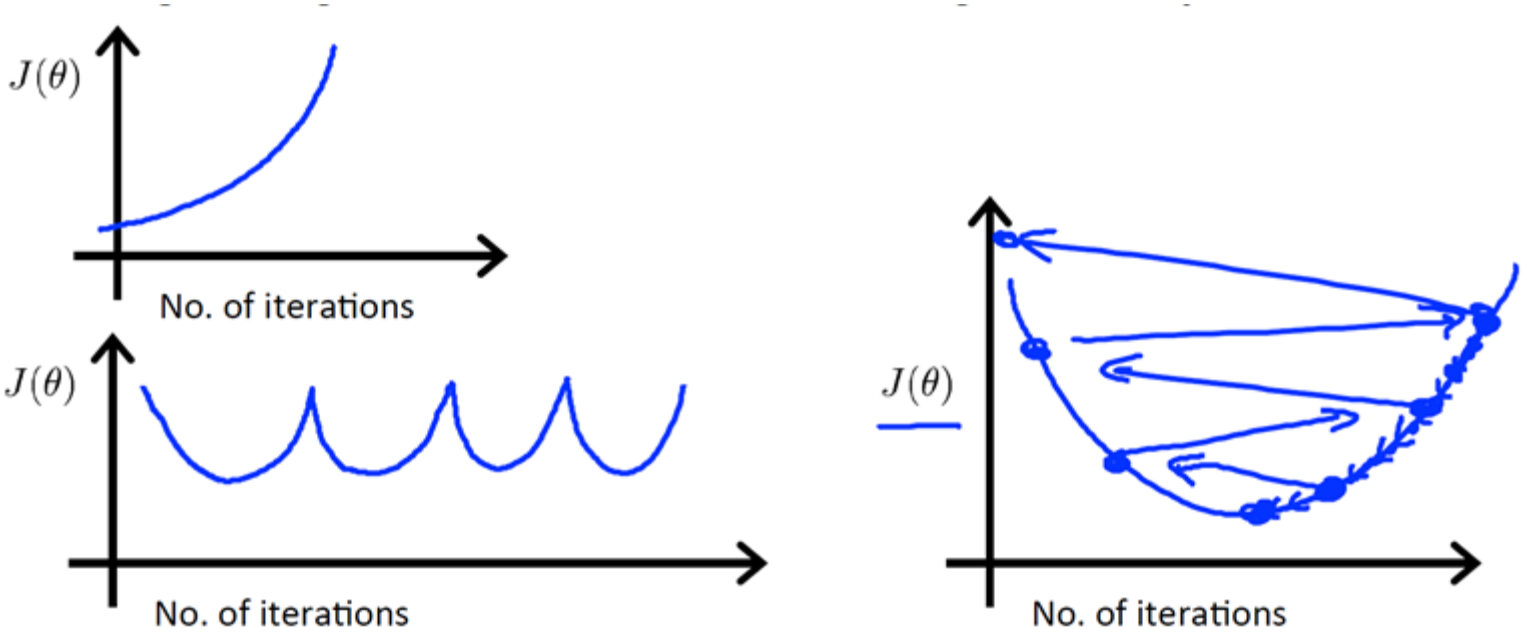
Hệ số học



- Kiểm tra J giảm qua các bước cập nhật hệ số
- J hội tụ khi J giảm ít hơn 0.001 (ε) sau mỗi lần lặp

Source: Andrew Ng

Hệ số học



- Hệ số học lớn, J có thể không hội tụ
- Hệ số học nhỏ, J hội tụ chậm
- Có thể thử: …, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, …

Source: Andrew Ng