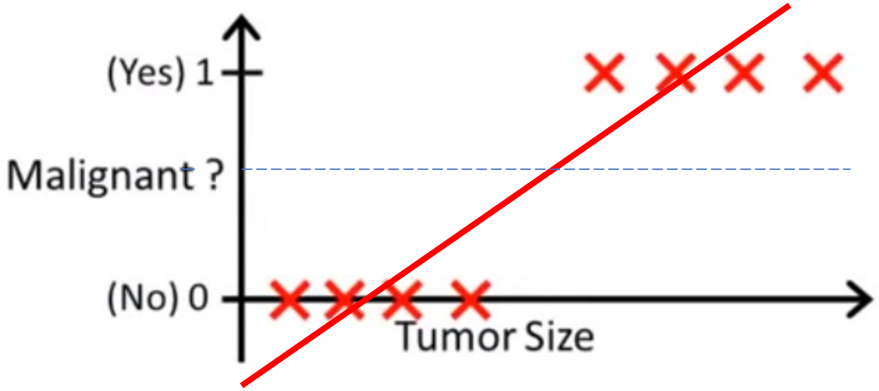
Logistic Regression
(supervised learning for classification)

Classification problem



Use linear regression, $h_\theta(x) = x^T\theta$?

Set a threshold for classification (at 0.5):

If $h_\theta(x) \geq 0.5, \text{predict "y} = 1\text{"}$

If $h_\theta(x) \leq 0.5, \text{predict "y} = 0\text{"}$

Inconvenience of linear regression:

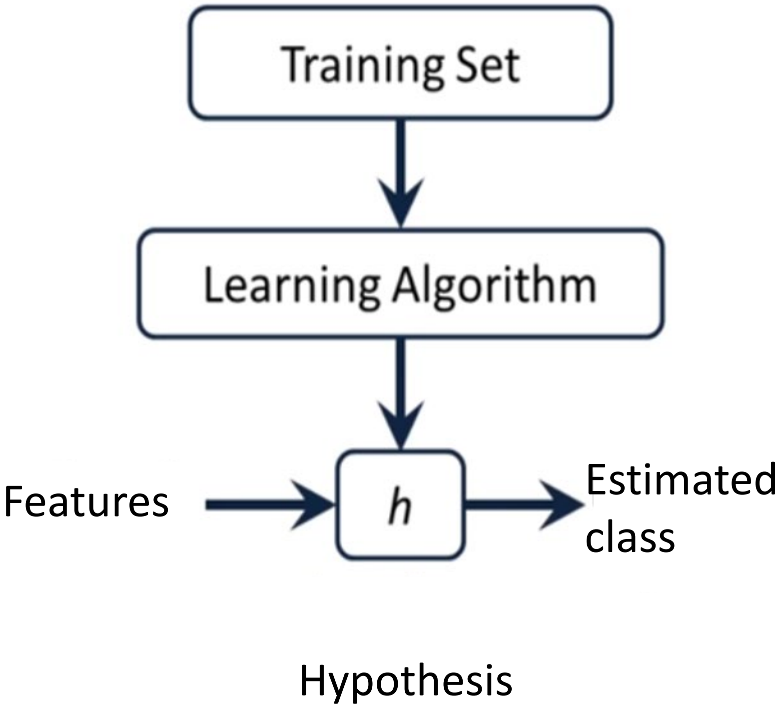
Classification: y = 0 or 1

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

Introduce:

Logistic Regression: $0 \leq h_\theta(x) \leq 1$

Supervised learning – classification problem



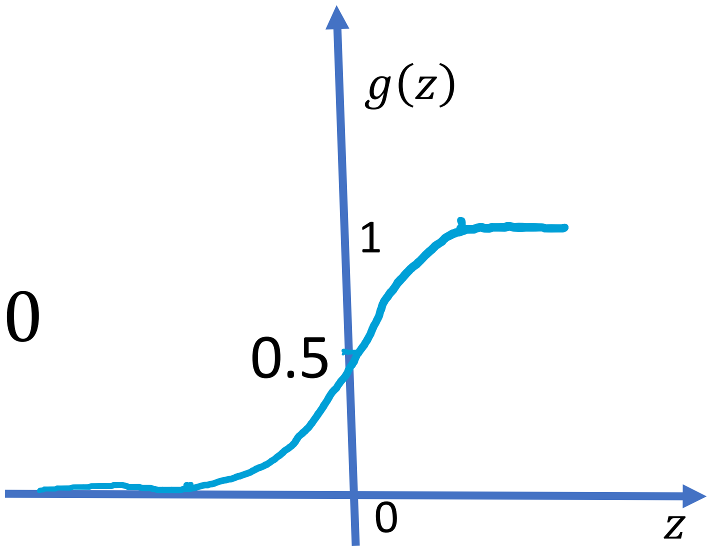
Hypothesis $h_\theta(x)$:

x: features

θ: parameter of model

Logistic regression model:

Hypothesis $h_\theta(x)$:
estimated probability that $y = 1\ or\ 0$
given input x, $h_\theta(x) = P(y|x, \theta)$



Sigmoid/logistic function

Realized by a learning model (logistic regression model)

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

thus $0 \le h_\theta(x) \le 1$

$$h_\theta(x) = P(y|x, \theta)$$

- Probability of predicting y, given x, parameterized by θ

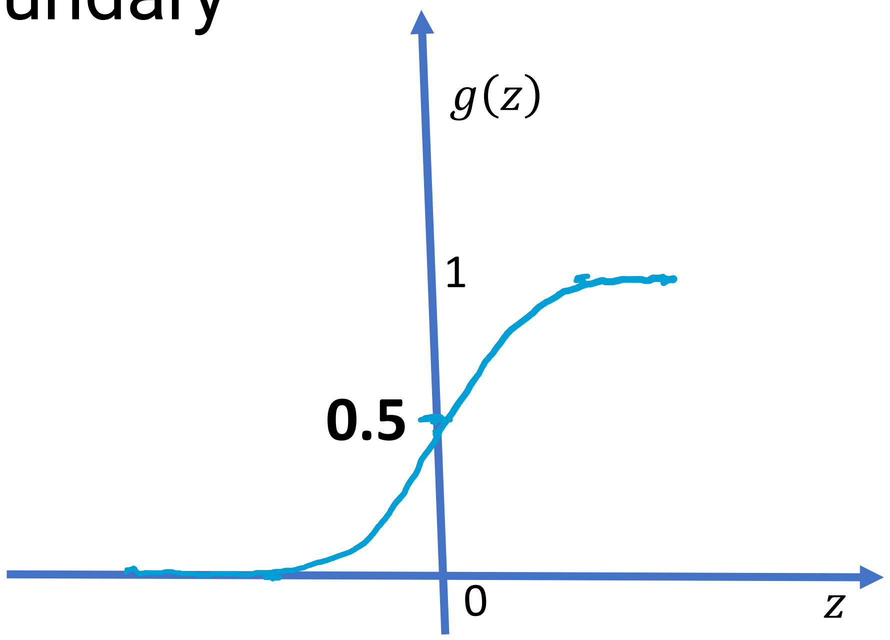
  Predict $y = 1$ if $h_\theta(x) \geq 0.5$
  Predict $y = 0$ if $h_\theta(x) < 0.5$

- For example, $h_\theta(x) = 0.85$, tell patient that 85% of chance of tumor being cancerous

Decision Boundary

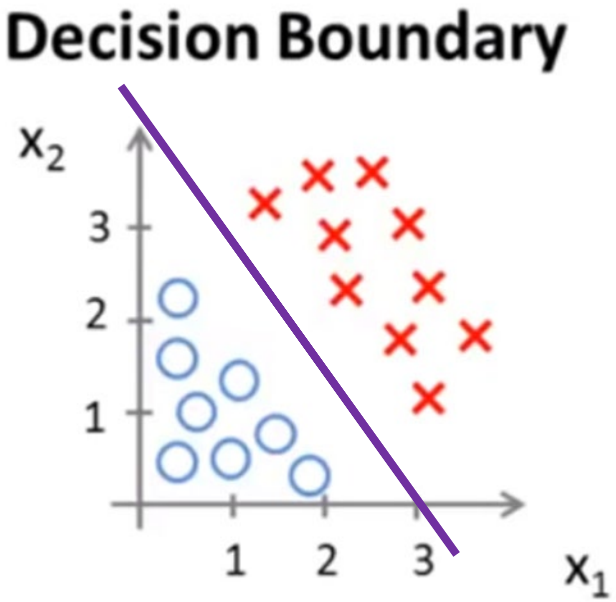
$$h_\theta(x) = g(x^T\theta)$$

(Let $z = x^T\theta$)



Predict $y = 1$ if $h_\theta(x) \geq 0.5$ *or* $z \geq 0$
Predict $y = 0$ if $h_\theta(x) < 0.5$ *or* $z < 0$

Decision boundary: $x^T\theta = 0$

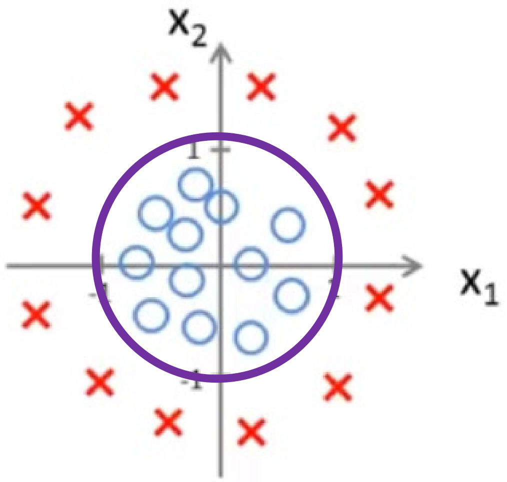


$$h_\theta(x) = g(-3 + x_1 + x_2)$$

Predict $y = 1$ if $h_\theta(x) \geq 0.5$ $or (-3 + x_1 + x_2) \geq 0$
Predict $y = 0$ if $h_\theta(x) < 0.5$ $or (-3 + x_1 + x_2) < 0$

Decision boundary:
$\boldsymbol{x_1 + x_2 = 3}$

Nonlinear decision boundaries



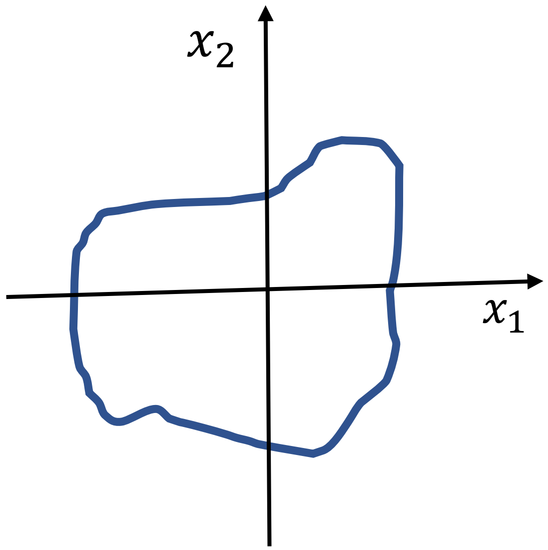
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

θ=[-1, 0, 0, 1, 1]

Decision boundary:

$$\boldsymbol{x_1}^2 + \boldsymbol{x_2}^2 = 1$$

More complicated nonlinear decision boundary



$$h_\theta(x) = g\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \cdots\right)$$

Decision boundary:

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \cdots = 0$$

Given training set for supervised learning:

$$\left\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \cdots, \left(x^{(m)}, y^{(m)}\right)\right\}$$

where

$$x \in \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_n \end{bmatrix} \qquad x_0 = 1, y \in \{0,1\}$$

In a logistic regression model,

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

How to choose parameter θ?

Cost function consideration:

If we use the cost formulation as in linear regression,

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

Recall:

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

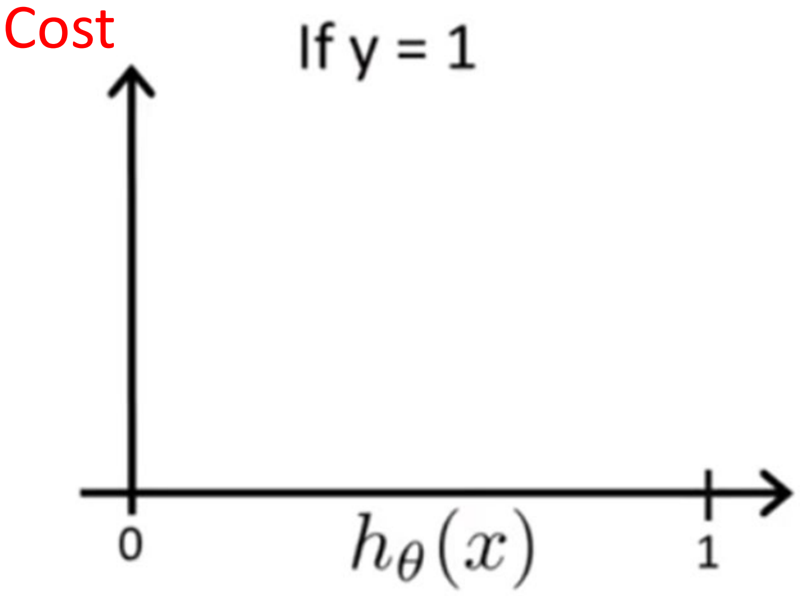
g is a sigmoid/logistic function

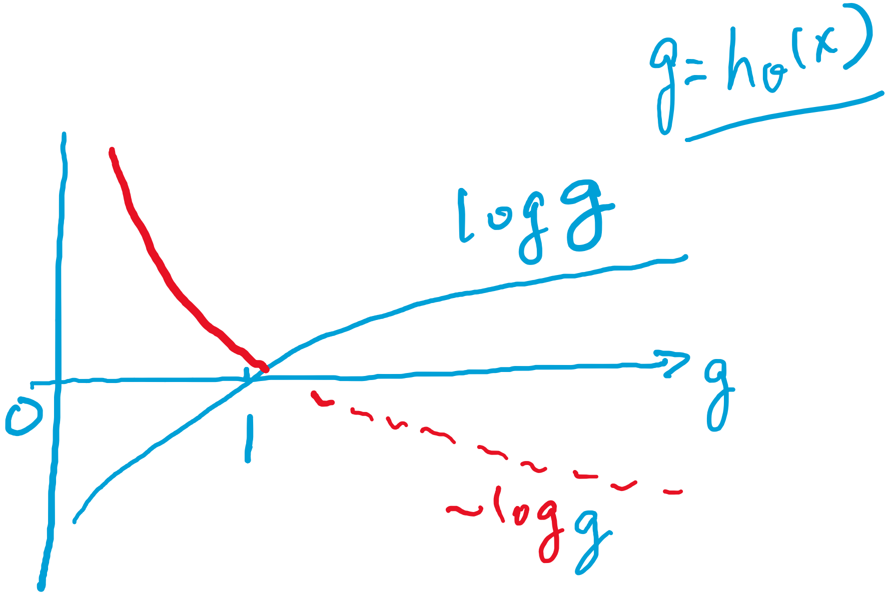
Look at the term:

$$\frac{1}{2}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 \Longrightarrow \text{Nonconvex, thus local minima}$$

Propose “logistic regression cost function” as follows:

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



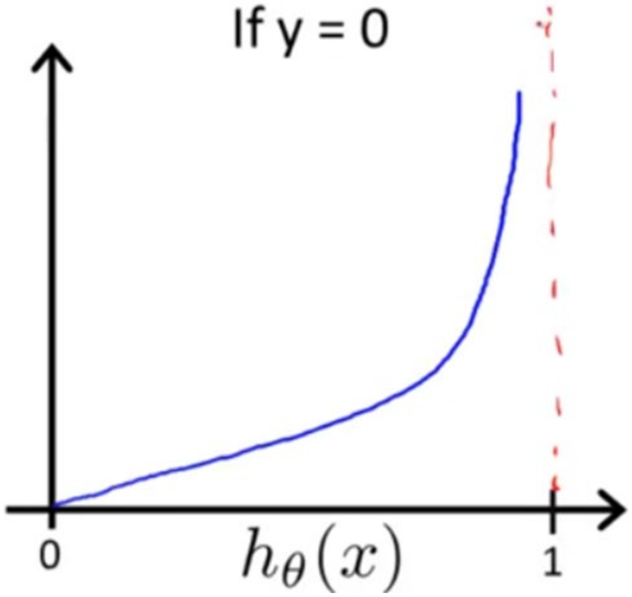


$$Cost\ = 0\ if\ y = 1, h_\theta(x) = 1$$

$$Cost\ \to \infty\ \ if\ y = 1, but\ \ h_\theta(x) \to 0$$

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

Cost



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

$$Cost\ \to \infty\ \ if\ y = 0, but\ \ h_\theta(x) \to 1$$

Logistic regression cost function

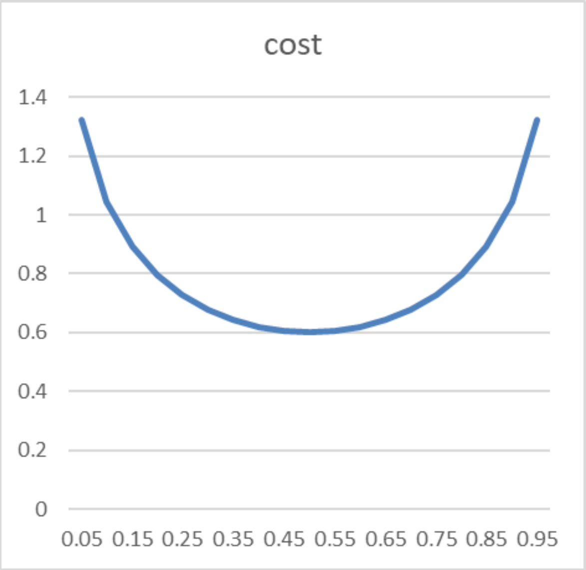
$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right)$$

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

Note: $y = 0$ or 1

Simplification:

$$\Rightarrow \quad \mathrm{Cost}(h_\theta(x), y) = -y\log\left(h_\theta(x)\right) - (1-y)\log\left(1 - h_\theta(x)\right)$$



Logistic regression cost function

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

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

To fit parameters θ:

$$\min_{\theta} J(\theta)$$

⇨ To get parameters

$$\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_n)$$

To make a prediction given new x:

Compute output:

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

Apply gradient descent to reduce the cost measure

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

Update parameters $\theta_0\ \theta_1 \ldots \theta_n$ using gradient descent until convergence

Simultaneously update θ_j , $j = 0, 1, \ldots, n$, according to

$$\theta_j \coloneqq \theta_j - \eta \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1, \ldots, \theta_n)$$

$$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j{}^{(i)}$$