KHOA CNTT & TRUYỀN THÔNG
BM KHOA HỌC MÁY TÍNH

HỒI QUY
REGRESSION

Giáo viên giảng dạy:
TS. TRẦN NGUYỄN MINH THƯ
tnmthu@cit.ctu.edu.vn

Quy ước

- Biến **đầu vào** (input variables)/đặc trưng (features), kí hiệu: $x^{(i)}$
- Biến **đầu ra** (output variable)/biến mục tiêu, kí hiệu y(i)
- Mẫu huấn luyện (training example)

  kí hiệu **$(x^{(i)}, y^{(i)})$**
- Tập huấn luyện **$X = \{(x^{(i)}, y^{(i)})\}$**, i = 1..m

Square meters	Bedrooms	Floors	Age of building (years)	Price in 1000€
x1	*x2*	*x3*	*x4*	*y*
200	5	1	45	460
131	3	2	40	232
142	3	2	30	315
756	2	1	36	178
…	…	…	…	…

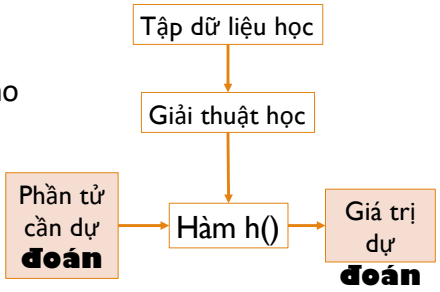
$$x^{(3)} = \begin{bmatrix} 142 \\ 3 \\ 2 \\ 30 \end{bmatrix}$$

$$x_1^{(4)} = 756$$

Phân loại học máy – **học có giám sát**

Từ tập dữ liệu huấn luyện {(x_1, y_1), (x_2,y_2),..,(x_m,y_m)

- Tìm hàm h (hypothesis) X=>Y sao cho h(x) dự báo được **y** từ **x**
- **Y là giá trị liên tục:** sử dụng pp hồi quy (regression)
- **Y là giá trị rời rạc:** sử dụng pp phân lớp (classification)

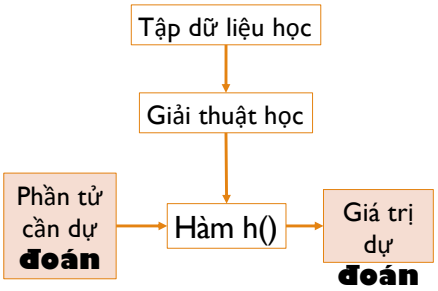


3

Phân loại học máy – **học có giám sát**

Ví dụ: bài toán dự báo giá nhà

Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



– **Đầu vào/thuộc tính**: ?

– **Đầu ra**: ?

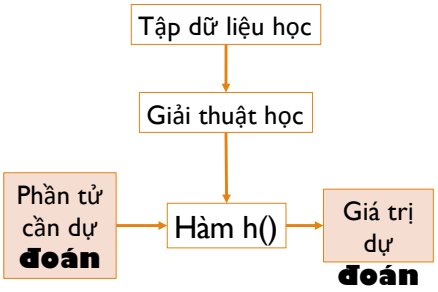
Xác định :
- Dự báo cái gì?
- Dựa trên thông tin gì?
- Giải thuật gì?

4

Phân loại học máy – **học có giám sát**

Ví dụ: bài toán dự báo giá nhà

Living area ($feet^2$)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



– **Đầu vào/thuộc tính**: diện tích

– **Đầu ra**: giá nhà - **giá trị liên tục**

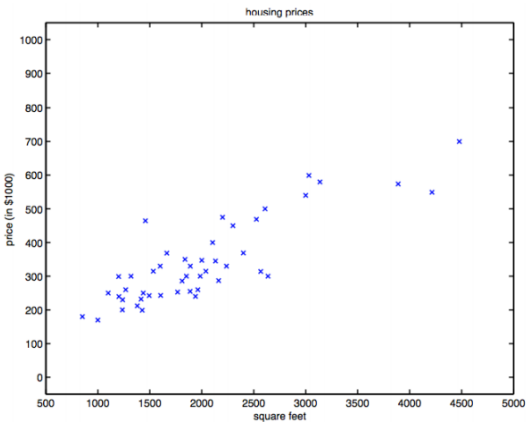
Xác định thuộc tính:
Dự báo cái gì? Y = giá nhà
Dựa trên thông tin gì?
Giải thuật gì? **Hồi quy**

5

Ví dụ dự đoán giá nhà

Dự báo giá nhà dựa vào diện tích

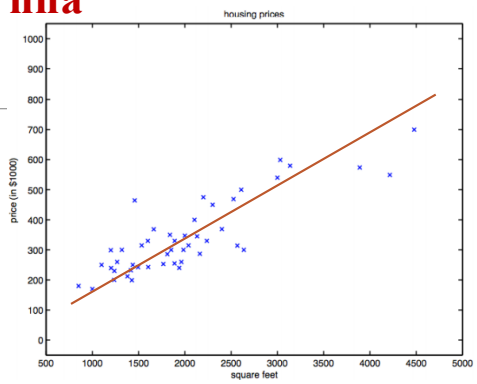
Living area ($feet^2$)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



Ví dụ dự đoán giá nhà

Dự báo giá nhà dựa vào diện tích

Living area ($feet^2$)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



➢**Biểu diễn giả thiết (hàm dự báo) h**

– Ví dụ h là một hàm tuyến tính 1 biến, h(x1) có dạng: $h_\theta(x) = \theta_0 + \theta_1 x_1$

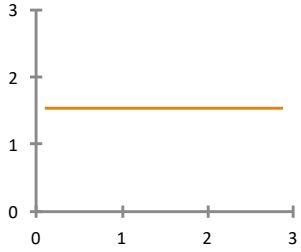
Trong đó, θ_0, θ_1 là các tham số cần mà ta phải tìm trong quá trình "dạy cho máy học" hay còn gọi là quá trình huấn luyện.

Ví dụ dự đoán giá nhà

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

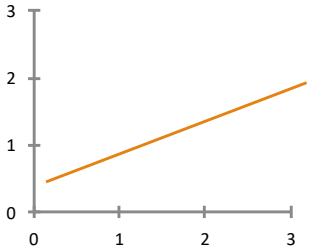
Cần xác định các tham số: θ_i

$h_\theta(x) = 1.5 + 0.x$



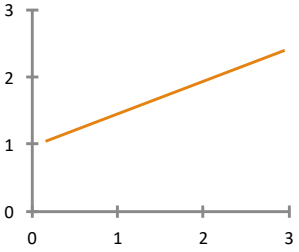
$\theta_0 = 1.5$
$\theta_1 = 0$

$h_\theta(x) = 0.5x$

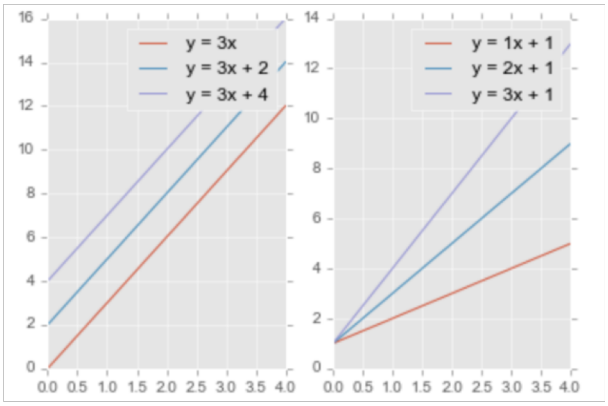


$\theta_0 = 0$
$\theta_1 = 0.5$

$h_\theta(x) = 1 + 0.5x$



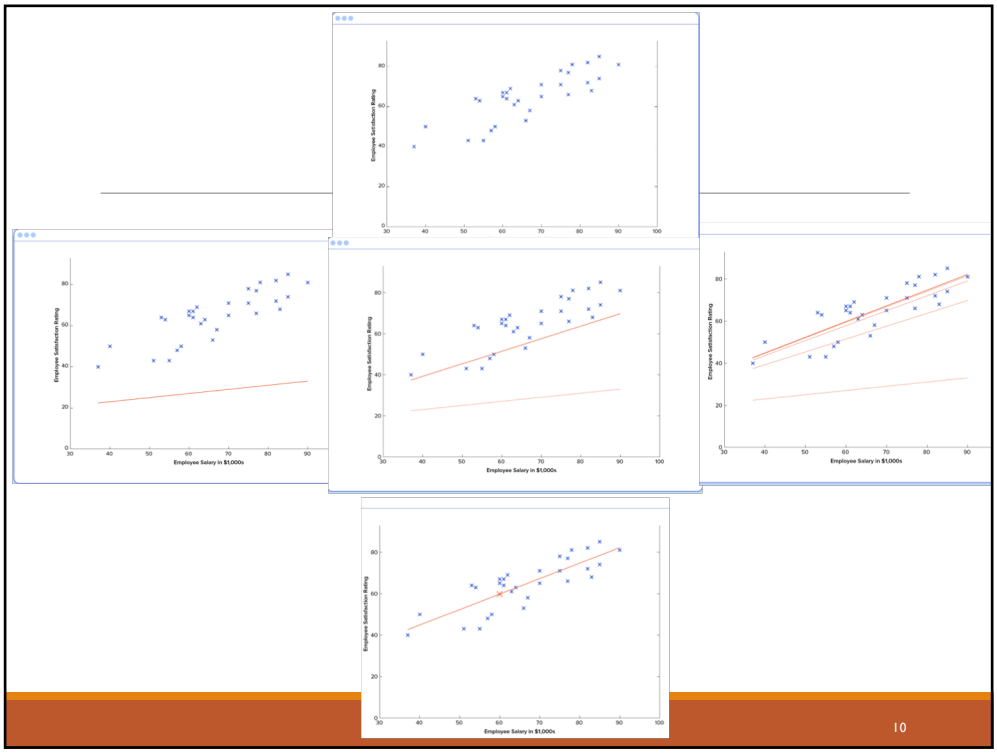
$\theta_0 = 1$
$\theta_1 = 0.5$



b: θ_0 quyết định điểm giao của đường thẳng với trục y, intercept/observation noise: điểm mà đường thẳng cắt trục Y.

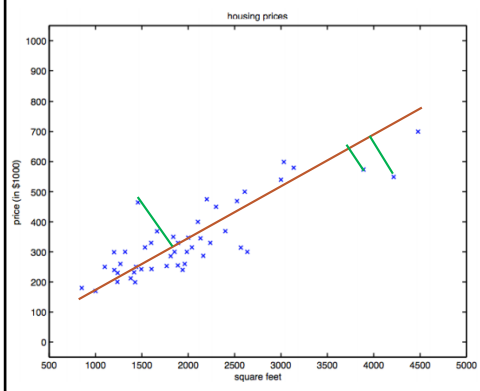
a quyết định góc của đường thẳng – $\theta_1$slope/coefficients: độ dốc của đường thẳng h(x)

9



10

Ví dụ dự đoán giá nhà



Living area ($feet^2$)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮

Trong khi sử dụng hồi quy tuyến tính, mục tiêu của chúng ta là để làm sao một đường thẳng có thể tạo được sự phân bố gần nhất với hầu hết các điểm. Do đó làm giảm khoảng cách (sai số) của các điểm dữ liệu cho đến đường đó.

Hồi quy tuyến tính

Dạy cho máy học/huấn luyện như thế nào ?

– Tìm các tham số từ tập huấn luyện sao cho lỗi huấn luyện (training error) nhỏ nhất.
– Ta phải tìm h sao cho **h(x)** gần với **y** nhất (sai số dự đoán)

Nói cách khác, chúng ta muốn giá trị sau đây càng nhỏ càng tốt:

$$h_\theta(x^{(i)}) - y^{(i)}$$

– Hàm chi phí/hàm lỗi (cost function/error function) của **m** phần tử

•Hàm lỗi **sai số tuyệt đối**: $\sum_{i=1}^{m} \left| h_\theta(x^{(i)}) - y^{(i)} \right|$

•Hàm lỗi **sai số bình phương**: $\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

Hồi quy tuyến tính

Dạy cho máy học/huấn luyện như thế nào ?

– Tìm các tham số từ tập huấn luyện sao cho lỗi huấn luyện (training error) nhỏ nhất.
– Ta phải tìm h sao cho **h(x)** gần với **y** nhất = $h_\theta(x^{(i)}) - y^{(i)}$
– Hàm chi phí/hàm lỗi (cost function/error function)

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

$\sum_{i=1}^{m}|h_\theta(x^{(i)}) - y^{(i)}|$

$\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$

Mục tiêu tìm θ sao cho J(θ) nhỏ nhất

Giảm gradient

- Tìm θ sao cho J(θ) nhỏ nhất
- Khởi tạo ngẫu nhiên θ
- Tăng/giảm θ một lượng Δθ sao cho J(θ +/- Δθ) nhỏ hơn J(θ)

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$

α: tốc độ học

LMS (Least mean square): bình phương trung bình nhỏ nhất

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

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

Giải thuật LMS

Tính đạo hàm riêng theo từng tham số:

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} (h_\theta(x) - y)^2 \\
&= 2 \cdot \frac{1}{2} (h_\theta(x) - y) \cdot \frac{\partial}{\partial \theta_j} (h_\theta(x) - y) \\
&= (h_\theta(x) - y) \cdot \frac{\partial}{\partial \theta_j} \left(\sum_{i=0}^{n} \theta_i x_i - y \right) \\
&= (h_\theta(x) - y)\, x_j
\end{aligned}$$

Đạo hàm riêng

Scalar multiple rule:	$\frac{d}{dx}(\alpha u) = \alpha \frac{du}{dx}$
Sum rule:	$\frac{d}{dx} \sum u = \sum \frac{du}{dx}$
Power rule:	$\frac{d}{dx} u^n = n u^{n-1} \frac{du}{dx}$
Chain rule:	$\frac{d}{dx} f(g(x)) = f'(g(x)) g'(x)$

$$\frac{d}{d\theta_1}\left(h_\theta(x^{(i)}) - y^{(i)}\right) = \frac{d}{d\theta_1}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) = x^{(i)}$$

16

Giải thuật LMS

Nếu chỉ có **1 mẫu huấn luyện**, ta cập nhật:

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

Nếu có **nhiều mẫu huấn luyện**, sử dụng luật cập nhật:

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

Hoặc:

for i=1 to m, {

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

}

Ví dụ

Cho tập dữ liệu gồm 3 phần tử như bảng bên, hãy thực hiện các công việc sau

- Biểu diễn tập dữ liệu lên mặt phẳng toạ độ Oxy
- Tìm hàm hồi quy h(x) với giá trị khởi tạo (0, 1), tốc độ học: 0.2, số bước lặp: 2
- Vẽ đường hồi quy lên mặt phẳng toạ độ
- Dự đoán giá trị y cho phần tử có **x = 3**

x	y
1	2
2	3
4	6

for i=1 to m, {

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

}

18

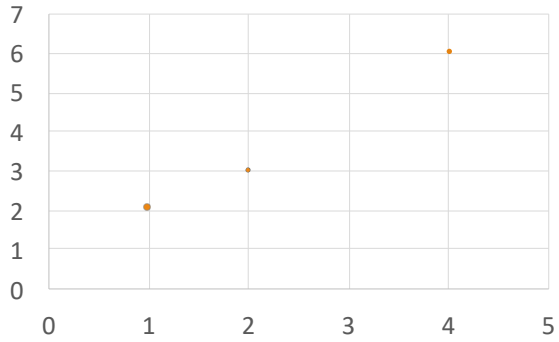
Ví dụ

for i=1 to m, {

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

}

- Biểu diễn tập dữ liệu lên mặt phẳng toạ độ Oxy



x	y
1	2
2	3
4	6

19

Ví dụ

for i=1 to m, {

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

}

- Tìm hàm hồi quy h(x) với giá trị khởi tạo (0, 1), tốc độ học: 0.2, số bước lặp: 2

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

$\theta_0 = 0, \theta_1 = 1; \alpha = 0.2 \quad h_\theta(x)_0 = 0 + 1.x$

Lần lặp 1:

Pt 1(1,2):($x^{(1)}$):

x	y
1	2
2	3
4	6

Tìm θ_0 $\theta_0 = \theta_{0+} \alpha*(y^1 - \{0 + 1* x^1{}_1\})*x^1{}_0$

$= 0 + 0.2(2- \{0 +1*1\})*1 = 0.2$

Tìm θ_1 $\theta_1 = \theta_{1+} \alpha*(y_1 - \{0 + 1*x_1\})*x_1$

$= 1 + 0.2(2- \{0 +1*1\})*1 = 1.2$

20

for i=1 to m, {

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

}

Ví dụ

- Tìm hàm hồi quy h(x) với giá trị khởi tạo (0, 1), tốc độ học: 0.2, số bước lặp: 2

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

$\theta_0 = 0, \theta_1 = 1; \alpha = 0.2$ $h_\theta(x)_0 = 0 + 1.x$

Lần lặp 1:

Pt 2(2,3): ($x^{(2)}$):

x	y
1	2
2	3
4	6

Tìm θ_0: $\theta_0 = \theta_{0+} \alpha*(y_2 - \{0.2 + 1*x^2{}_1\})*x^2{}_0$

$= \mathbf{0.2} + 0.2(3 - \{0.2 + \mathbf{1.2}*2\}) *1 = \mathbf{0.28}$

Tìm θ_1 $\theta_1 = \theta_{1+} \alpha*(y_2 - \{0.2 + 1.2*x_2\})*x_2$

$= 1.2 + 0.2*(3 - (0.2 + 1.2*2)) *2 = \mathbf{1.36}$

for i=1 to m, {

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

}

Ví dụ

- Tìm hàm hồi quy h(x) với giá trị khởi tạo (0, 1), tốc độ học: 0.2, số bước lặp: 2

$h_\theta(x) = \theta_0 + \theta_1 x$; $\theta_0 = 0, \theta_1 = 1; \alpha = 0.2$ $h_\theta(x)_0 = 0 + 1.x$

x	y
1	2
2	3
4	6

Lần lặp 1: Tìm θ_0

Pt 3(4,6): ($x^{(3)}$)

Tìm θ_0: $\theta_0 = \theta_{0+} \alpha*(y_1 - (0.36 + 1* x^3{}_1))* x^3{}_0$

$= \mathbf{0.28} + 0.2*(6 - (\mathbf{0.28} + \mathbf{1.44}*4))*1 = \mathbf{0.336}$

Tìm θ_1 $\theta_1 = \theta_{1+} \alpha*(y_1 - \{0.36 + 1*x_1\})*x_1$

$= 1.44 + 0.2(6 - (\mathbf{0.28} + \mathbf{1.44}*4))*4 = \mathbf{1.58}$

$h_\theta(x)_0 = \mathbf{0.336} + \mathbf{1.58}.x$

Ví dụ

for i=1 to m, {

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

}

- Tìm hàm hồi quy h(x) với giá trị khởi tạo (0, 1), tốc độ học: 0.2, số bước lặp: 2

$h_\theta(x) = \theta_0 + \theta_1 x \Rightarrow h_\theta(x)_0 = 0.336 + 1.58.x$

x	y
1	2
2	3
4	6

Lần lặp 2: Tìm θ_0

Tiếp tục với giá trị $\theta_0 = 0.336$, $\theta_1 = 1.58$;

23

Đánh giá mô hình hồi quy

Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Mean Squared Error (MSE) is the mean of the squared errors:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

24

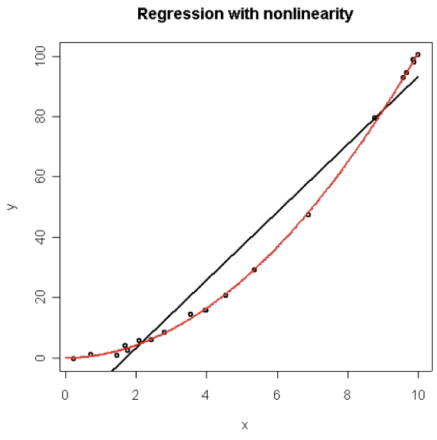
Phân loại hồi quy

- Phân loại
  - Hồi qui tuyến tính (linear) và phi tuyến (nonlinear)
  - Hồi qui đơn biến (single) và đa biến (multiple)
  - Hồi qui có thông số (parametric), phi thông số (nonparametric), và thông số kết hợp (semiparametric)
  - Hồi qui đối xứng (symmetric) và bất đối xứng (asymmetric)

25

Phân loại hồi quy

- Phân loại
  - Hồi qui tuyến tính (linear) và phi tuyến (nonlinear)
    - Linear in parameters: kết hợp tuyến tính các thông số tạo nên Y
    - Nonlinear in parameters: kết hợp phi tuyến các thông số tạo nên Y



26

Phân loại hồi quy – hồi quy đơn biến

Cho N đối tượng đã được quan sát, mô hình hồi qui tuyến tính đơn biến được cho dưới dạng sau với ε_i dùng giữ phần biến thiên của đáp ứng Y không được giải thích từ X:

-Dạng đường thẳng

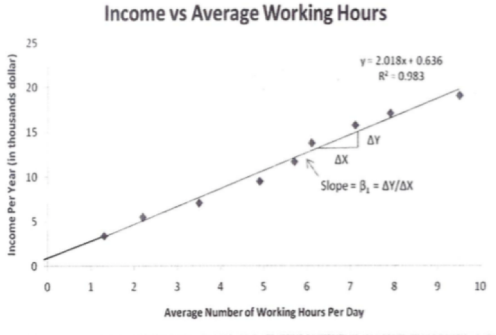
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, N.$$

-Dạng parabola

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \ i = 1, \ldots, N.$$

27

Hồi quy tuyến tính đơn biến



- Y= β0 + **β1***X1 → Y = 0.636 + 2.018*X
- Dấu của β1 cho biết sự ảnh hưởng của X đối với Y.

28

Hồi quy tuyến tính **đa biến**

Ví dụ: bài toán dự báo giá nhà

Living area ($feet^2$)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

Hồi quy tuyến tính đa biến

Thiết lập bài toán

- **Xác định thuộc tính**:
  - Dự báo cái gì
  - Dựa trên thông tin gì?

Living area ($feet^2$)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

- **Biểu diễn giả thiết (hàm dự báo) h**

h là một hàm tuyến tính 2 biến, h(x1, x2) có dạng:

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

Trong đó, θ_0 , θ_1, θ_2 là các tham số mà ta cần phải tìm trong qua trình "dạy cho máy học" hay còn gọi là qua trình huấn luyện.

Bài toán phân lớp và hồi quy logistic

Bài toán phân lớp:

- Giống như bài toán hồi quy, ngoài trừ y có giá trị rời rạc
- Ví dụ bài toán 2 lớp:
  - 0: lớp âm
  - 1: lớp dương
- Ta có thể giải bài toán phân lớp này bằng giải thuật hồi quy tuyến tính như trên. Tuy nhiên, vì y chỉ có thể có giá trị là 0 hoặc 1, nên không cần thiết phải định nghĩa h có nhiều giá trị.

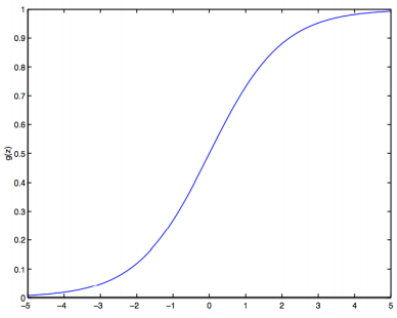
Hồi quy logistic

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

Trong đó

$$g(z) = \frac{1}{1 + e^{-z}}$$

Hàm logistic hay hàm sigmoid



Hồi quy logistic

Đạo hàm của g(z)

$$\begin{aligned} g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= \frac{1}{(1+e^{-z})^2}\left(e^{-z}\right) \\ &= \frac{1}{(1+e^{-z})}\cdot\left(1-\frac{1}{(1+e^{-z})}\right) \\ &= g(z)(1-g(z)). \end{aligned}$$

Hồi quy logistic

Tìm tham số như thế nào với hàm hồi quy logistic ?

Bỏ qua các công thức phức tạp, ta thu được luật cập nhật tham số y hệ như trường hợp hồi quy tuyến tính !

$$\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}$$

Tuy nhiên cần phải chú ý rằng: hàm h(x) trong trường hợp này là hàm logistic.