Classification

VD. Email. Spam / not spam?

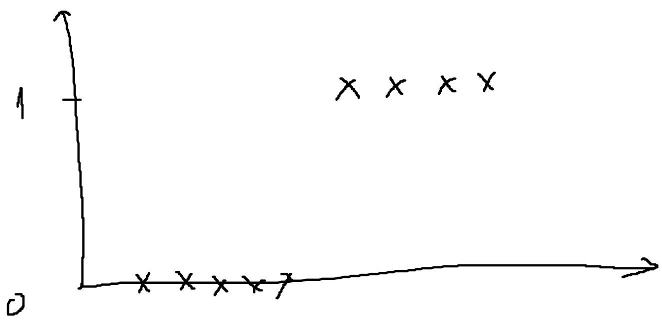
Tế bào: lành tính / ác tính?

$y \in \{0, 1\}$ 0. negative

1. Positive class

$y \in \{0, 1, 2, 3, 4, \ldots\} \rightarrow$ multiple classification

* đồ thị



có thể sử dụng linear regression nhưng không chính xác

classification: 0 or 1

$h_\theta(x)$: can be > 1 or < 0

Logistic regression: $0 < h_\theta(x) < 1$

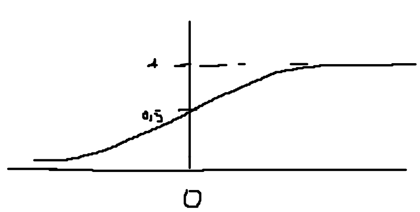
Model

Mong muốn. $0 \leq h_\theta(x) \leq 1$

$h_\theta(x) = g(\theta^T x)$

$g(z) = \dfrac{1}{1 + e^{-z}}$. Sigmoid function / Logistic function

$\Leftrightarrow h_\theta(x) = \dfrac{1}{1 + e^{-(\theta^T x)}}$



giải thích toán học

$h_\theta(x) =$ xác suất để $y = 1$ tại mẫu x

VD: $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \end{bmatrix}$

$h_\theta(x) = 0,7 = 70\%$

$\Rightarrow$ 70% ước tính

$h_\theta(x) = P(y=1 \mid x, \theta)$ — tỉ lệ để $y=1$ tại x với tham số là θ

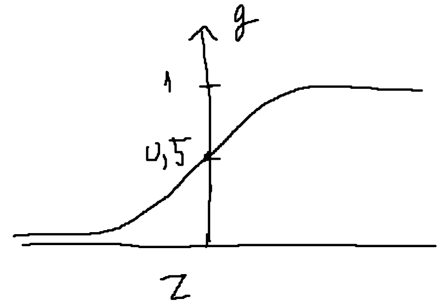
Decision boundary

$$h_\theta(x) = g(\theta^T x) = P(y=1|x;\theta)$$

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

dự đoán $y=1$ nếu $h_\theta(x) \geq 0,5$

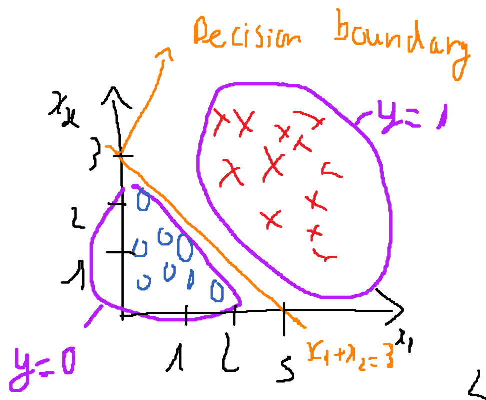
$y=0$ nếu $h_\theta(x) < 0,5$



$g(z) \geq 0,5$ khi $z \geq 0$

$\Leftrightarrow h_\theta(x) \geq 0$ hay $\theta^T x \geq 0$

VD



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

giả sử $\theta_0 = -3$, $\theta_1 = 1$, $\theta_2 = 1$

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

$y=1 \Leftrightarrow h_\theta(x) \geq 0,5 \Leftrightarrow -3 + x_1 + x_2 \geq 0$

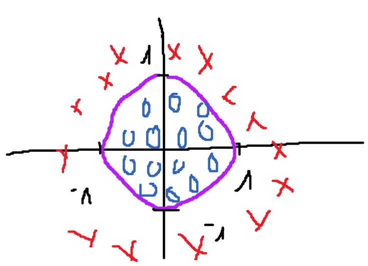
$x_1 + x_2 \geq 3$

$x_1 + x_2 = 3 \Leftrightarrow h_\theta(x) = 0,5$

$\Downarrow$

decision boundary

Non-linear decision boundaries



$$h_\theta(x) = g(-1 + x_1^2 + x_2^2)$$

decision boundary $x_1^2 + x_2^2 = 1$

x

Cost function

training set $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots (x^{(m)}, y^{(m)})\}$

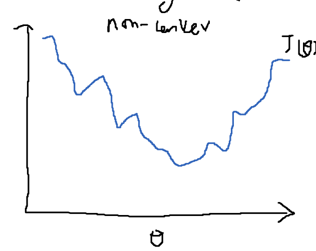
m examples $\quad x \in \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \quad y \in \{0, 1\}$

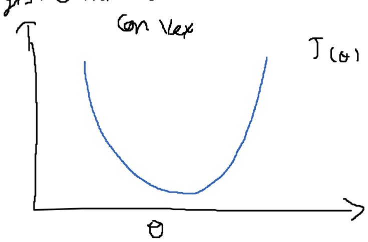
$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Linear regress. $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{\frac{1}{2} \left(h_\theta(x^i) - y^i \right)^2}_{= \, Cost \left(h_\theta(x^i), y^i \right)}$

$$Cost\left(h_\theta(x^i), y^i\right) = \frac{1}{2} \left(h_\theta(x^i) - y^i \right)^2$$

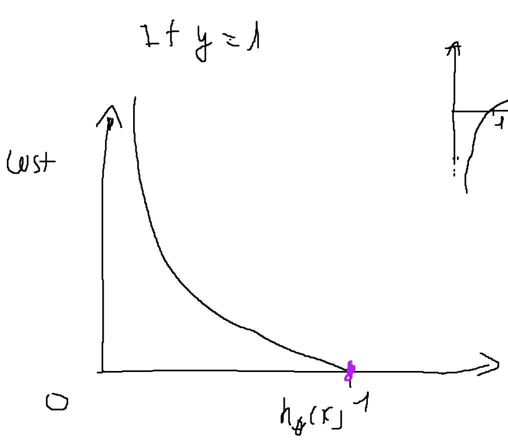
nếu cứ dùng $J(\theta)$ của linear cho logistic thì sẽ là một hàm non-convex

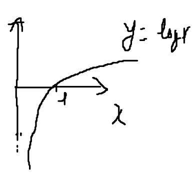


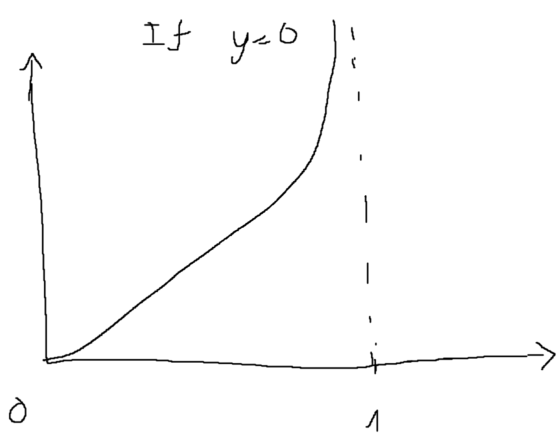


Cost function for logistic

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$







Cost = 0 Khi y = 1 và $h_\theta(x) = 1$

nhưng $h_{\theta(x)} \to 0$ $\Rightarrow$ large cost
cost $\to \infty$

note. nếu $h_\theta(x) = 0$, (Predict $P(y=1|x;\theta)$)
$y = 1$ thì sẽ trả giá lớn

note. nếu $h_\theta(x) = 0$, (Predict $P(y=1|x;\theta)$)
$y=1$ thì sẽ trả giá lớn

$cost(h_\theta(x), y) = 0$ if $h_\theta(x) = y$

$cost(h_\theta(x), y) \to \infty$ if $\begin{cases} y=0, & h_\theta x \to 1 \\ y=1, & h_\theta x \to 0 \end{cases}$

Simplifield cost function

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

$$\Rightarrow J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y\log(h_\theta(x)) + (1-y)(\log(1-h_\theta(x))) = \frac{1}{m}\sum_{i=1}^{m} cost(h_\theta(x), y)$$

tìm bằng Maximum Like lihood estimatino
(MLE)
* còn nhiều cost function khác

$\min_\theta J(\theta)$ tìm θ để dự đoán với x mới

output: $h_\theta(x) = \dfrac{1}{1+e^{-\theta^T x}}$ $\quad P(y=1|x;\theta)$

$$J_\theta(x) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}.\log h(x) + (1-y^{i}).\log(1-h(x^i))$$

dạng Vecto
$h = g(X\theta)$
$J(\theta) = \frac{1}{m}.(-y^T\log h - (1-y)^T\log(1-h))$

Lặp lại đến khi hội tụ

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

$$\Leftrightarrow \theta_j = \theta_j - \alpha.\frac{1}{m}\sum_{i=1}^{m}(h(x^i) - y^i)x^i$$

Vecto

$$\theta = \theta - \frac{\alpha}{m}X^T(g(X\theta) - \vec{y})$$

Prove gradient descent

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

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left(y\log(h_\theta(x)) + (1-y)\log(1-h_\theta(x))\right)$$

Gradient descent

$\min_\theta J(\theta)$.

Repeat {

$$\theta_j = \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta) = \theta_j - \alpha\cdot\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^i) - y^i\right)x_j^{(i)}$$

}

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

Prove

$$\frac{\partial}{\partial\theta}J(\theta) = \frac{\partial}{\partial\theta}\left(-y\log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))\right) \quad \text{With } h_\theta(x) = \frac{1}{1+e^{\theta^T x}} \quad \text{and } h'_\theta(x) = x.(1-h_\theta(x))\,h_\theta(x)$$

$$= -y.x(1-h_\theta) + (1-y)\,h_\theta(x)\,x$$

$$= (-y + yh_\theta(x) - yh_\theta(x) + h_\theta(x)).x$$

$$= (h_\theta(x) - y).x$$

optimization

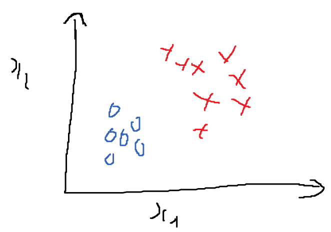
- conjugate gradient
- BFGS
- left BFGS

One vs all

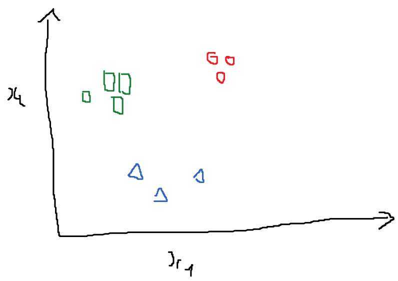
multi class classification

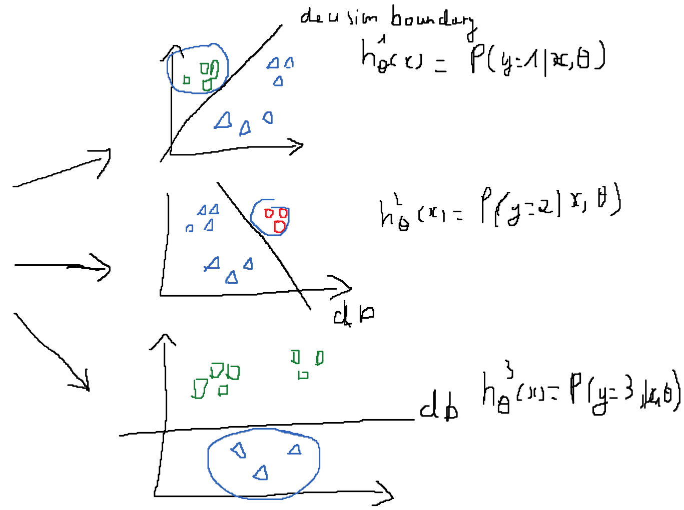
eg: email: work, Friend, family, hobby
$y=1$ 2 3 4

binary classification



multi class

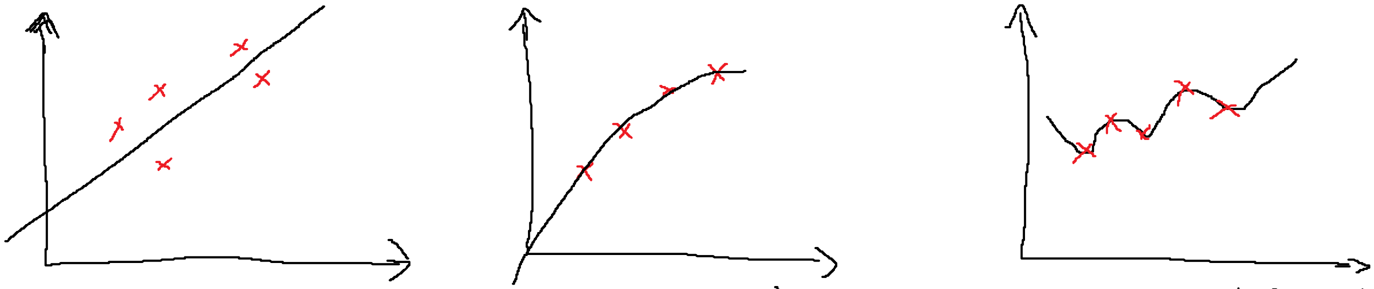




train model lg $h_\theta^{(i)}(x)$ cho tất cả y để dự đoán $y = i$

$$\max_i h_\theta^{(i)}(x)$$

overfitting



$\theta_0 + \theta_1 x$ $\theta_0 + \theta_1 x + \theta_2 x^2$ $\theta_0 + \theta_1 x_1 + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

under fit   high bias   over fitting   high variance

- under fit: mô hình dự đoán sai ở cả train set và thực tế
- over fit: mô hình dự đoán tốt ở trainning set nhưng tệ ở thực tế
    - khi có quá nhiều feature và model fit với data set nhưng dự đoán sai với data mới
- cách xử lý: 1 - giữ lại những feature quan trọng
    - thuật toán lựa chọn mô hình

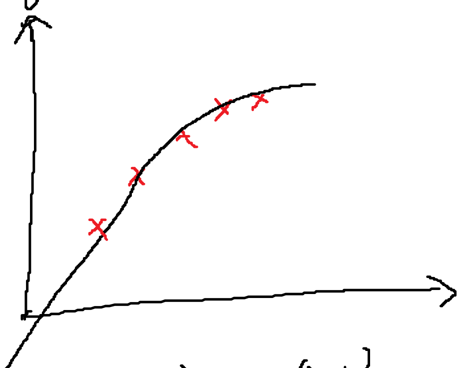
    2 regularization
    - giữ tất cả feature nhưng giảm độ lớn/giá trị của tham số
    - tất cả các feature đóng góp vào việc dự đoán y

Regularization - cost function

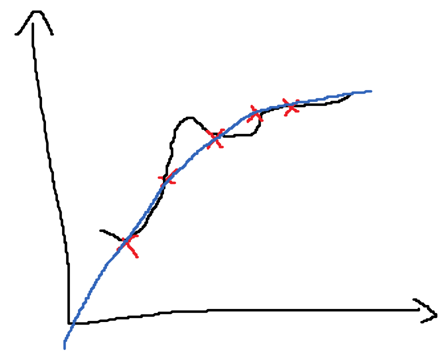
Small value for parameter $\theta_0, \theta_1, .., \theta_n$ (Penalize)

- "simpler" hypothesis
- Less prone to over fitting

eg



$\theta_0 + \theta_1 x + \theta_2 x^2$



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

when $\theta_3, \theta_4 \approx 0$

$\Leftrightarrow \theta_0 + \theta_1 x + \theta_2 x^2$

housing.

- Feature $x_1, x_2, .. x_{100}$
- parameter: $\theta_0, \theta_1, \theta_2, .. \theta_{100}$

$$J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^{m} \left(h_\theta(x^i) - y^i \right)^2 + \lambda \sum_{j=1}^{m} \theta_j^2 \right]$$

regularization Parameter

$$\sum_{j=1}^{m} \theta_j^2 = \|\theta\|_2^2$$

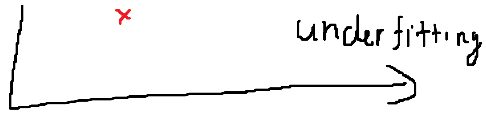
$$\min_\theta J(\theta)$$

nếu λ quá lớn $\Leftrightarrow \theta_1, \theta_2, \theta_3, \theta_4, ... \approx 0$ (quá nhỏ)

model

$h_\theta(x) = \theta_0$

underfitting


underfitting

Regularizaion for Linear

Regularization

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=0}^{m}\left(h_\theta(x^i) - y\right)^2 + \sum_{j=1}^{m}\theta_j^2\right]$$

$$\min_\theta J(\theta)$$

Gradient desent

Repeat {

$$\theta_0 = \theta_0 - \alpha\frac{1}{m}\sum_{i=0}^{m}\left(h_\theta(x^i) - y^i\right)x_0^{(i)}$$

$$\theta_j = \theta_j - \alpha\left[\frac{1}{m}\sum_{i=0}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right) \cdot x_j^{(i)} + \frac{\lambda}{m}\theta_j\right]$$

$$\Updownarrow \qquad \frac{\partial}{\partial\theta_j}J(\theta) \qquad j = (1, 2, \ldots, n$$

$$\theta_j = \underbrace{\theta_j\left(1 - \alpha\cdot\frac{\lambda}{m}\right)}_{\text{thing } < 1} - \alpha\cdot\frac{1}{m}\sum_{i=0}^{m}\left(h_\theta(x^i) - y^i\right)x_j^i$$

Normal equation

$$X = \begin{bmatrix} (x^1)^T \\ (x^2)^T \\ \vdots \\ (x^{m})^T \end{bmatrix}_{m \times (n+1)} \qquad y = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{bmatrix} \quad R^m$$

$$\theta = \left(X^T X + \lambda\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots \\ \vdots & 0 & 1 & & \\ 0 & 0 & \cdots & & 1 \end{bmatrix}\right)^{-1} \cdot X^T Y$$

$$\theta = \left(X^T X + \lambda \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & & \\ 0 & 0 & \cdots & & 1 \end{bmatrix} \right) \cdot X^T y$$

Regularization for logistic

$$J(\theta) = -\left[\frac{1}{m} \sum_{i=1}^{m} y^i \log(h_\theta(x^i)) + (1-y^i) \log(1-h_\theta(x^i)) + \frac{\lambda}{m} \sum_{j=1}^{m} \theta_j^2 \right]$$

Gradient decent

$$\left\{ \theta_0 = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i) x_0^i \right.$$

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