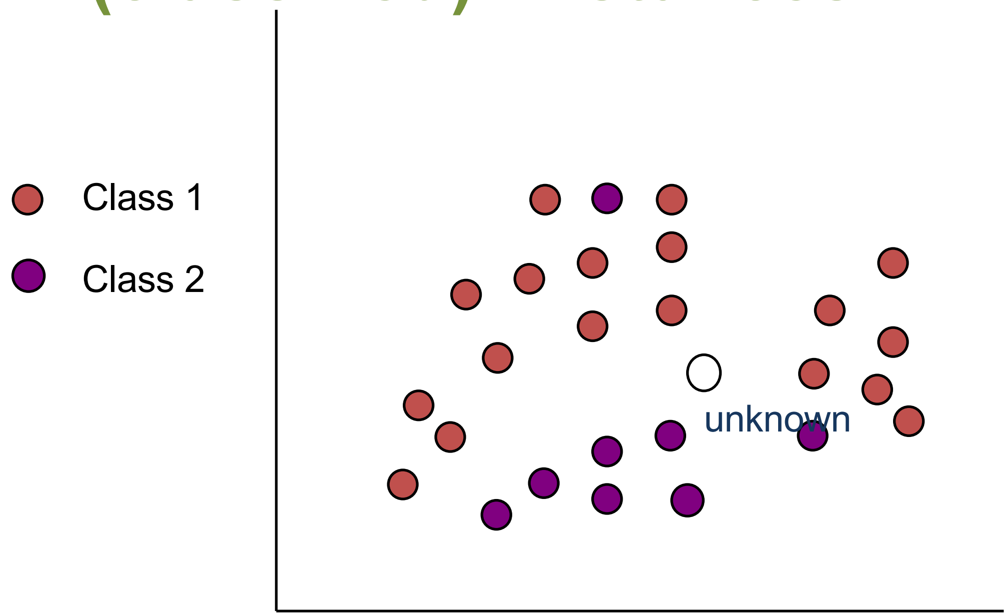
Data Mining

Logistic Regression

Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances



Many approaches: Regression, Decision Trees, Nearest Neighbor, Support Vector Machines, Neural Networks,
...

Classification Problem

Loan approval problem with a single variable

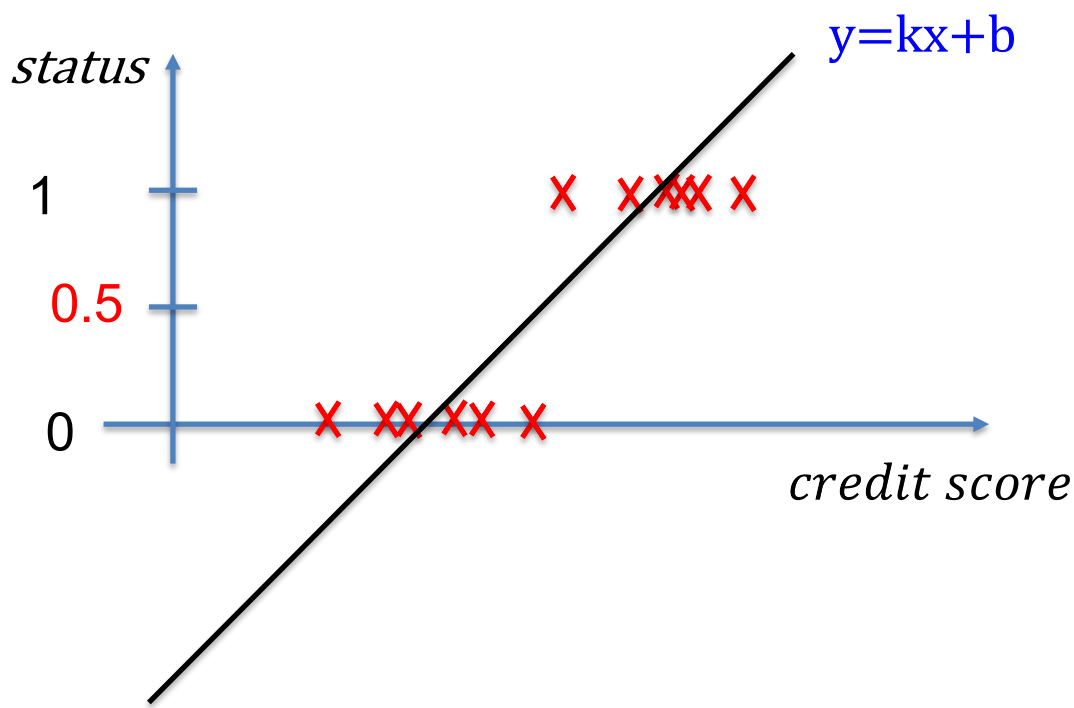
x_1: credit score (FICO score)
y: 1-approve, 0-deny

Credit Score	Loan Status
750	1
725	0
700	0
650	0
726	1
645	0
800	1
...	...

Classification Problem

Loan approval problem with a single variable

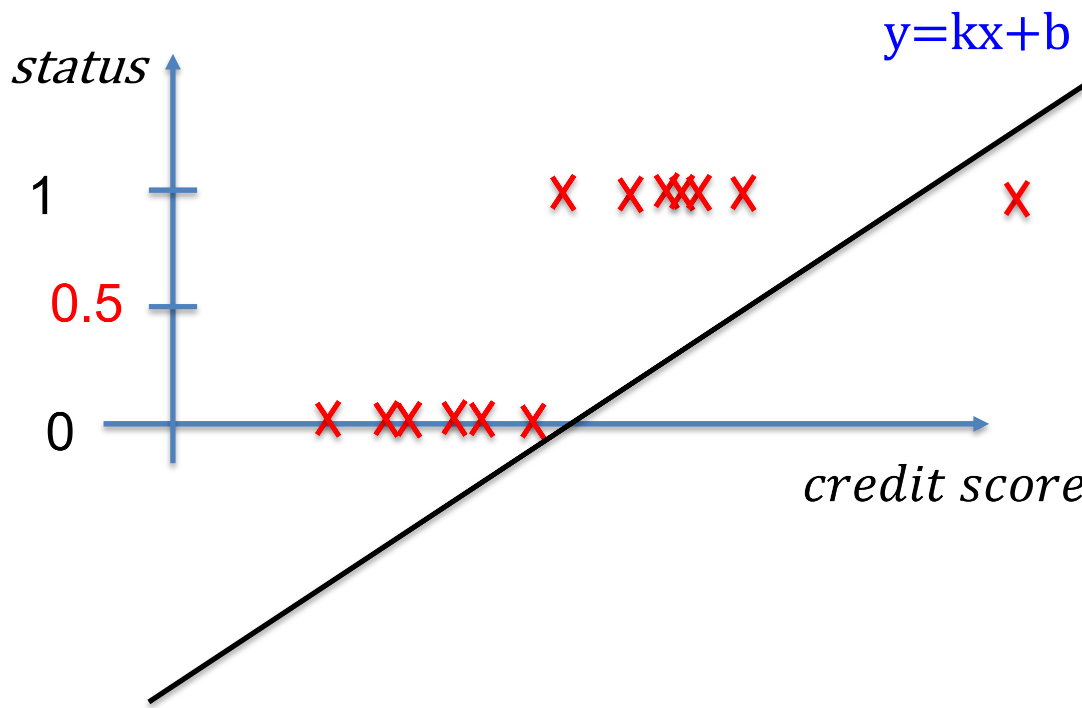
Credit Score	Loan Status
750	1
725	0
700	0
650	0
726	1
645	0
800	1
...	...



Classification Problem

Loan approval problem with a single variable

Credit Score	Loan Status
750	1
725	0
700	0
650	0
726	1
645	0
800	1
...	...



Classification Problem

Loan approval problem

x_1: credit score (FICO score)
x_2: income
(may include other features)

y: 1-approve, 0-deny

Training Data

Credit Score	Income	Loan Status
750	113000	1
725	26000	0
700	54000	0
650	45000	0
726	89500	1
645	78500	0
800	87050	1
...	...	...

Test data:
for a new applicant with credit score 715 and income 68500, will the loan application be approved?

Binary Classification Data

Given:

Training data set:

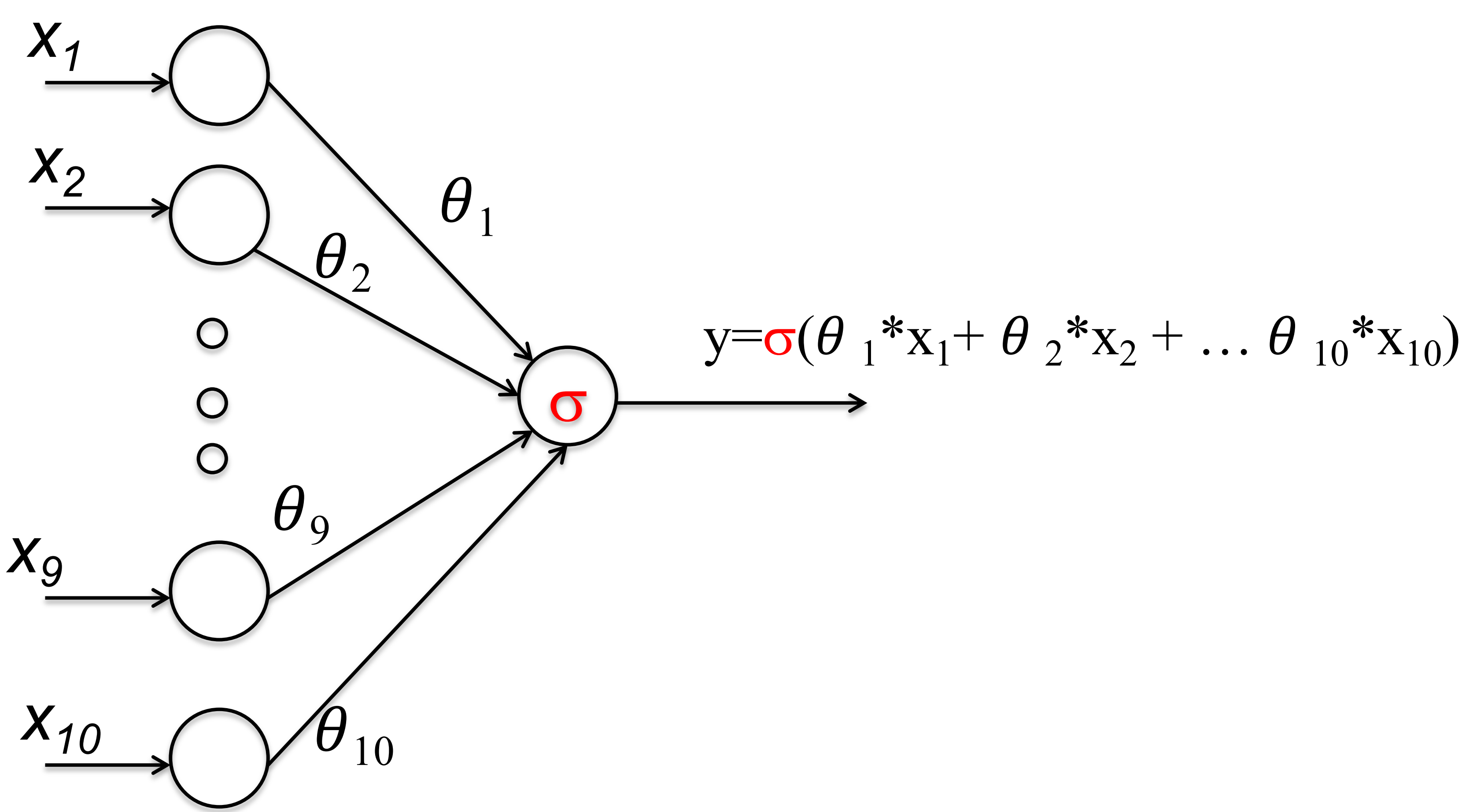
$\{ \{x^1, y^1\},$
$\{x^2, y^2\},$
$\{x^3, y^3\},$
$\ldots$
$\{x^m, y^m\}\}$

$$\mathrm{x} = \begin{bmatrix} x_0 \\ x_1 \\ \ldots \\ x_n \end{bmatrix}$$

$x_0=1$, $y \in \{0, 1\}$

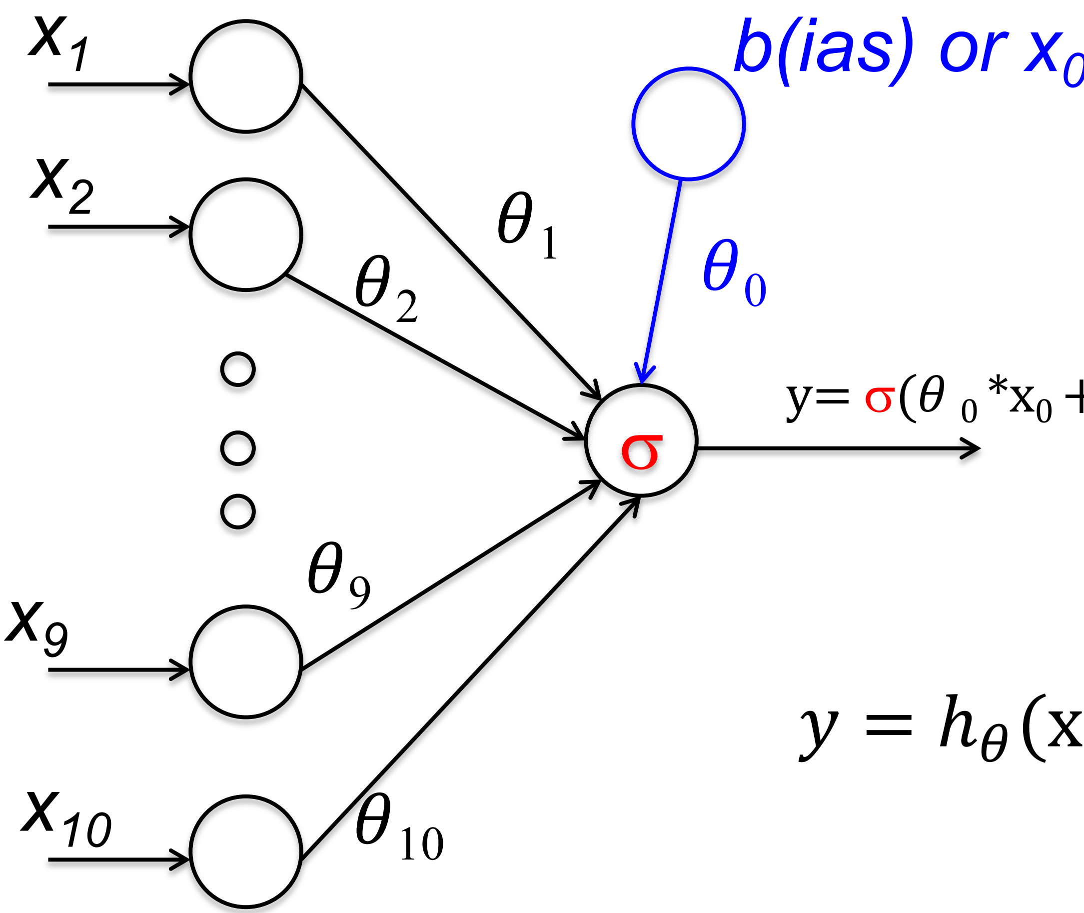
m examples

Logistic Regression For Binary Classification (1)



10 features

Logistic Regression For Binary Classification (2)



$$y = \sigma(\theta_0 * x_0 + \theta_1 * x_1 + \theta_2 * x_2 + \ldots \theta_{10} * x_{10})$$

$$y = h_\theta(\mathrm{x}) = \sigma\left(\sum_{i=0}^{10} \theta_i * x_i\right)$$
$$= \sigma(\theta^{\mathrm{T}}\mathrm{x})$$

10 features

Activation Function σ

- Tanh()

$$f(x) = \frac{2}{1+e^{-2x}} - 1$$

$$f'(x) = 1 - f(x)^2$$

- Sigmoid/Logistic

$$f(x) = \frac{1}{1+e^{(-x)}}$$

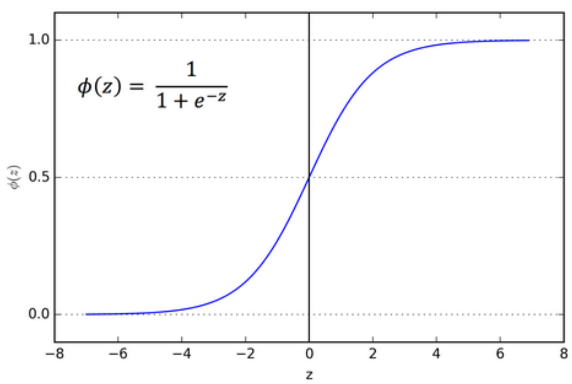
$$f'(x) = f(x)[1 - f(x)]$$

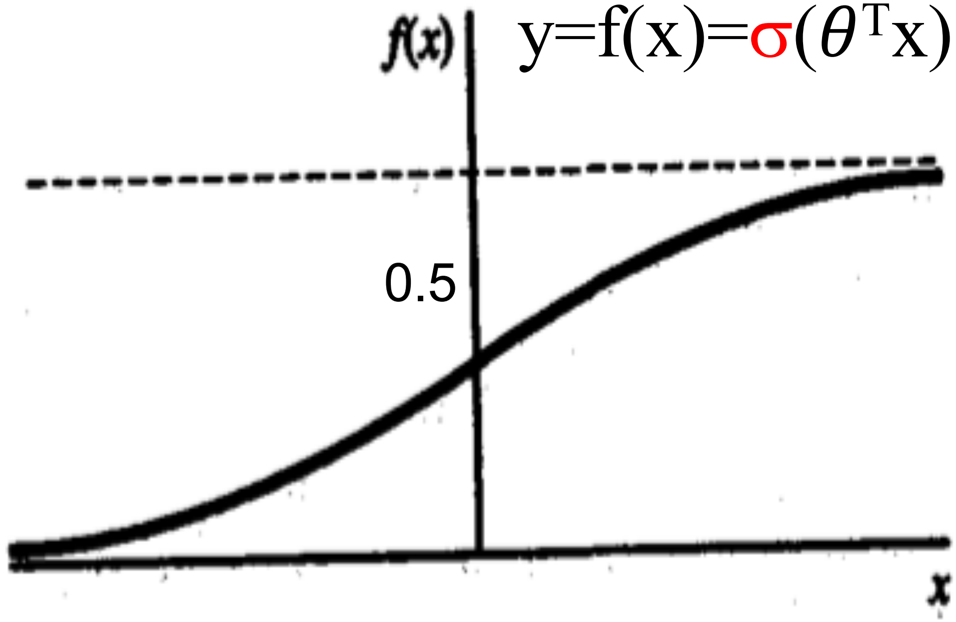
- Bipolar Sigmoid

$$f(x) = \frac{2}{1+e^{(-x)}} - 1$$

$$f'(x) = \frac{1}{2}[1 + f(x)][1 - f(x)]$$

Sigmoid Function for Classification





if $\sigma(\theta^T x) < 0.5$,
predict class 0

if $\sigma(\theta^T x) > 0.5$,
predict class 1

($\theta^T x < 0$,
predict class 0)

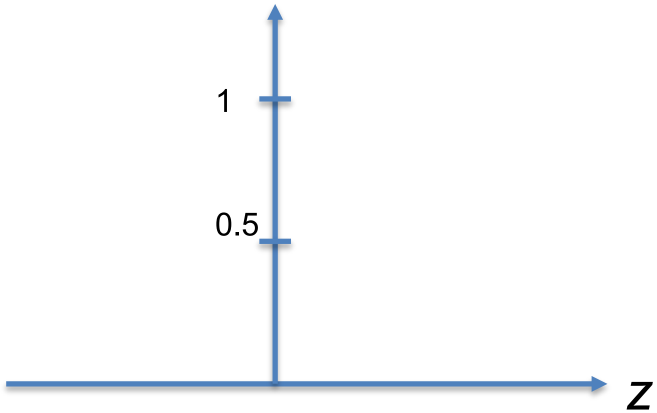
($\theta^T x \geq 0$,
predict class 1)

Logistic Regression Model

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

let $z = \theta^T x$, $\sigma(z) = \frac{1}{1+e^{-z}}$

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



Sigmoid/logistic function

$h_\theta(x)$: $estimated\ probablity\ that\ y = 1\ on\ input\ x$

p(y=1 | x, θ)

p(y=0 | x, θ) = 1 - p(y=1 | x, θ)

How to use it in credit assignment or medical diagnosis problems?

Estimate the Parameters θ

Given:
Training data set:

$\{ \{x^1, y^1\},$
$\{x^2, y^2\},$
$\{x^3, y^3\},$
...
$\{x^m, y^m\}\}$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix}$$

x_0=1, y∈ $\{0, 1\}$

m examples

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

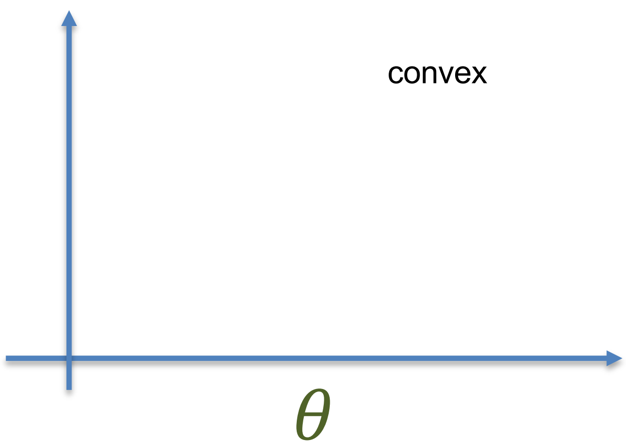
How to estimate the parameters θ from data?

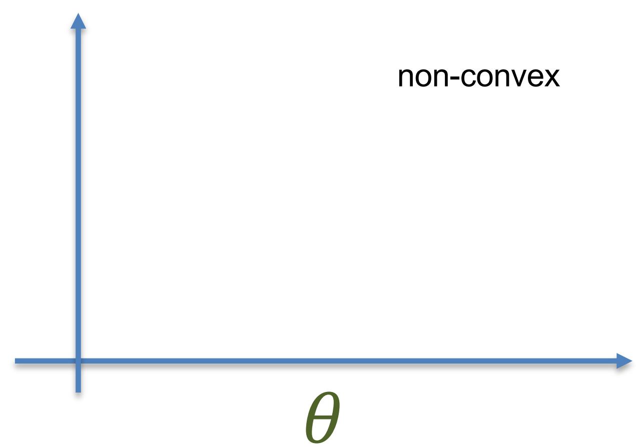
Cost Function

- Linear Regression:

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

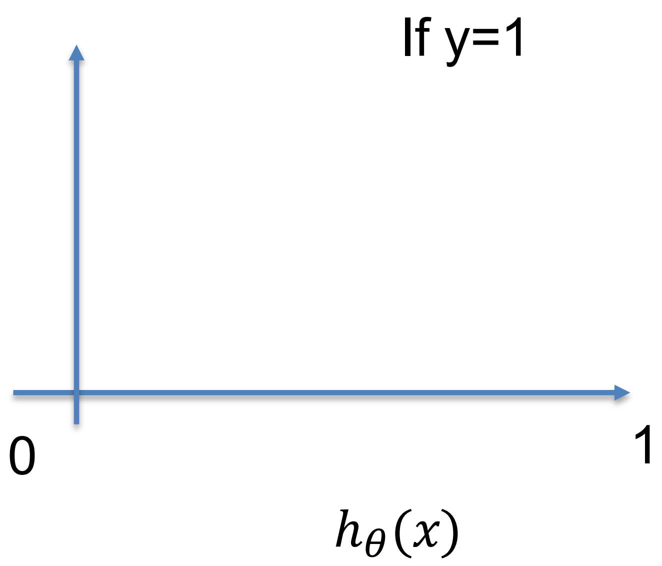
- In logistic regression, $(h_\theta(x^{(i)}) - y^{(i)})^2$ is not a convex curve, not suitable for gradient descent approximation approach.

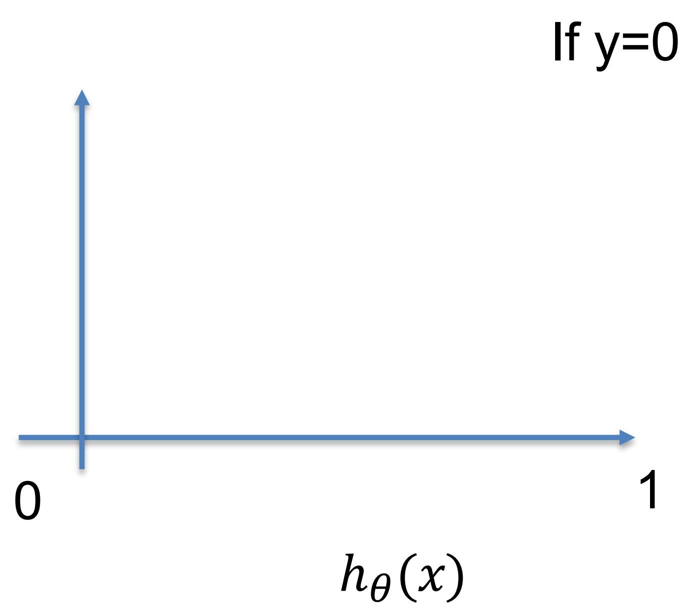




Logistic Regression Cost Function

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





Logistic Regression Cost Function

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

Combine these two into one single cost function:

$$Cost(h_\theta(x), \text{y}) = -y * \log(h_\theta(x)) - (1\text{-y})^* \log(1 - h_\theta(x))$$

If y=1, $Cost(h_\theta(x), \text{y}) = -\log(h_\theta(x))$
If y=0, $Cost(h_\theta(x), \text{y}) = -\log(1 - h_\theta(x))$

Gradient Descent

- *To minimize the Cost function:*

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

- To minimizing the cost function over the entire data set
  - Generally, there is no closed form solution for this minimization problem, except for special cases
  - Approach: Gradient descent

Repeat for each iteration:

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

where:

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

Weight Updates with Gradient Descent

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

Want to minimize $J(\theta)$:

Repeat for each iteration:

$$\theta_j := \theta_j \text{ - } \lambda \sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

Simultaneously update all θ_j

λ: Learning Rate → step size

Cross Entropy Error Cost Function

- Logistic Regression Error
  - 0 if correct, >0 if not correct, more wrong → bigger cost
- Cross-Entropy Error cost function

$$Cost(h_\theta(x), \mathrm{y}) = -y * \log(h_\theta(x)) - (1\text{-y})^* \log(1 - h_\theta(x))$$

y is the target, $h_\theta(x)$ is the predicted value

y	$h_\theta(x)$	cost
1	1	0
0	0	0
1	0.9	0.11
1	0.5	0.69
1	0.1	2.3