CS60050
Machine Learning

Logistic Regression

Somak Aditya
Assistant Professor

Sudeshna Sarkar

Department of CSE, IIT Kharagpur

August 11, 2023

Logistic Regression

for Classification

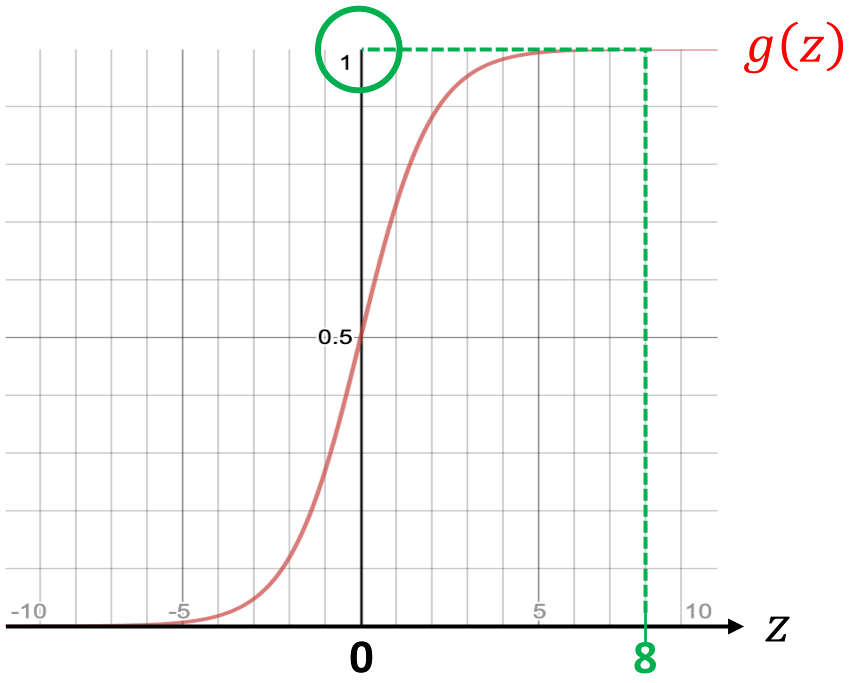
Regression vs. Classification

We want the possible outputs of $\boldsymbol{f_{\theta}(x)} = \boldsymbol{\theta^T x}$ to be discrete-valued

Use an ***activation function*** (e.g., ***sigmoid or logistic function***)

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

$z \in \mathbb{R}, but$
$g(z) \in [0,1]$



If y = **1**, we want $g(z) \approx 1$ (i.e., we want a correct prediction)
For this to happen, $\boldsymbol{z} \gg \boldsymbol{0}$

If y = **0**, we want $g(z) \approx 0$ (i.e., we want a correct prediction)
For this to happen, $\boldsymbol{z} \ll \boldsymbol{0}$

Classification

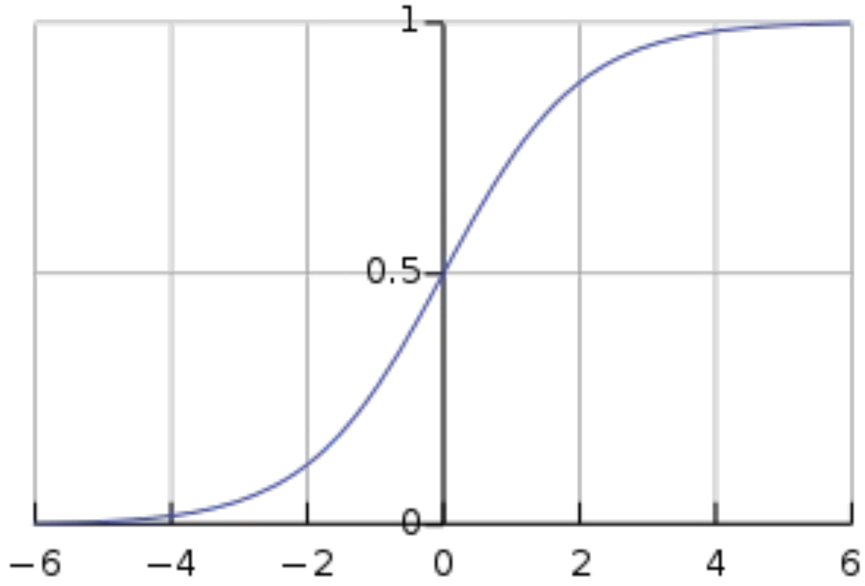
$$\boldsymbol{x} = [\boldsymbol{x}_0, \dots, \boldsymbol{x}_m]$$

$$g(\boldsymbol{\theta}^T \boldsymbol{x})$$

$$\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}) \in [0,1]$$

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

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



Thresholding:
predict "y = 1" if $h_\theta(x) \geq 0.5$

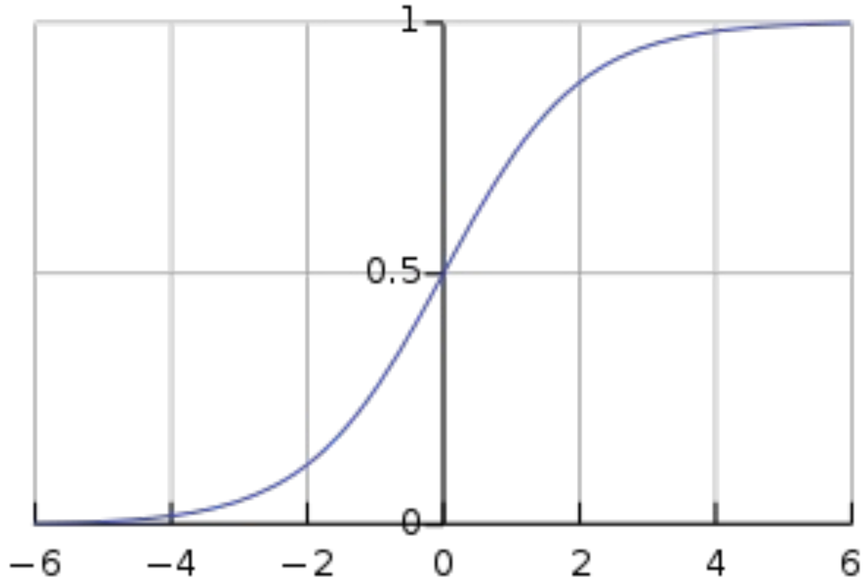
predict "y = 0" if $h_\theta(x) < 0.5$

Classification

$\boldsymbol{x} = [\boldsymbol{x_0}, \ldots, \boldsymbol{x_m}]$ → $\boldsymbol{g(\theta^T x)}$ → $\boldsymbol{h_\theta(x)} \in [0,1]$

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

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



Thresholding:

predict "y = 1" if $h_\theta(x) \geq 0.5$

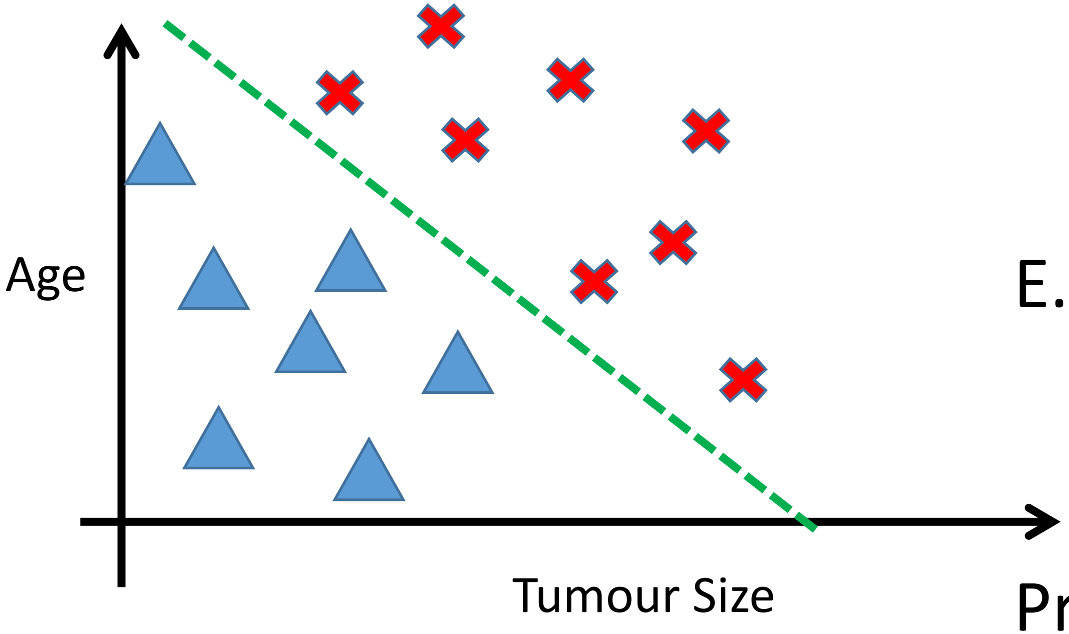
$\boldsymbol{z = \theta^\top x \geq 0}$

predict "y = 0" if $h_\theta(x) < 0.5$

$\boldsymbol{z = \theta^\top x < 0}$

Alternative Interpretation: $h_\theta(x) =$ estimated probability that $y = 1$ on input x

Decision boundary



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

E.g., $\theta_0 = -3, \theta_1 = 1, \ \theta_2 = 1$

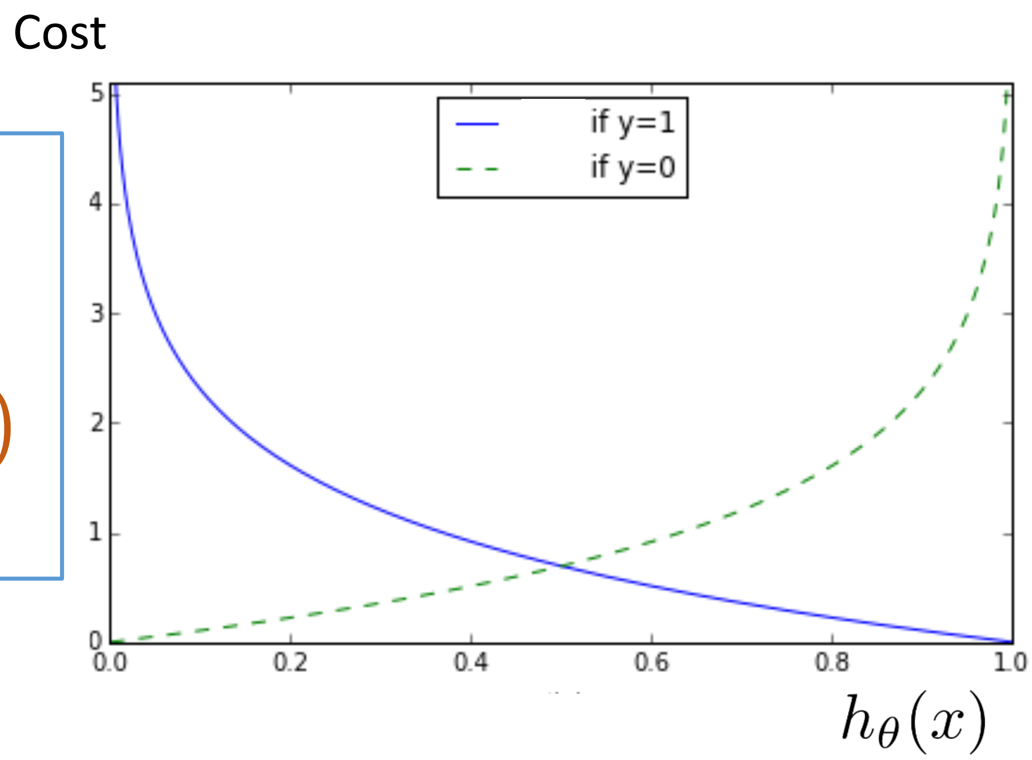
Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

Cost function for Logistic Regression

Logistic Regression

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

$$= -\boldsymbol{y}\ \mathbf{log}(\boldsymbol{h_\theta}(\mathbf{x})) - (\mathbf{1} - \mathbf{y})\ \mathbf{log}(\mathbf{1} - \boldsymbol{h_\theta}(\boldsymbol{x}))$$



$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(\boldsymbol{h}_\theta(x^{(i)}), y^{(i)}))$$

$$= -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\ \log\left(\boldsymbol{h}_\theta\left(x^{(i)}\right)\right) + (1 - y^{(i)}) \log\left(1 - \boldsymbol{h}_\theta\left(x^{(i)}\right)\right)\right]$$

Gradient descent

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

Goal: $\min_{\theta} loss(\theta)$

Good news: Convex function!
Bad news: No analytical solution

Gradient descent

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

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

Gradient descent

Repeat {

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

}

(Simultaneously update all θ_j)

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

Gradient descent for Linear Regression

Repeat {

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

}

$$\boxed{h_\theta(x) = \theta^\top x}$$

Gradient descent for Logistic Regression

Repeat {

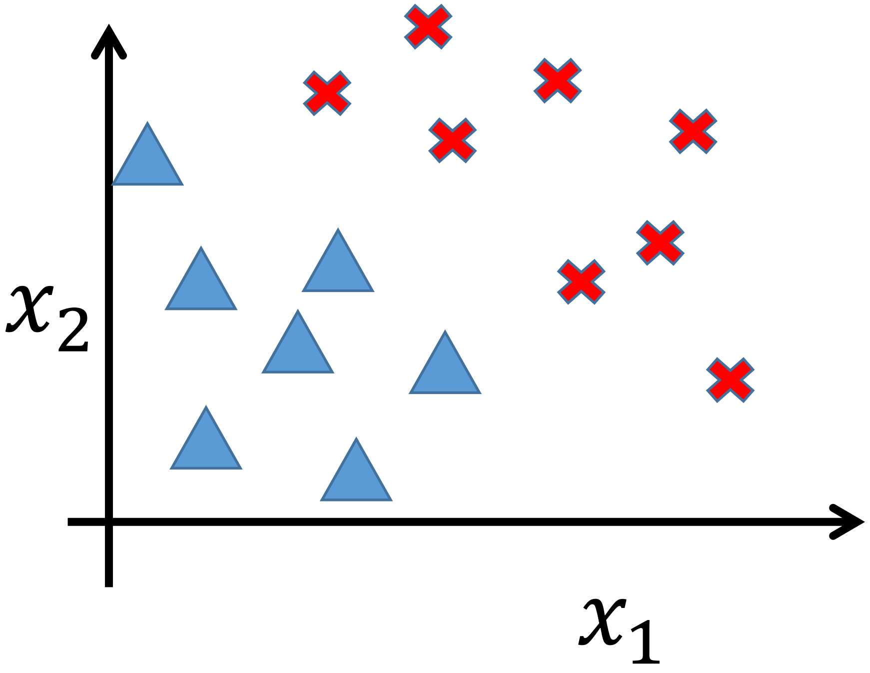
$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$$

}

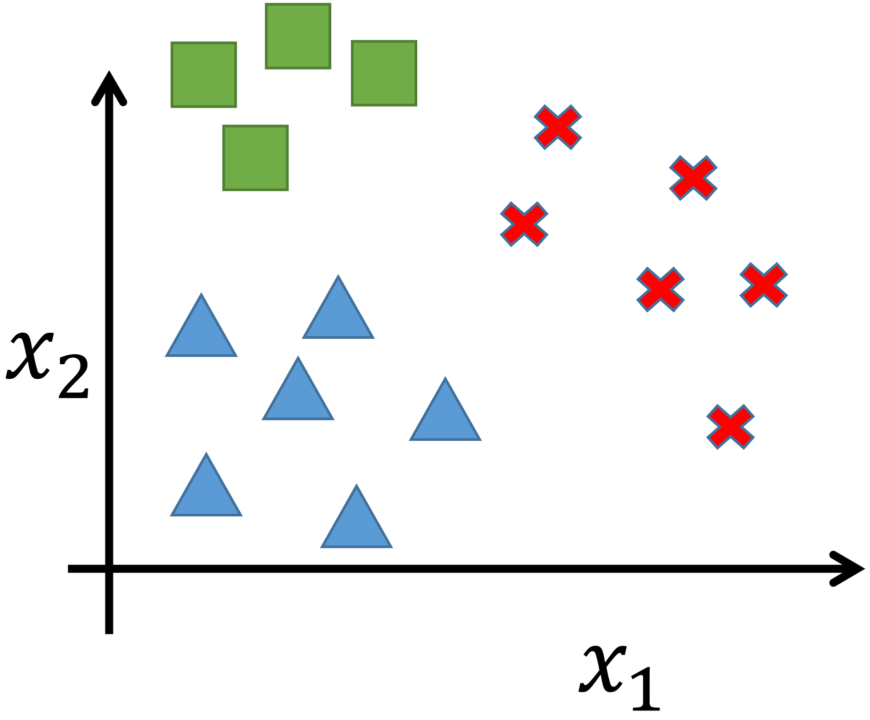
$$\boxed{h_\theta(x) = \frac{1}{1 + e^{-\theta^\top x}}}$$

Multiclass classification

Binary classification



Multiclass classification



Multi-class Classification

- Multi-class Classification: y can take on K different values $\{1,2,\ldots,k\}$
- $f_\theta(x)$ estimates the probability of belonging to each class

$$P(y=k|x,\theta) \propto \exp\left(\theta_k^T x\right)$$

$$\theta = \begin{bmatrix} \vdots & \vdots & \vdots \\ \theta_1 & \theta_2 & \theta_k \\ \vdots & \vdots & \vdots \end{bmatrix}$$

$$P(y=k|x,\theta) = \frac{\exp\left(\theta_k^T x\right)}{\sum_{j=1}^{K} \exp\left(\theta_j^T x\right)}$$

$$\boxed{\mathrm{J}(\theta) = -\left[\sum_{i=1}^{m} \sum_{j=1}^{K} 1\{y^{(i)} = k\} \log \frac{\exp\left(\theta_k^T x^{(i)}\right)}{\sum_{j=1}^{K} \exp\left(\theta_j^T x^{(i)}\right)}\right]}$$