ED5340 - Data Science: Theory and Practise

L19 - Logistic Regression (Credit to Andrew Ng)

Ramanathan Muthuganapathy (https://ed.iitm.ac.in/~raman)
Course web page: https://ed.iitm.ac.in/~raman/datascience.html
Moodle page: Available at https://courses.iitm.ac.in/

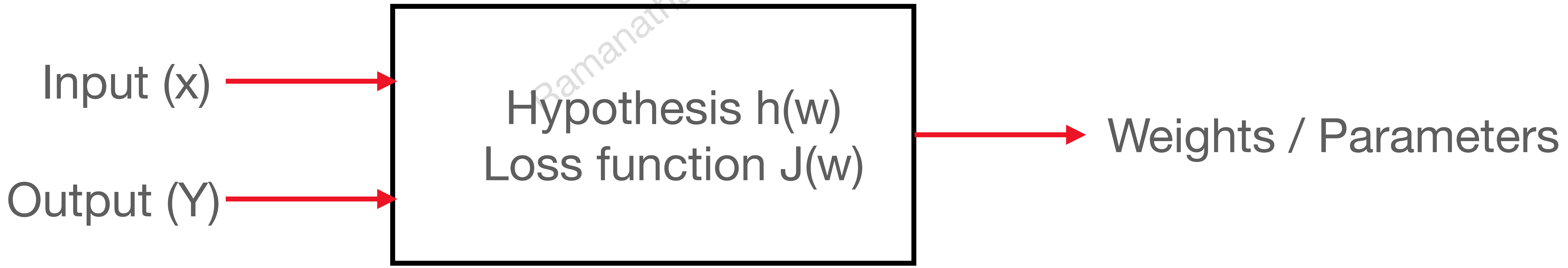
Linear Regression

Predictive problem - Continuous input / output

- Ground truth data - Input feature / output $(\mathbf{x}, \mathbf{y})$ are the knowns
- Use a model / hypothesis as $h(w)$
- Develop an error / cost / loss function $J(w) = J(\mathbf{y}, \bar{\mathbf{y}}) = J(\mathbf{y}, h(w))$
- The weights are identified by
  - min $J(w)$
- Essentially, ML problem is now reduced to an optimization problem.
- Weights are identified using Optimization.

Linear Regression
Predictive

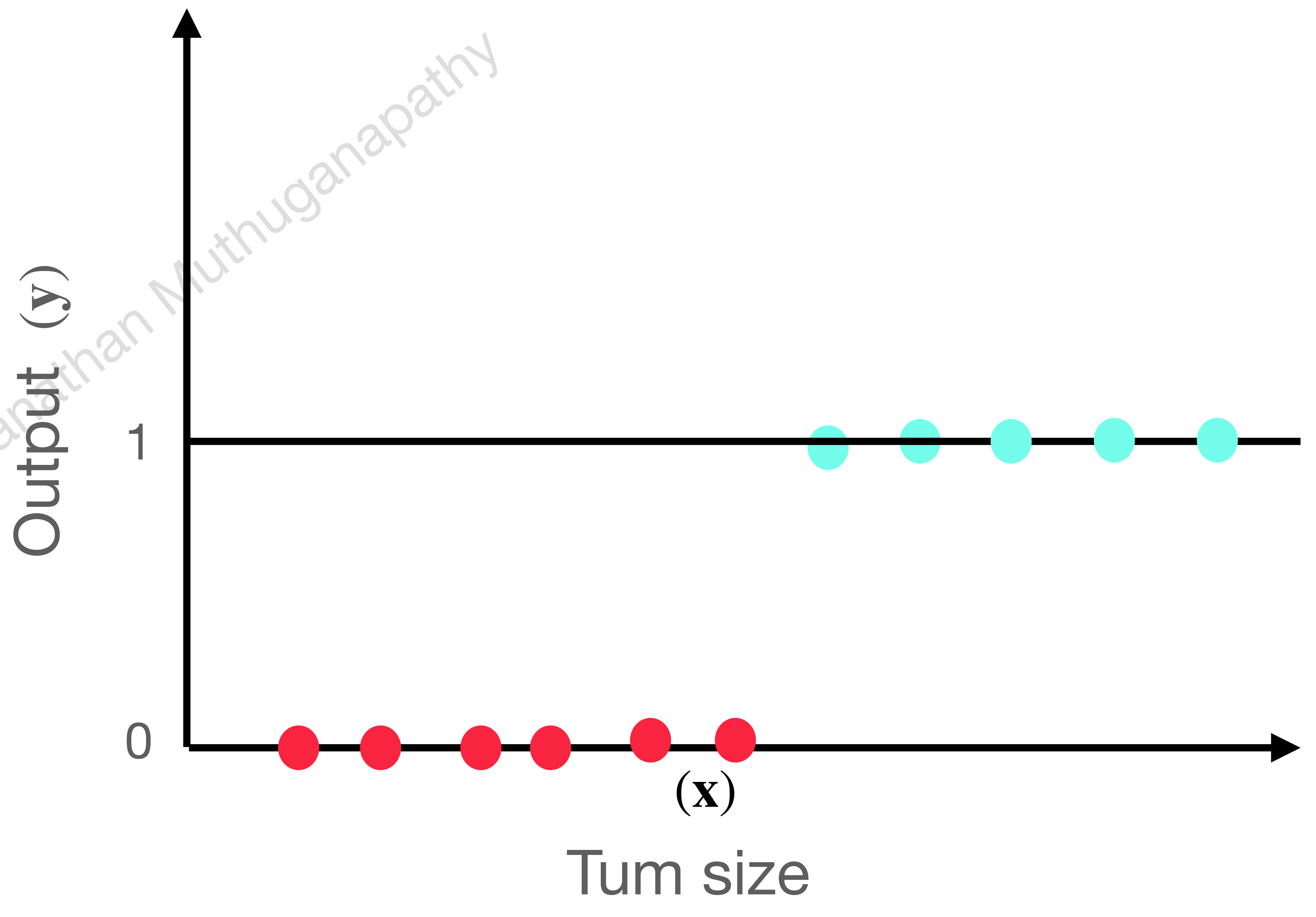
- Ground truth data - Input feature / output $(\mathbf{x}, \mathbf{y})$ are the knowns
- Use a model / hypothesis as $h(w)$ and cost function $J(w)$
-



Logistic Regression

Classification (binary)

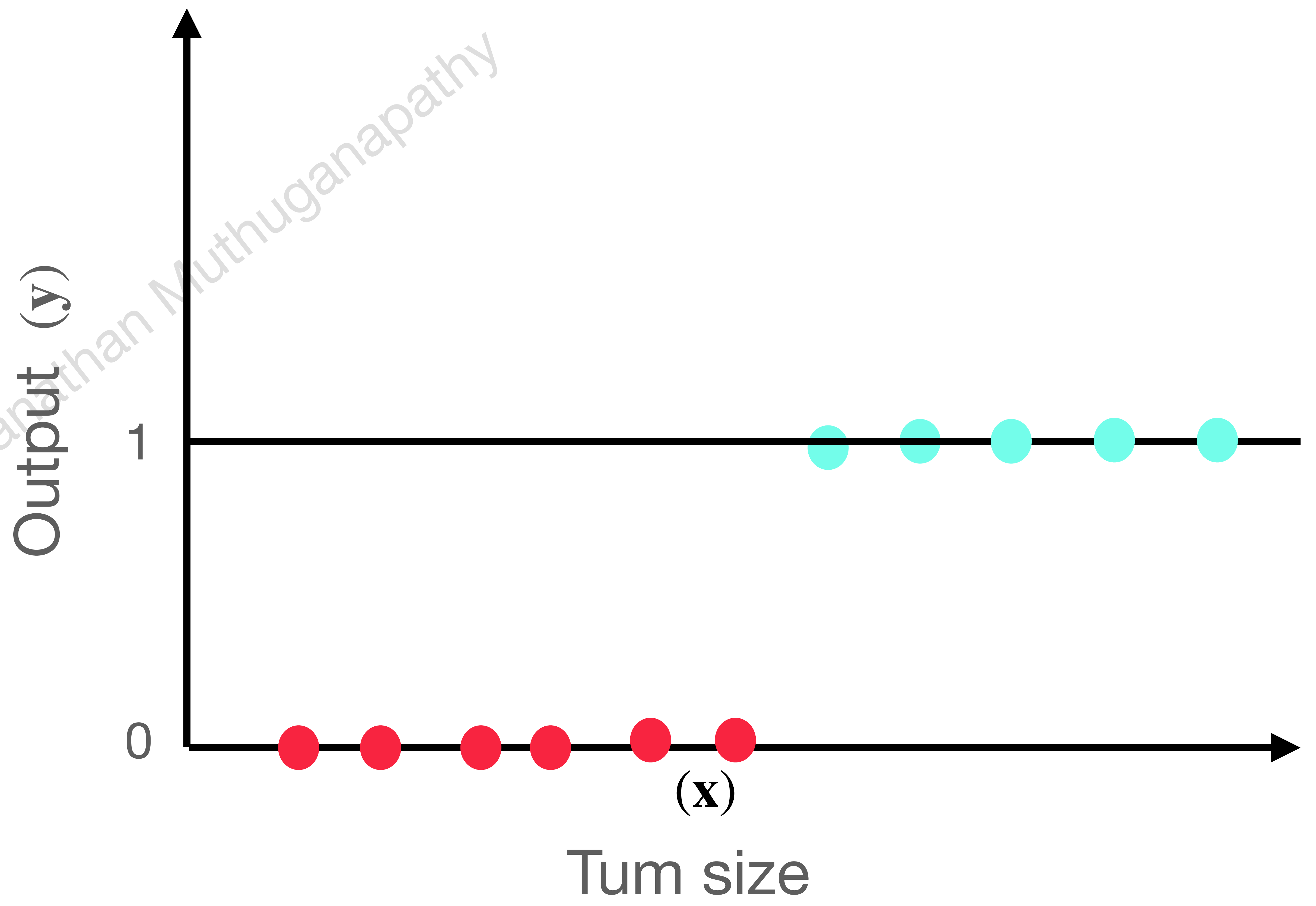
- Ground truth data - Input feature / output $(\mathbf{x}, \mathbf{y})$ are the knowns
- Output is either 0 or 1



Logistic Regression

Classification (binary) - Examples

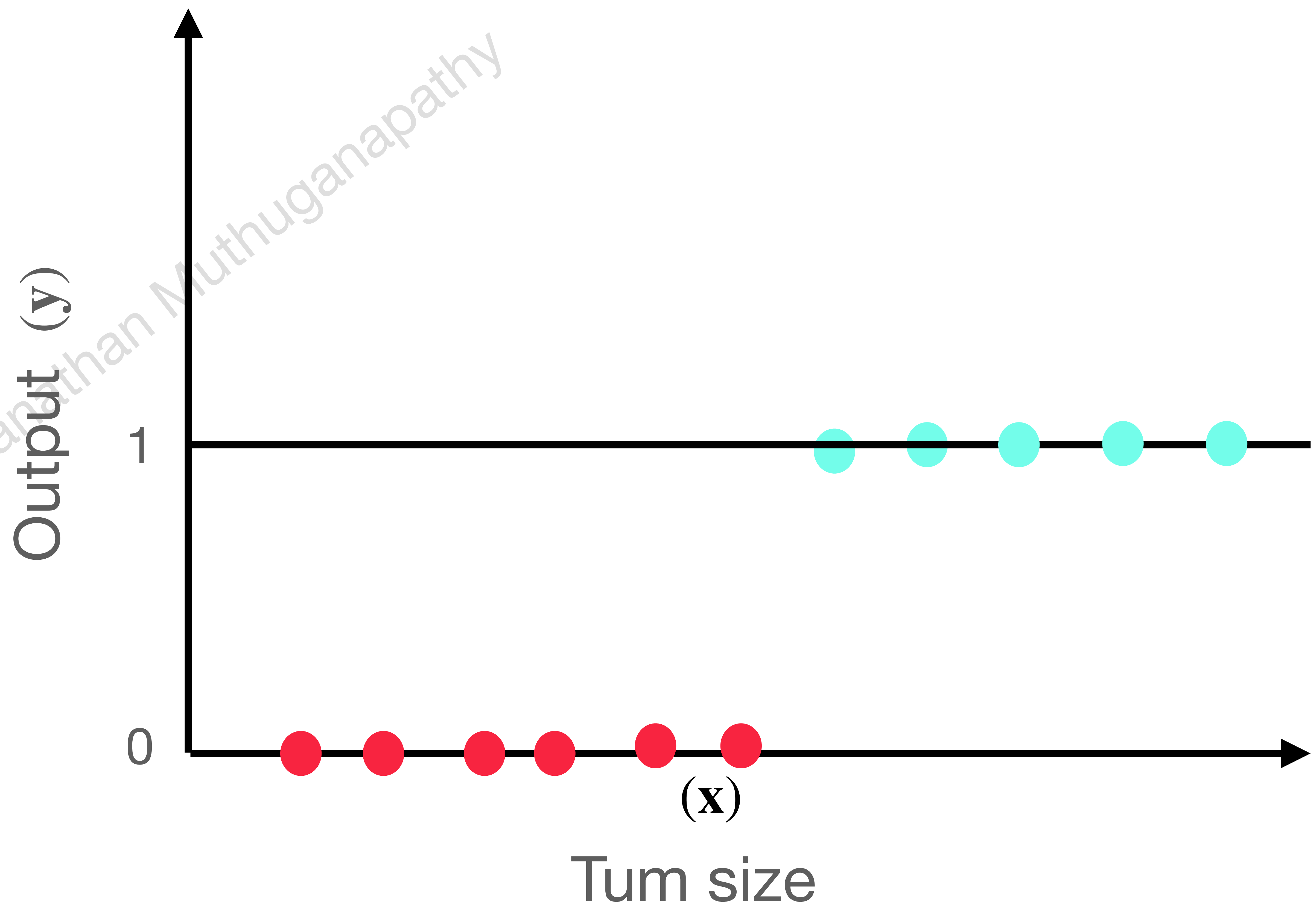
- Spam / Not spam
- Malignant / benign
- Fraud / No fraud
- Good / bad grades



Logistic Regression

Classification (binary)

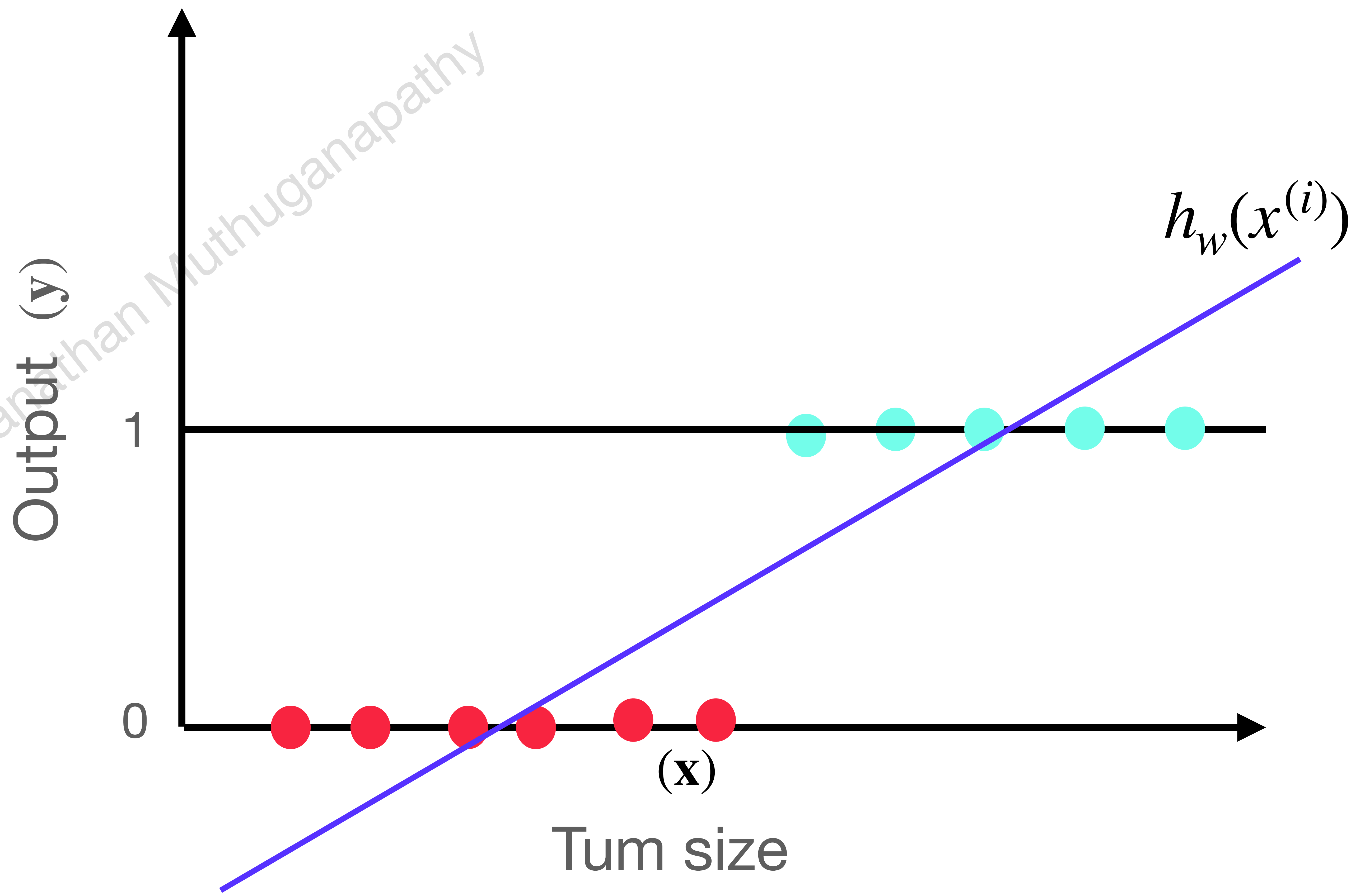
- Ground truth data - Input feature / output $(\mathbf{x}, \mathbf{y})$ are the knowns
- Output is either 0 or 1
- 🔴 - Benign
- 🔵 - Malignant



Logistic Regression

Hypothesis - Linear Regression Model

- Ground truth data - Input feature / output $(\mathbf{x}, \mathbf{y})$ are the knowns
- Output is either 0 or 1
- 🔴 - Small
- 🔵 - Large
- $\bar{y}^{(i)} = h_w(x^{(i)}) = w_0 + w_1 x^{(i)}$

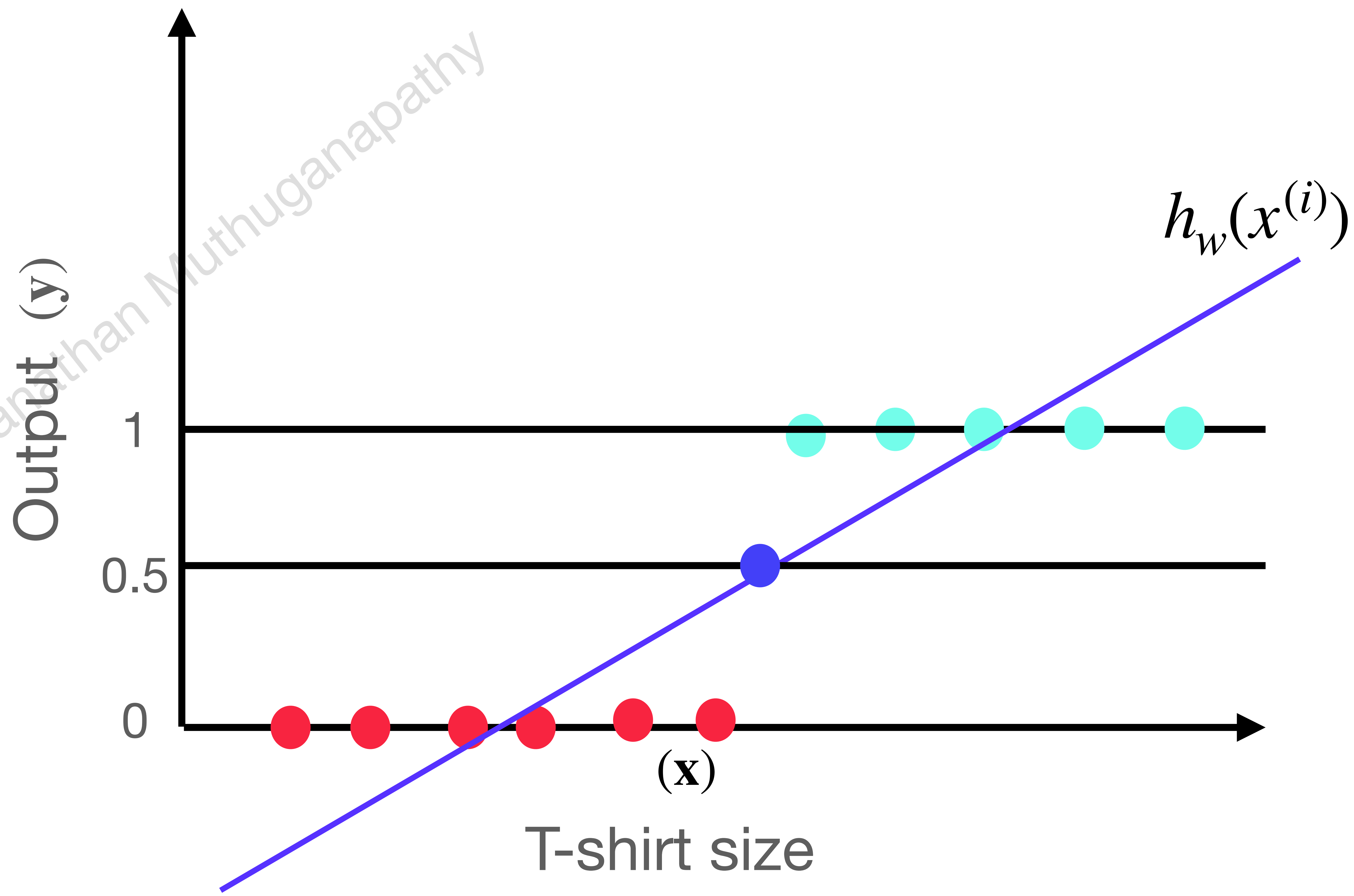


Ramanathan Muthuganapathy, Department of Engineering Design, IIT Madras

Logistic Regression

Hypothesis - Linear Regression Model with thresholding

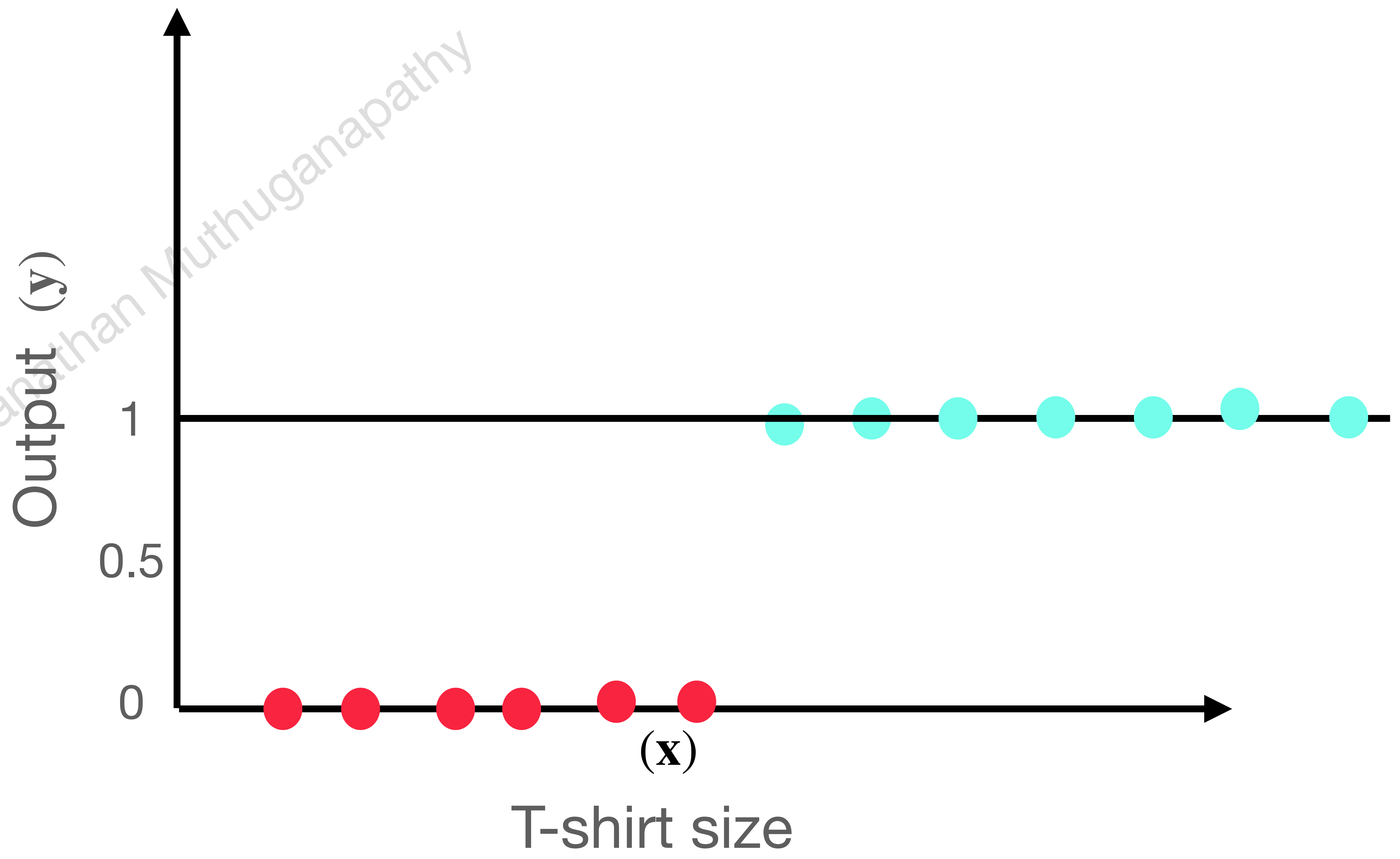
- $h_w(x^{(i)}) \geq 0.5, y = 1$
- $h_w(x^{(i)}) < 0.5, y = 0$



Logistic Regression

Hypothesis - Increase the training data

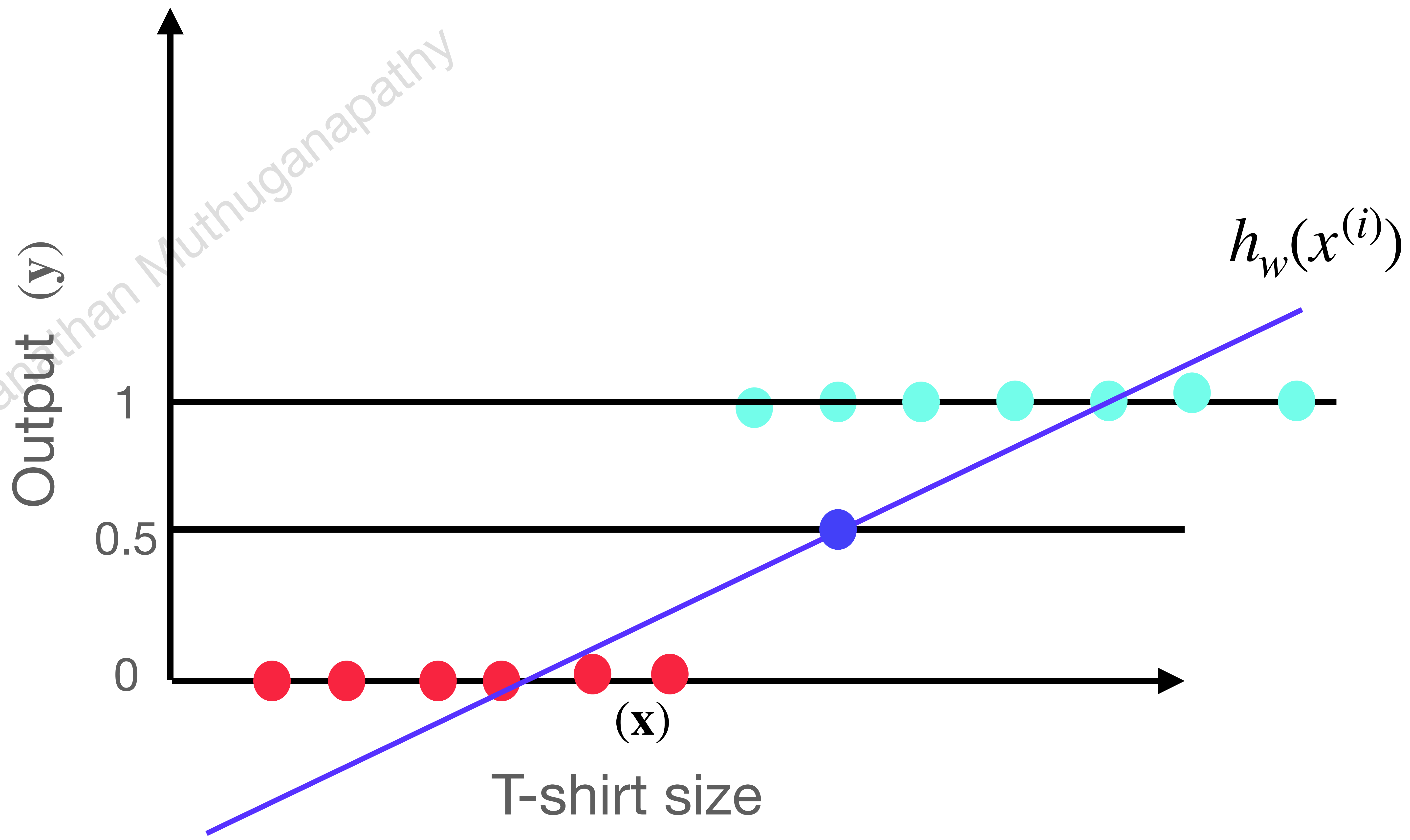
- $h_w(x^{(i)}) \geq 0.5, y = 1$
- $h_w(x^{(i)}) < 0.5, y = 0$



Logistic Regression

Hypothesis - Increase the training data

- $h_w(x^{(i)}) \geq 0.5, y = 1$
- $h_w(x^{(i)}) < 0.5, y = 0$
- Misclassification starts happening
- Not a good idea to use Linear Regression
- $y < 0$ or $y > 1$



Logistic Regression

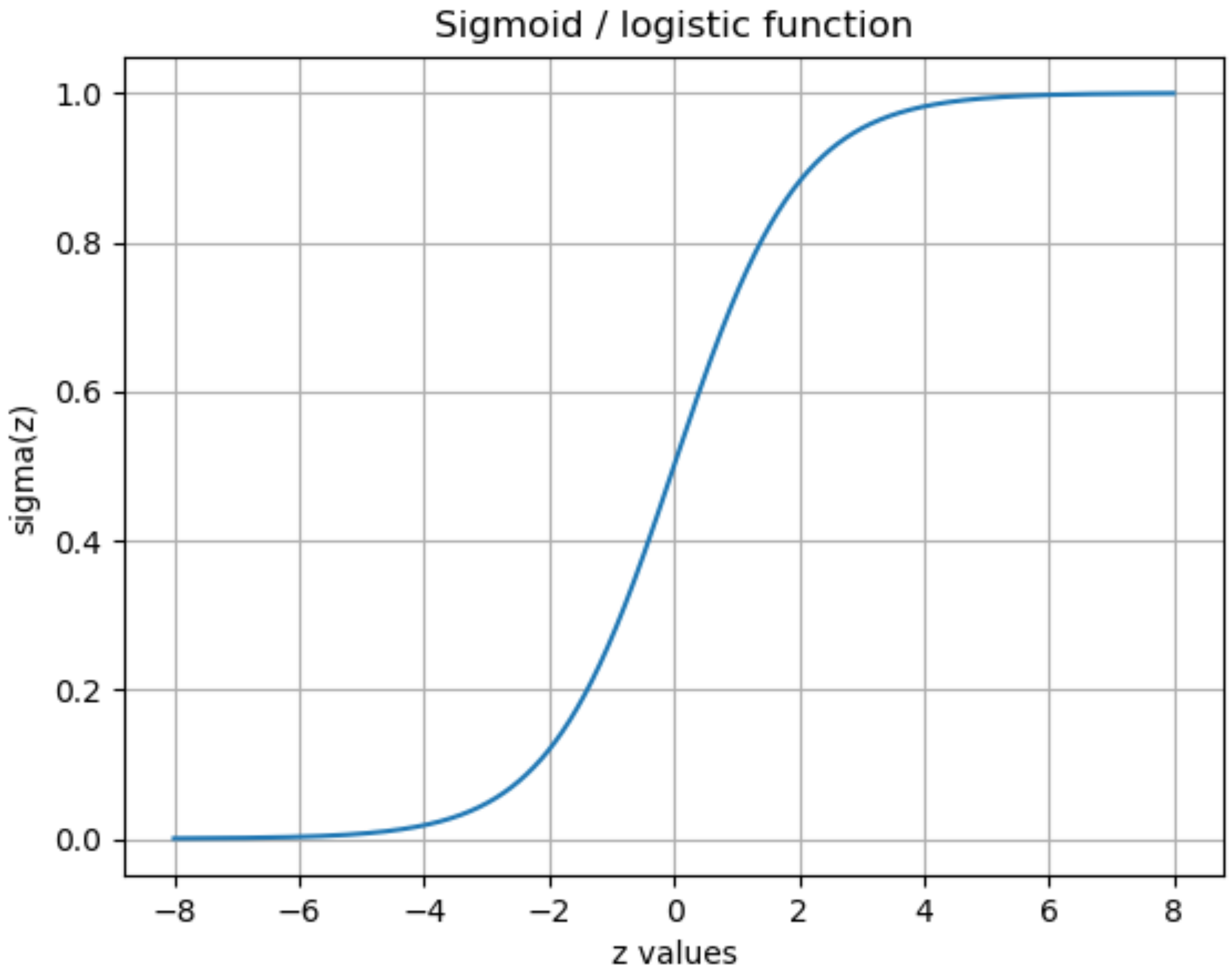
Sigmoid function

- $h_w(x) = \mathbf{w}^T\mathbf{x}$
- $h_w(x) = \sigma(\mathbf{w}^T\mathbf{x})$
- $\sigma(z) = \dfrac{1}{1 + e^{-z}}$
- $\sigma(z)$ is called Sigmoid or Logistic function.

Logistic Regression

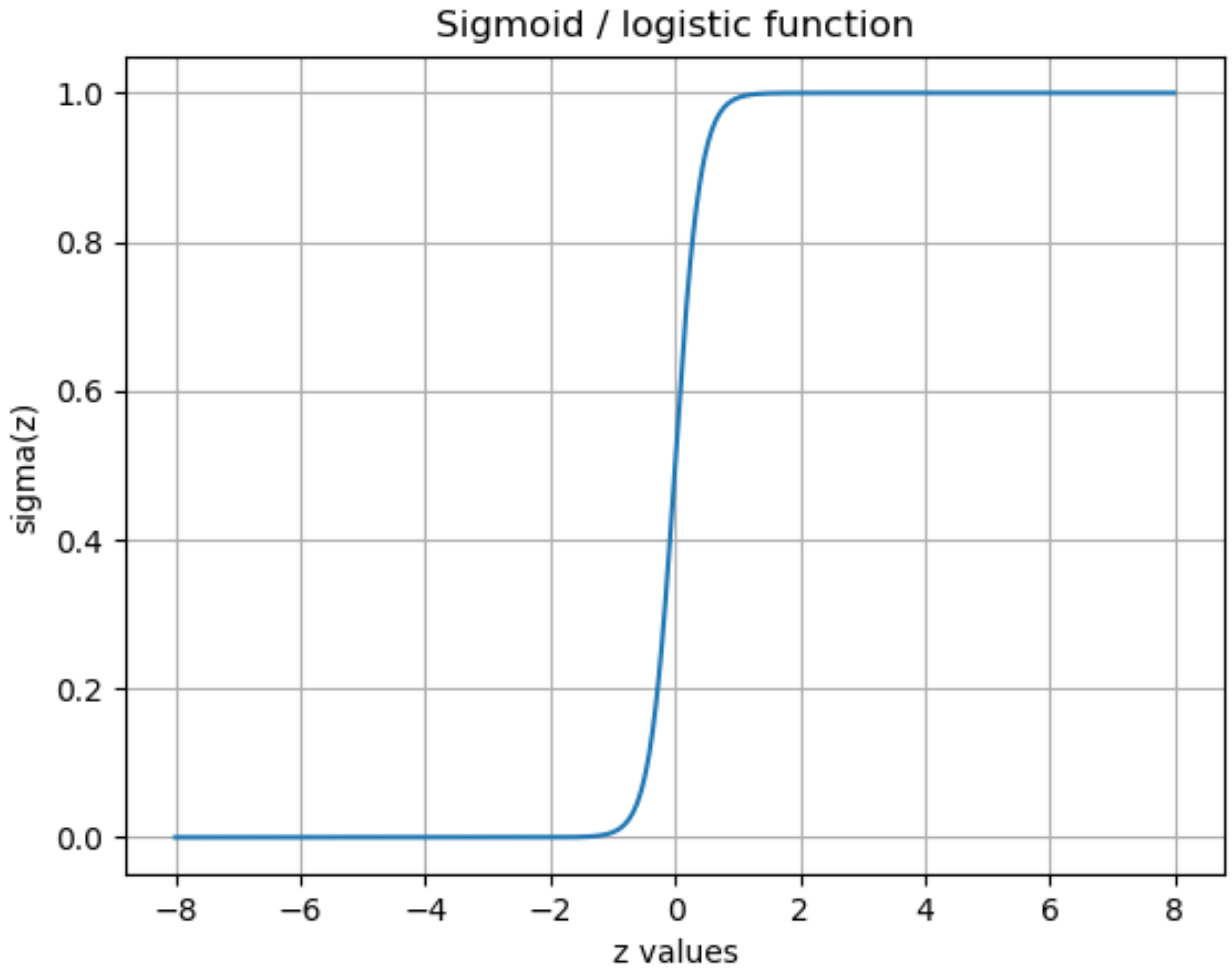
Sigmoid function

- $\sigma(z) = \dfrac{1}{1 + e^{-z}}$
- $\sigma(z)$ is called Sigmoid or Logistic function.



Logistic Regression
Sigmoid function

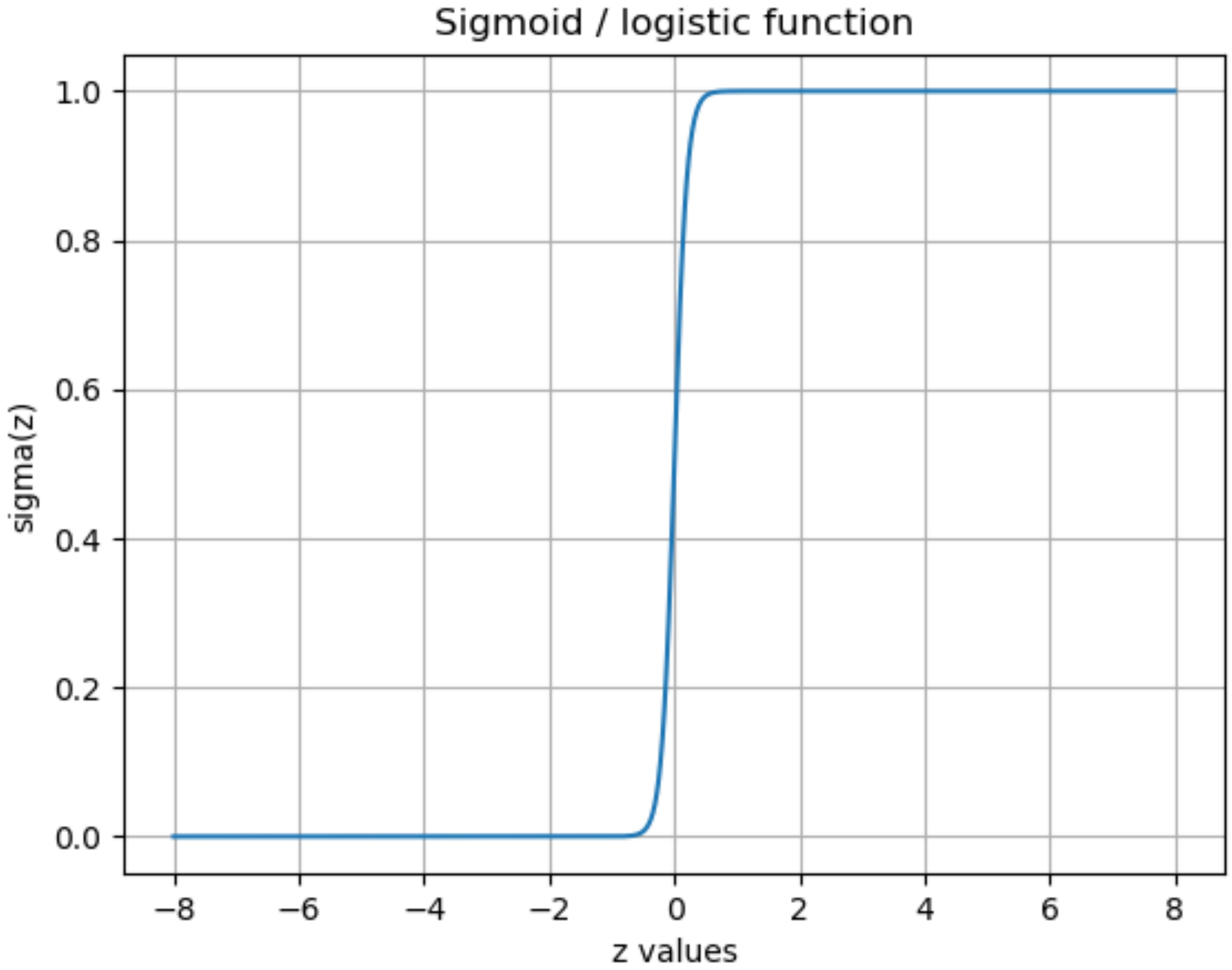
- $\sigma(z) = \dfrac{1}{1 + e^{-5z}}$
- $\sigma(z)$ with 5



Logistic Regression

Sigmoid function

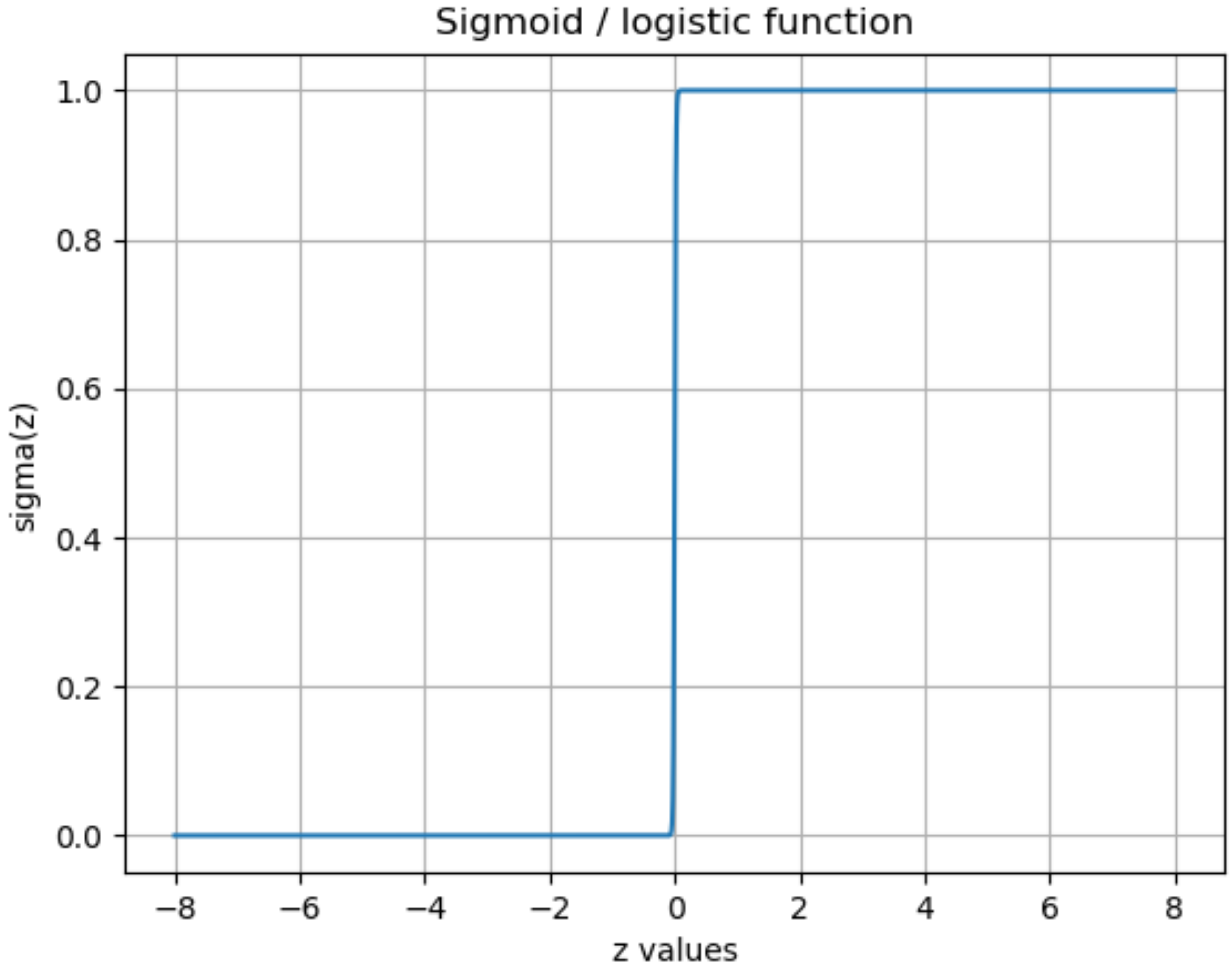
- $\sigma(z) = \dfrac{1}{1 + e^{-10z}}$
- $\sigma(z)$ with 10.



Logistic Regression

Sigmoid function

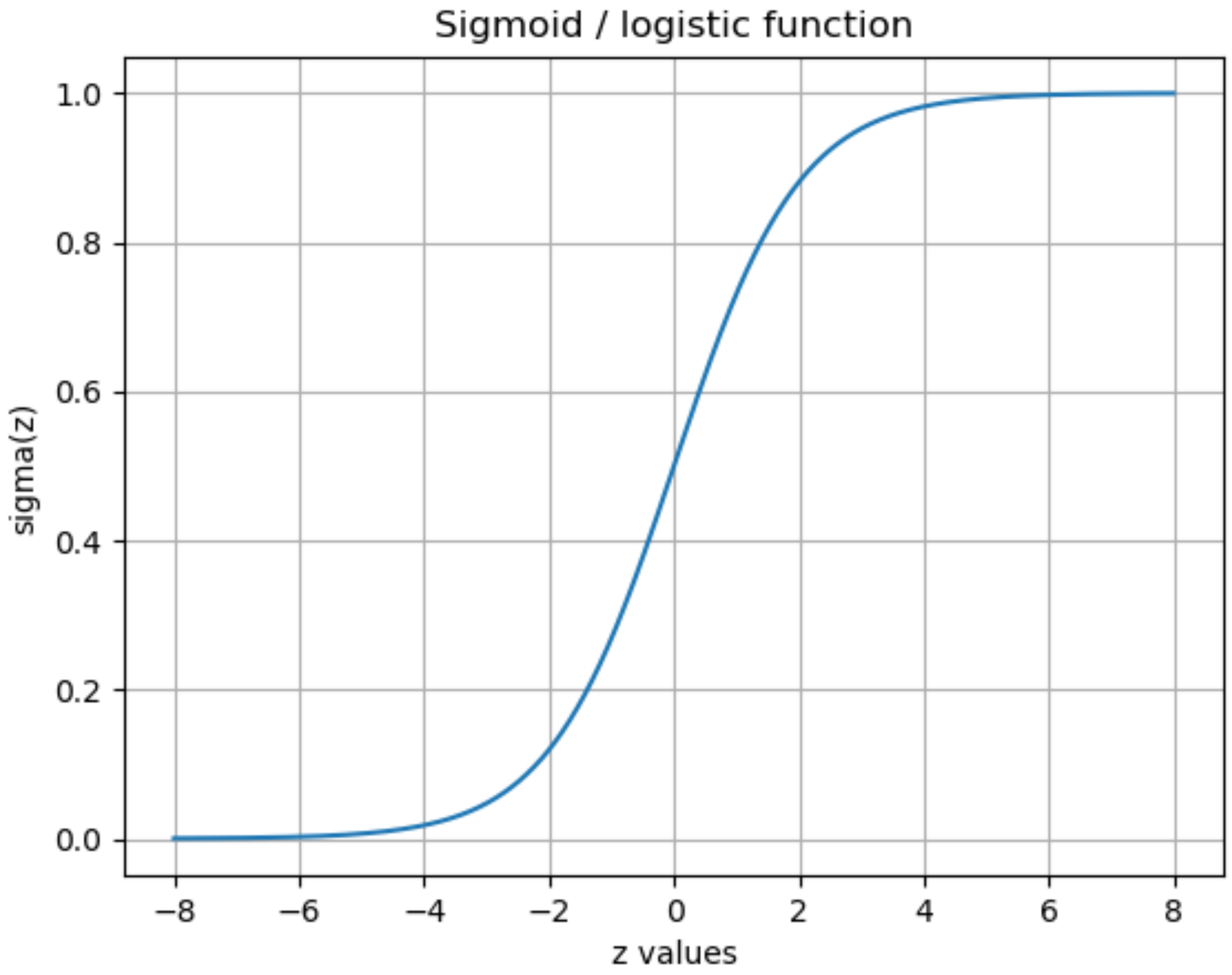
- $\sigma(z) = \dfrac{1}{1 + e^{-100z}}$
- $\sigma(z)$ with 100



Logistic Regression

Sigmoid function

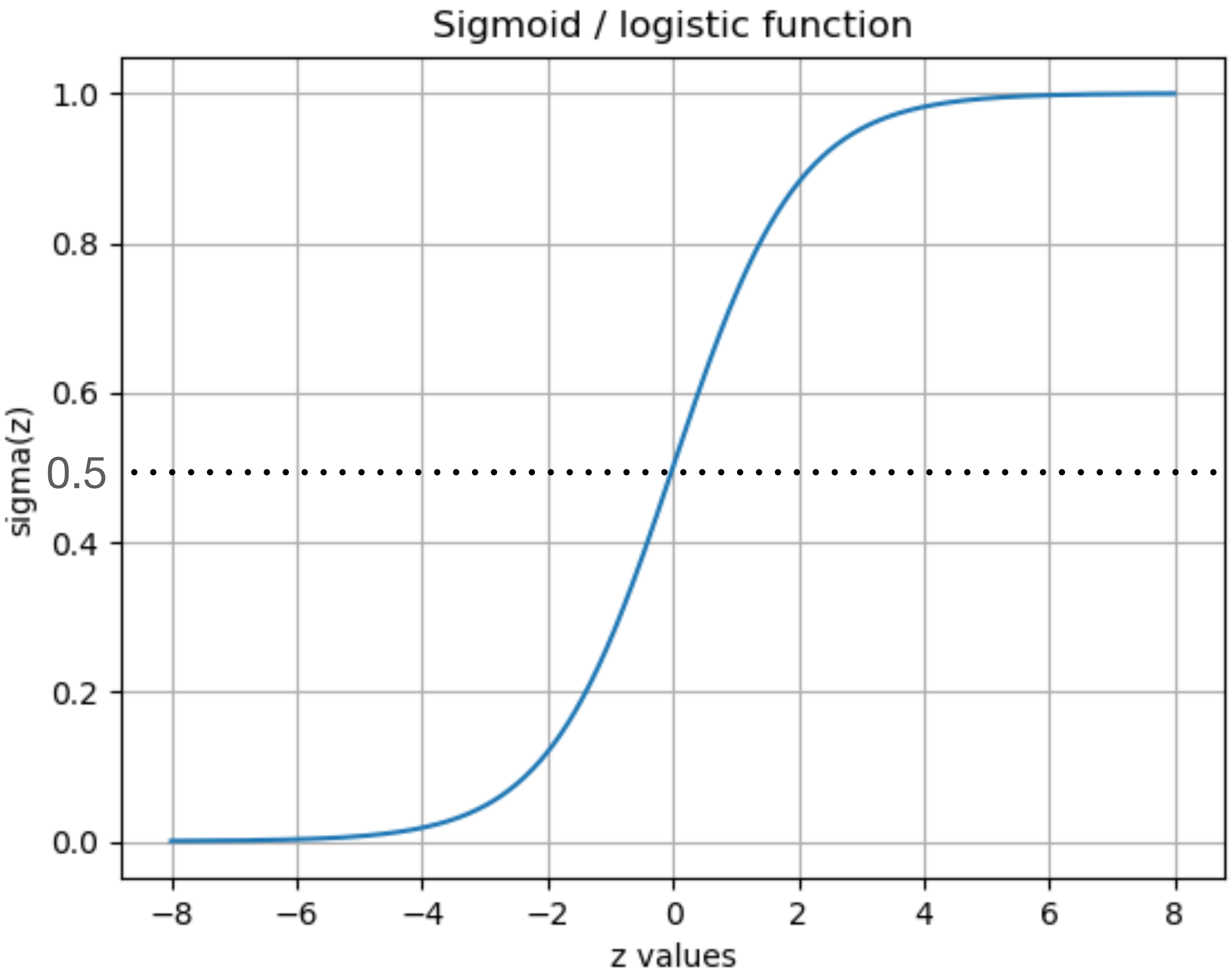
- $\sigma(z) = \dfrac{1}{1 + e^{-z}}$
- Smoother approximation of step function
- This means what?



Logistic Regression

Sigmoid - Observations

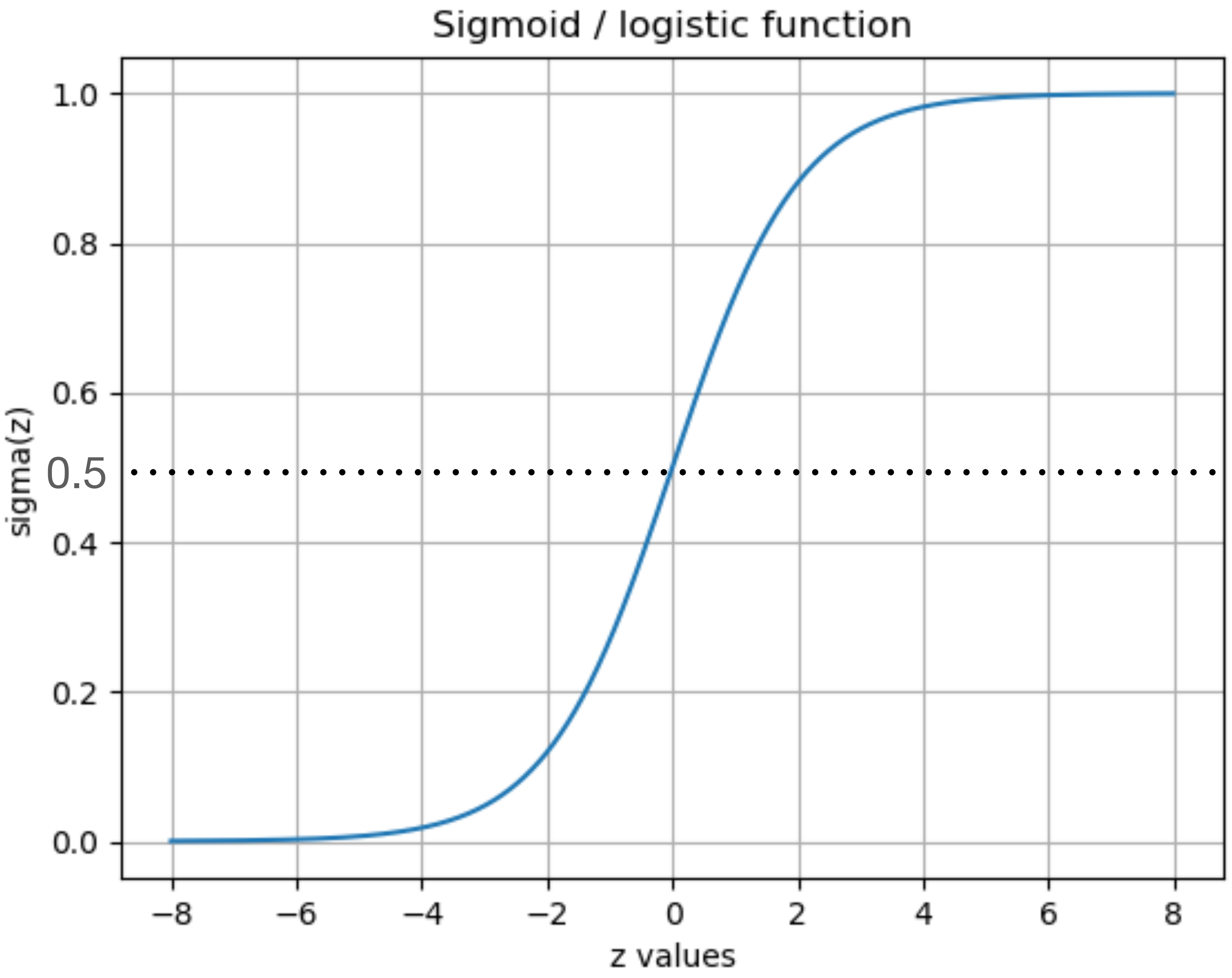
- $\sigma(z) = \dfrac{1}{1 + e^{-z}}$
- $0 \leq \sigma(z) < = 1$



Logistic Regression

Sigmoid - Observations

- $\sigma(z) = \dfrac{1}{1 + e^{-z}}$
- value of $\sigma(z)$ at $z = 0$?

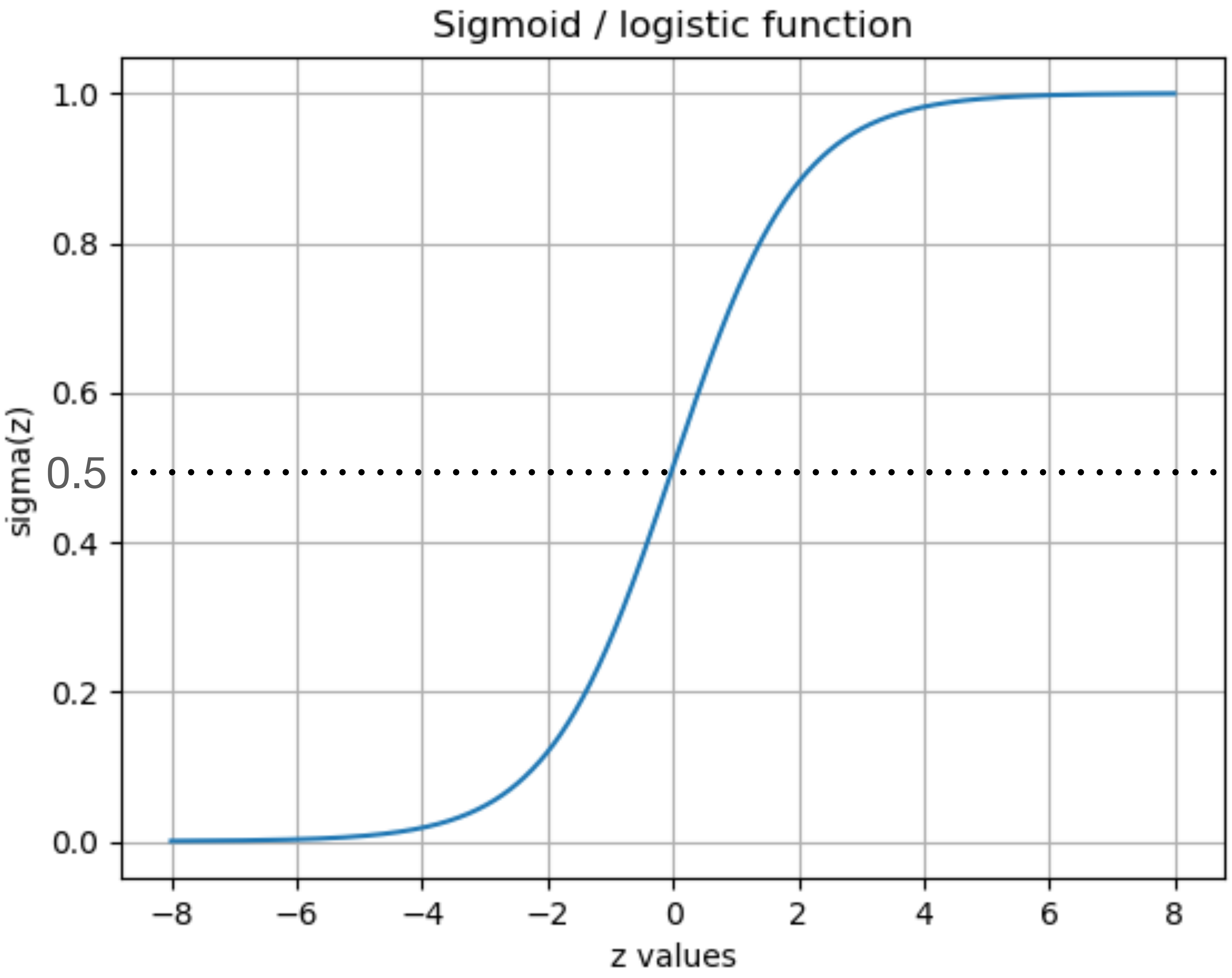


Ramanathan Muthuganapathy, Department of Engineering Design, IIT Madras

Logistic Regression

Sigmoid - Observations

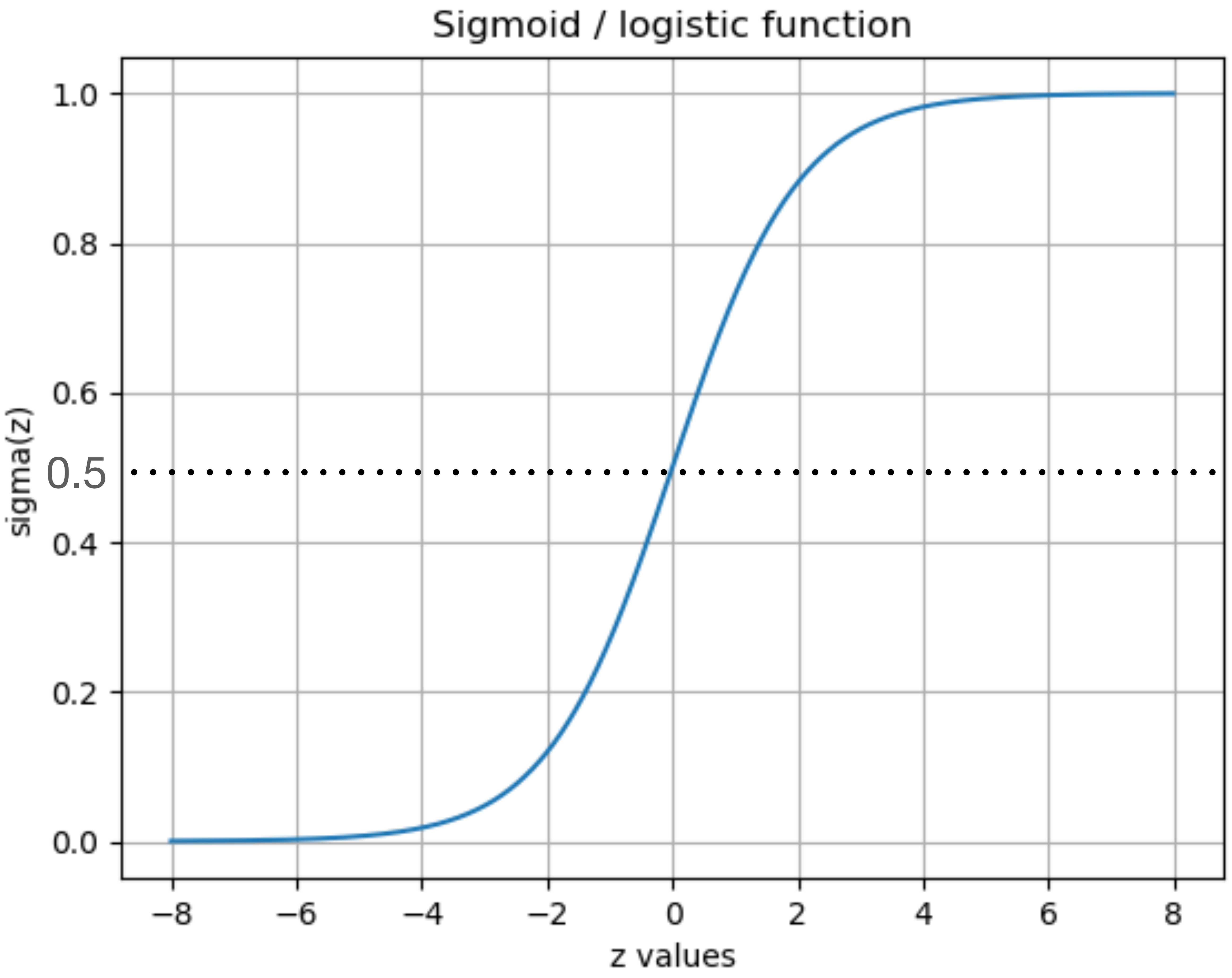
- $\sigma(z) = \dfrac{1}{1 + e^{-z}}$
- $z \geq 0, \sigma(z) \geq 0.5$
- $z < 0, \sigma(z) < 0.5$



Logistic Regression

Sigmoid - Observations

- $\sigma(z) = \dfrac{1}{1 + e^{-z}}$
- $z \geq 0, \sigma(z) \geq 0.5$
- $z < 0, \sigma(z) < 0.5$
- $\sigma(z)$ sign changes at 0.5

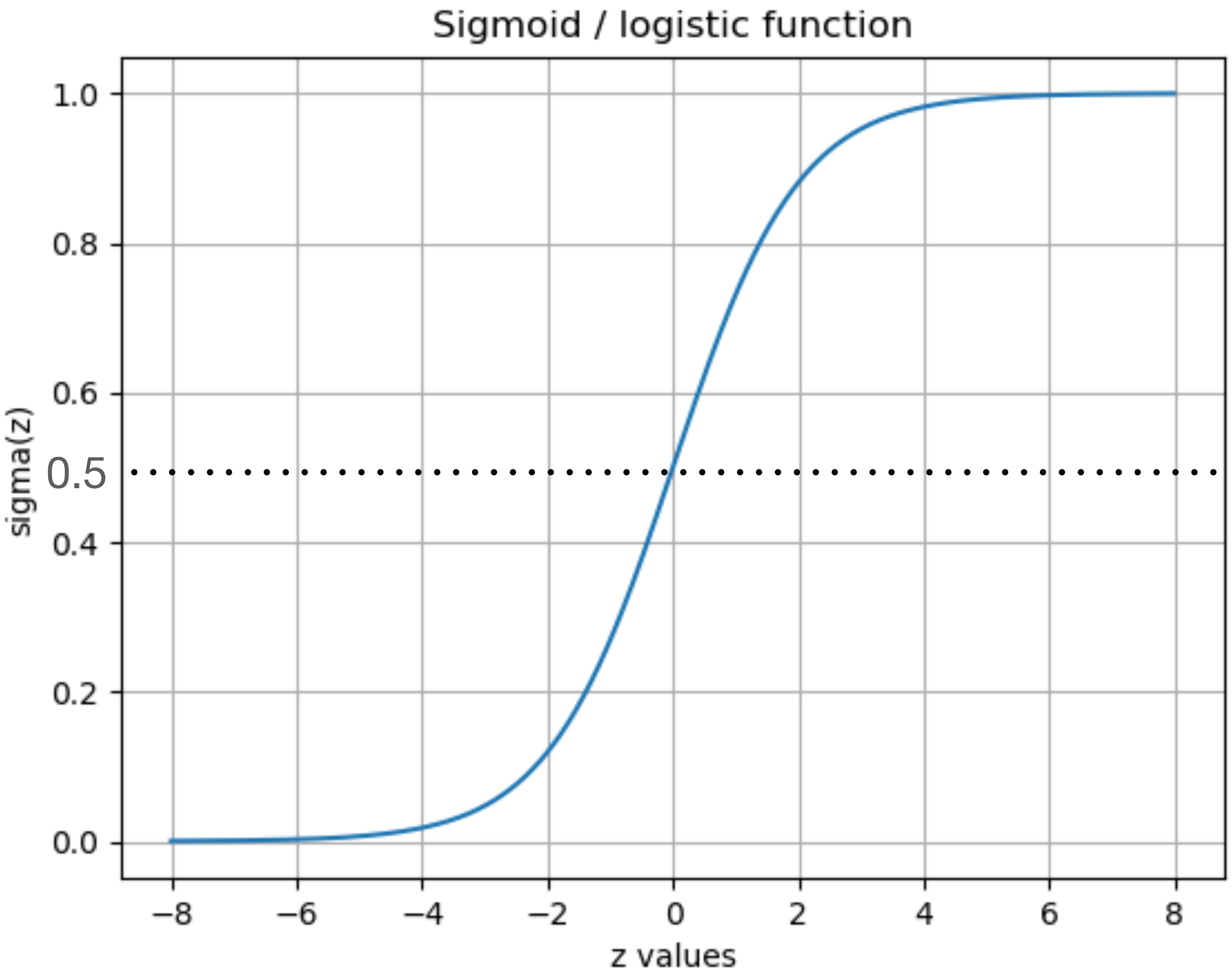


Ramanathan Muthuganapathy, Department of Engineering Design, IIT Madras

Logistic Regression

Sigmoid - Observations

- $h_w(x) = \sigma(\mathbf{w}^T\mathbf{x})$
- $h_w(x) = \dfrac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$

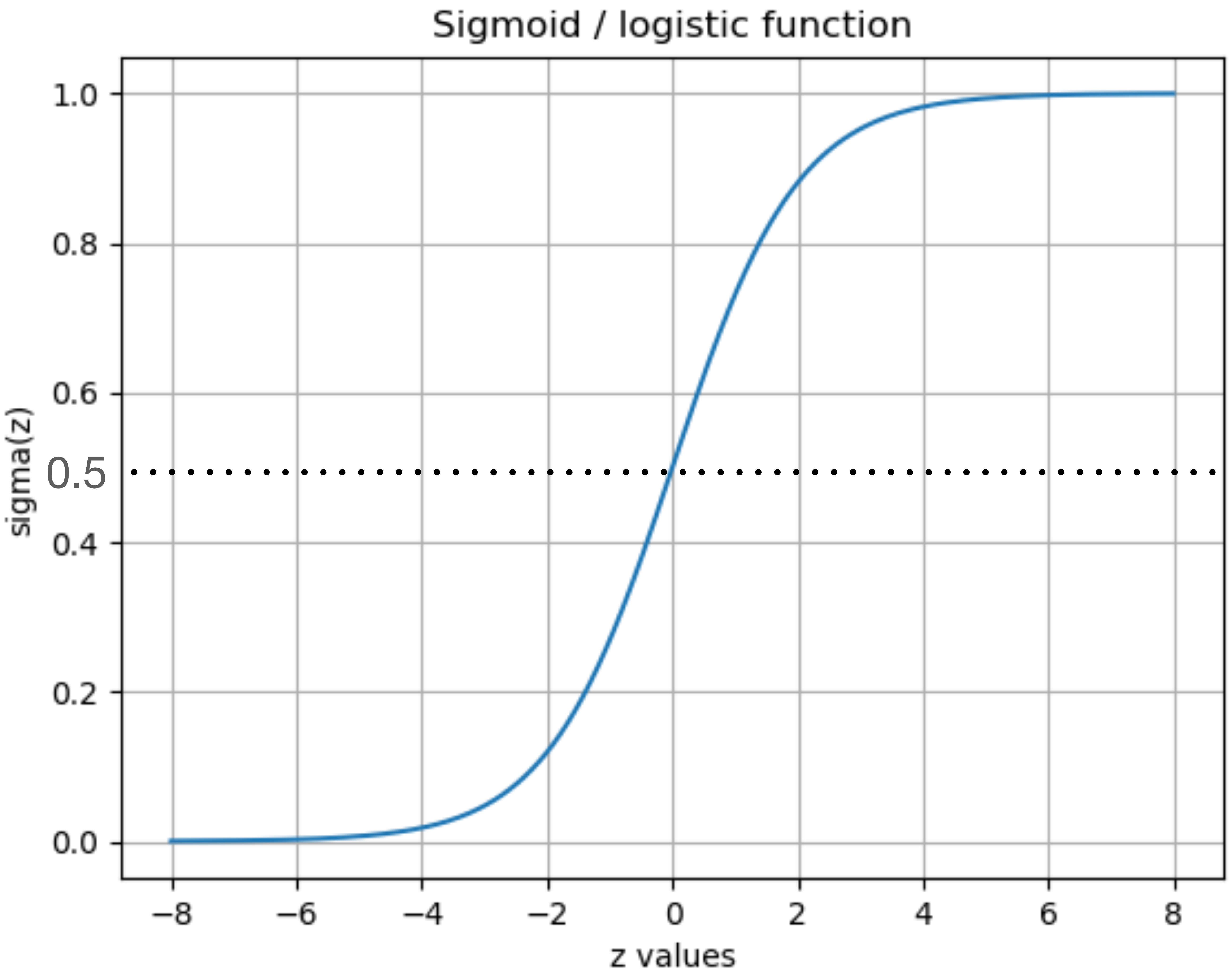


Ramanathan Muthuganapathy, Department of Engineering Design, IIT Madras

Logistic Regression

Sigmoid - Observations

- $h_w(x) = \sigma(\mathbf{w}^T\mathbf{x})$
- $h_w(x) = \dfrac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$
- $\mathbf{w}^T\mathbf{x} \geq 0, \sigma(\mathbf{w}^T\mathbf{x}) \geq 0.5$
- $\mathbf{w}^T\mathbf{x} < 0, \sigma(\mathbf{w}^T\mathbf{x}) < 0.5$



Logistic Regression

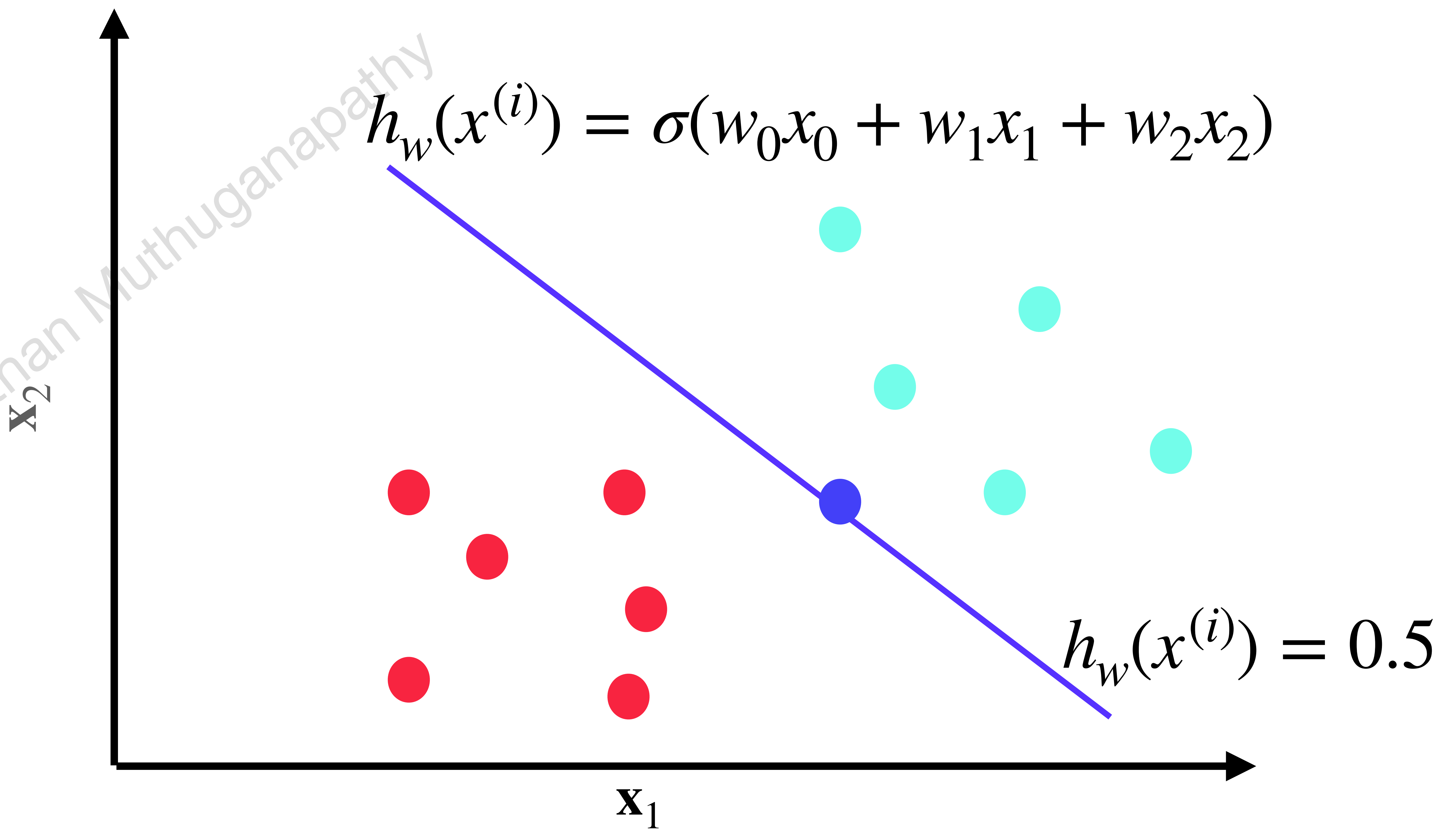
Sigmoid - Interpretation

- $h_w(x)$ - Estimated probability that y = 1 at x
- $h_w(x) = 0.85$, probably that the size is large is 85% and hence y = 1
- $y = 1$ if $h_w(x) \geq 0.5$
- $y = 0$ if $h_w(x) < 0.5$
- $\mathbf{w}^T\mathbf{x} \geq 0, \sigma(\mathbf{w}^T\mathbf{x}) \geq 0.5$
- $\mathbf{w}^T\mathbf{x} < 0, \sigma(\mathbf{w}^T\mathbf{x}) < 0.5$

Logistic Regression

Decision boundary

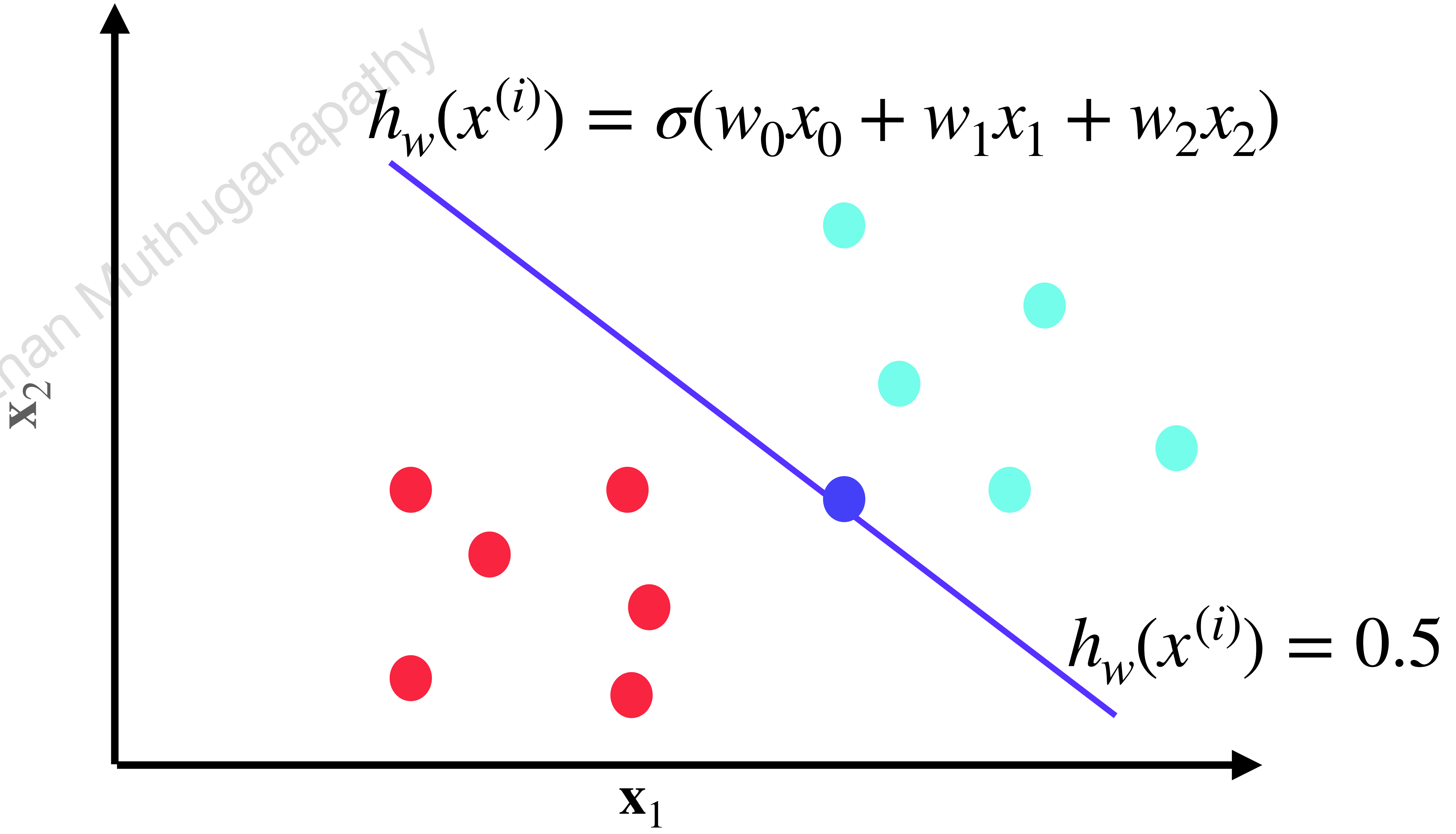
- $h_w(x^{(i)}) \geq 0.5, y = 1$
- $h_w(x^{(i)}) < 0.5, y = 0$
- $w_0 = -5, w_1 = 1, w_2 = 1$
- Apply $\mathbf{w}^T\mathbf{x} \geq 0$
- Linear decision boundary.
- You can also get non-linear decision boundary.



Logistic Regression

Cost function

- We need $h_w(x^{(i)})$
- We need to find the weights $w_i's$
- Cost function.



Logistic Regression

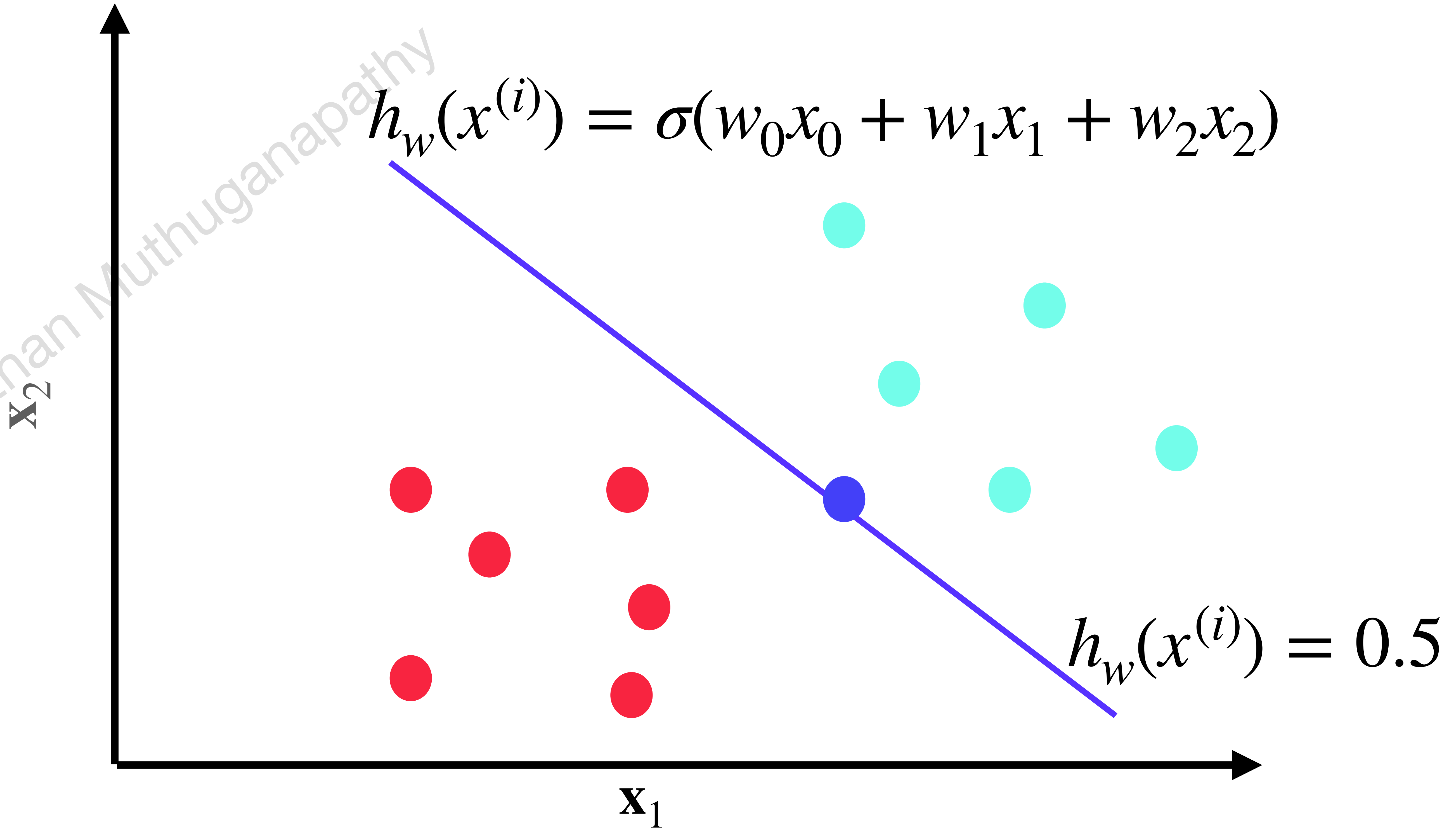
Cost function - Squared cost function

- Let us look at squared distance cost function.

- Assume we have $h_w(x)$

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

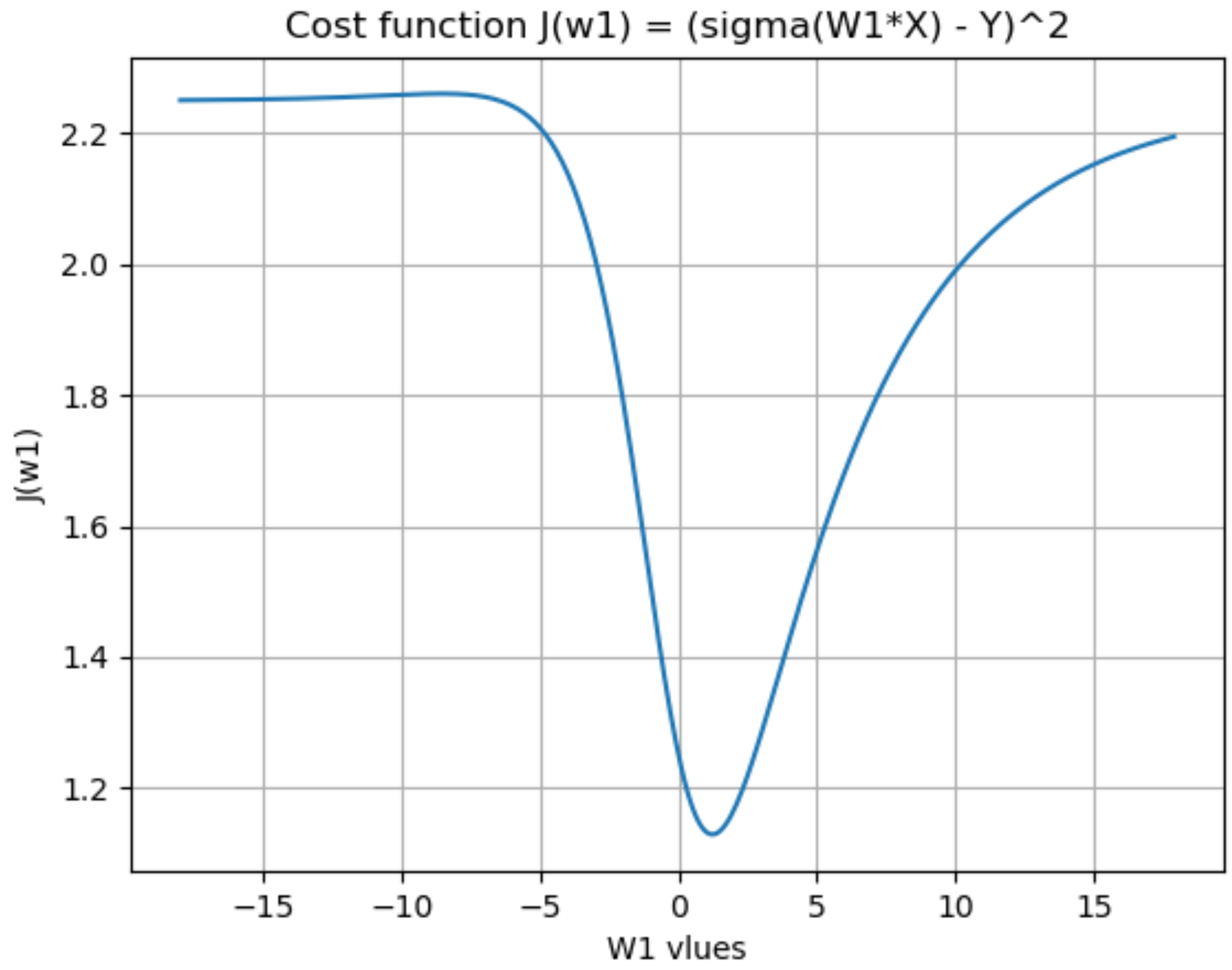
- $$h_w(x) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$$



Logistic Regression

Cost function - Squared cost function

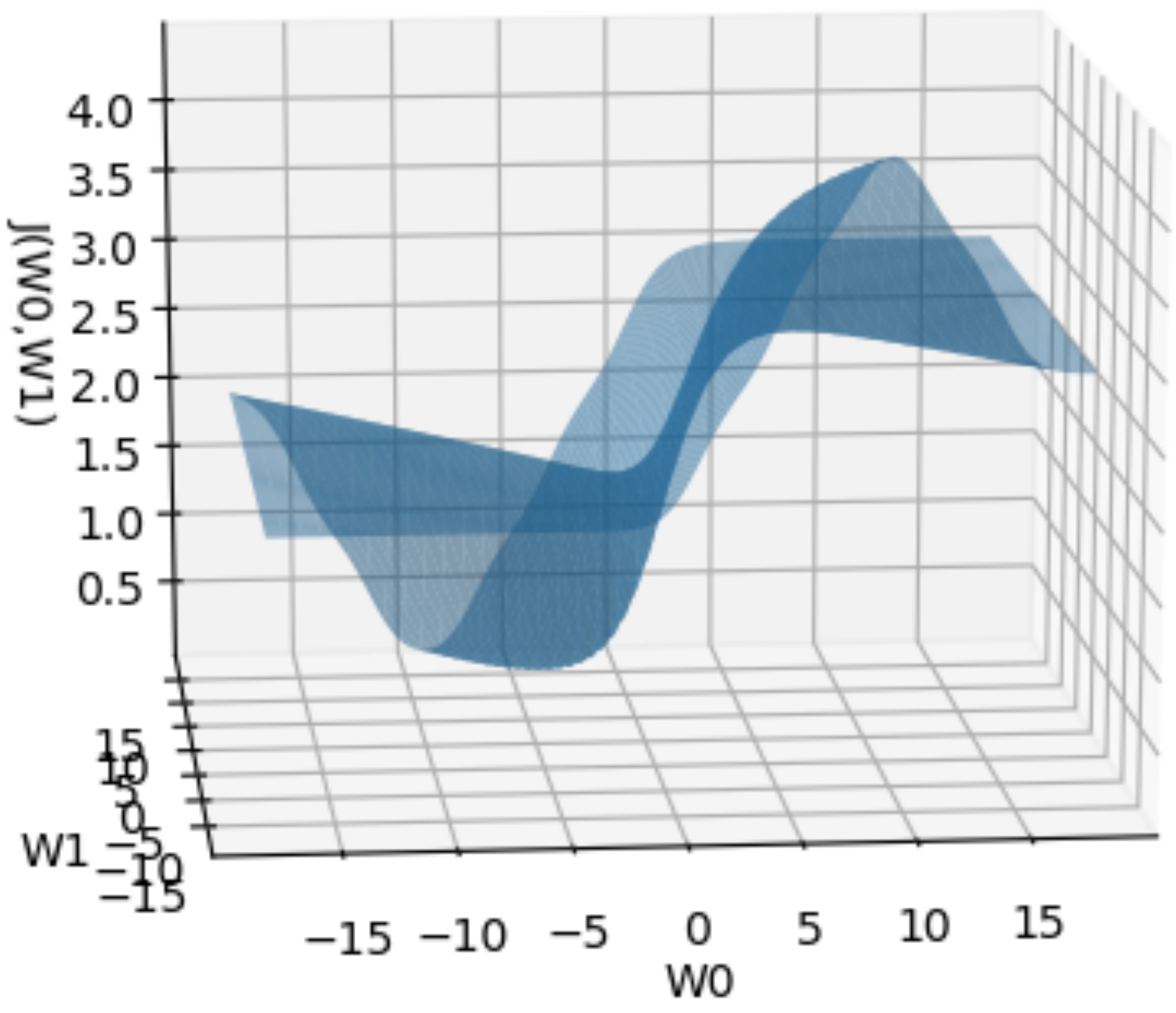
- In one variable.
- Not very desirable

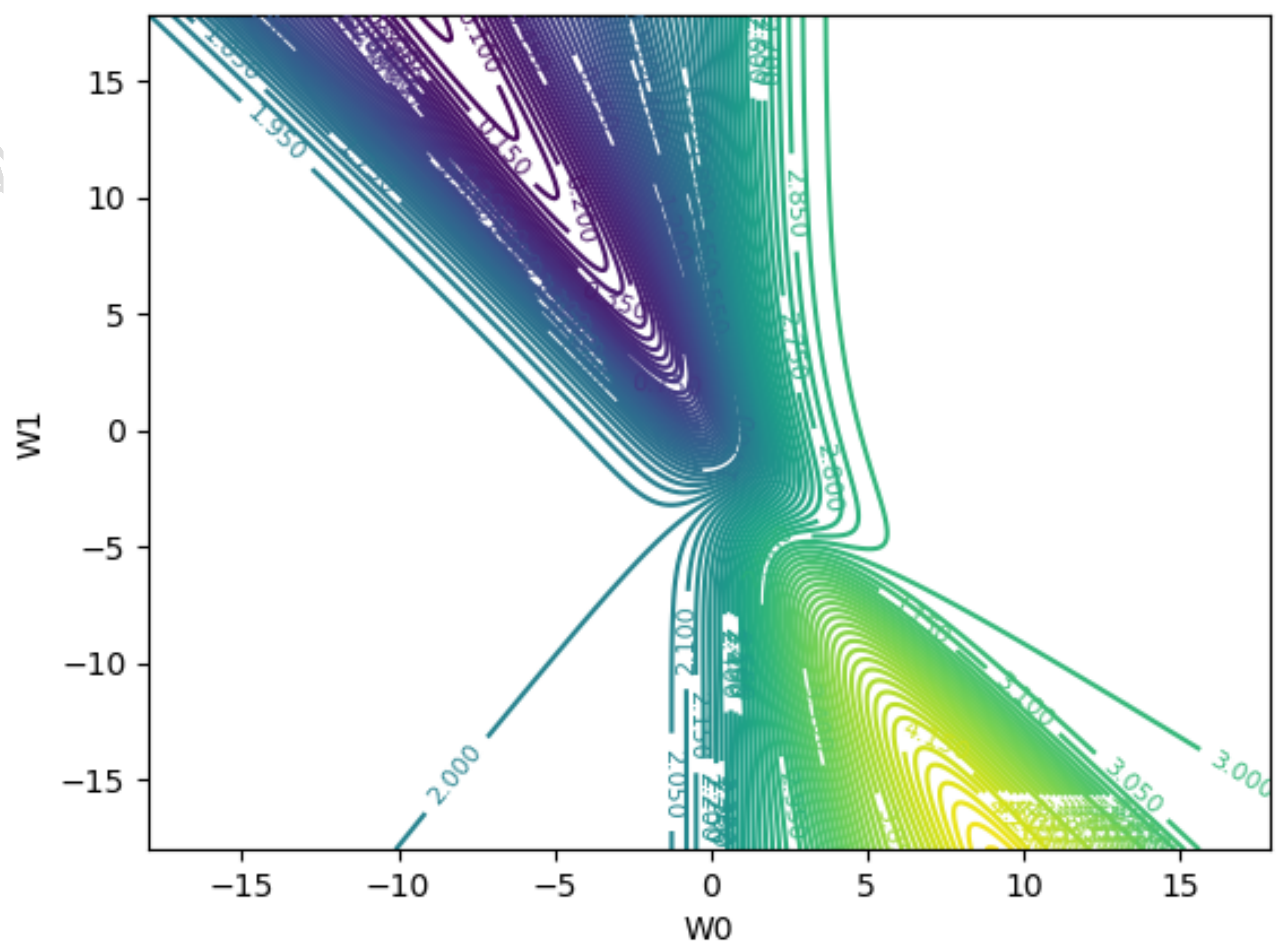


Ramanathan Muthuganapathy, Department of Engineering Design, IIT Madras

Logistic Regression
Cost function - Squared cost function (two variables)

- Non-convex, CP looks pretty bad!
- Not very desirable

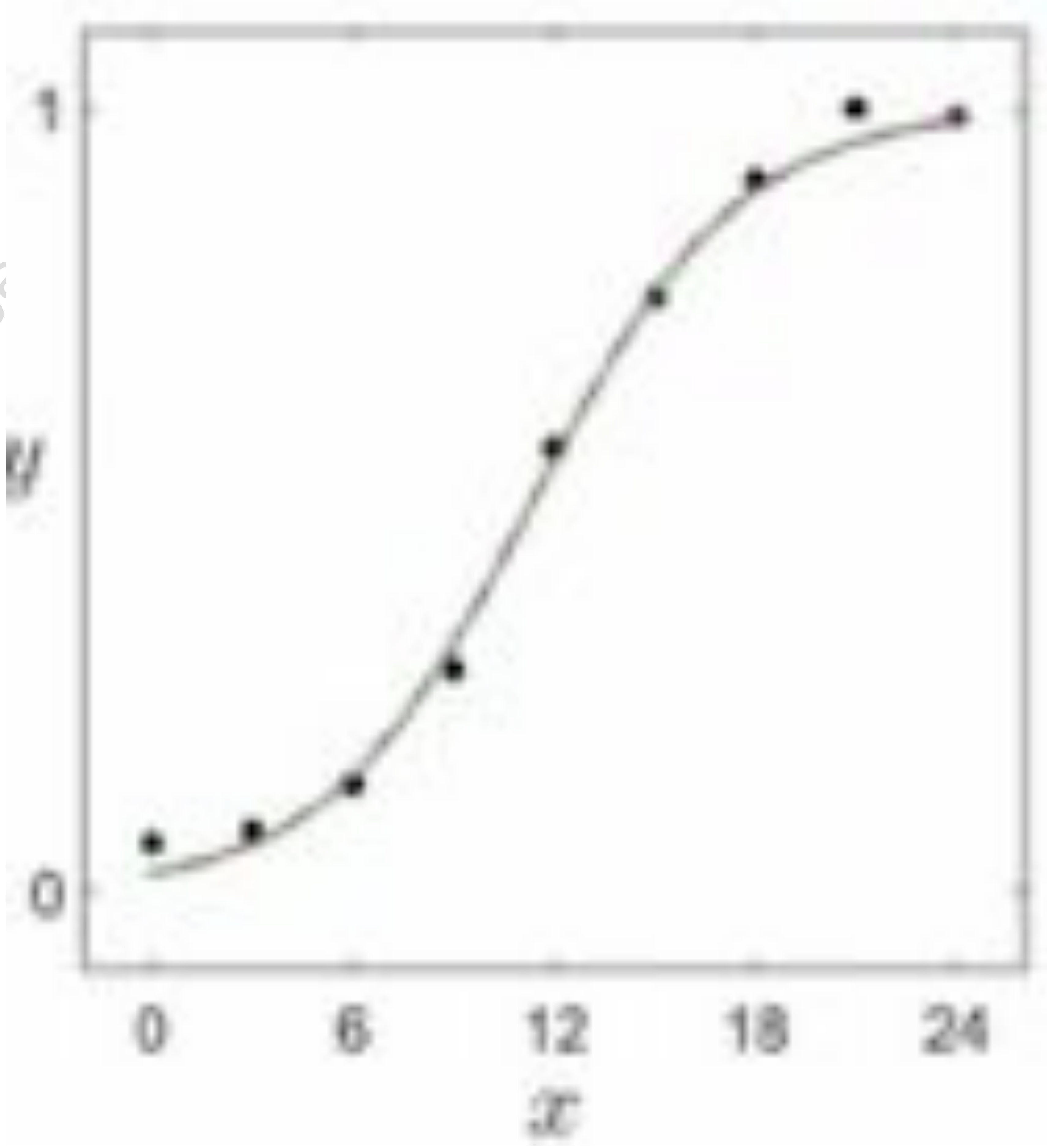




Ramanathan Muthuganapathy, Department of Engineering Design, IIT Madras

Note on Logistic Regression
for prediction (MLR book)

- To model population growth
- To get to a saturation level
- Squared distance cost function
- Now it is synonymous with classification



Logistic Regression

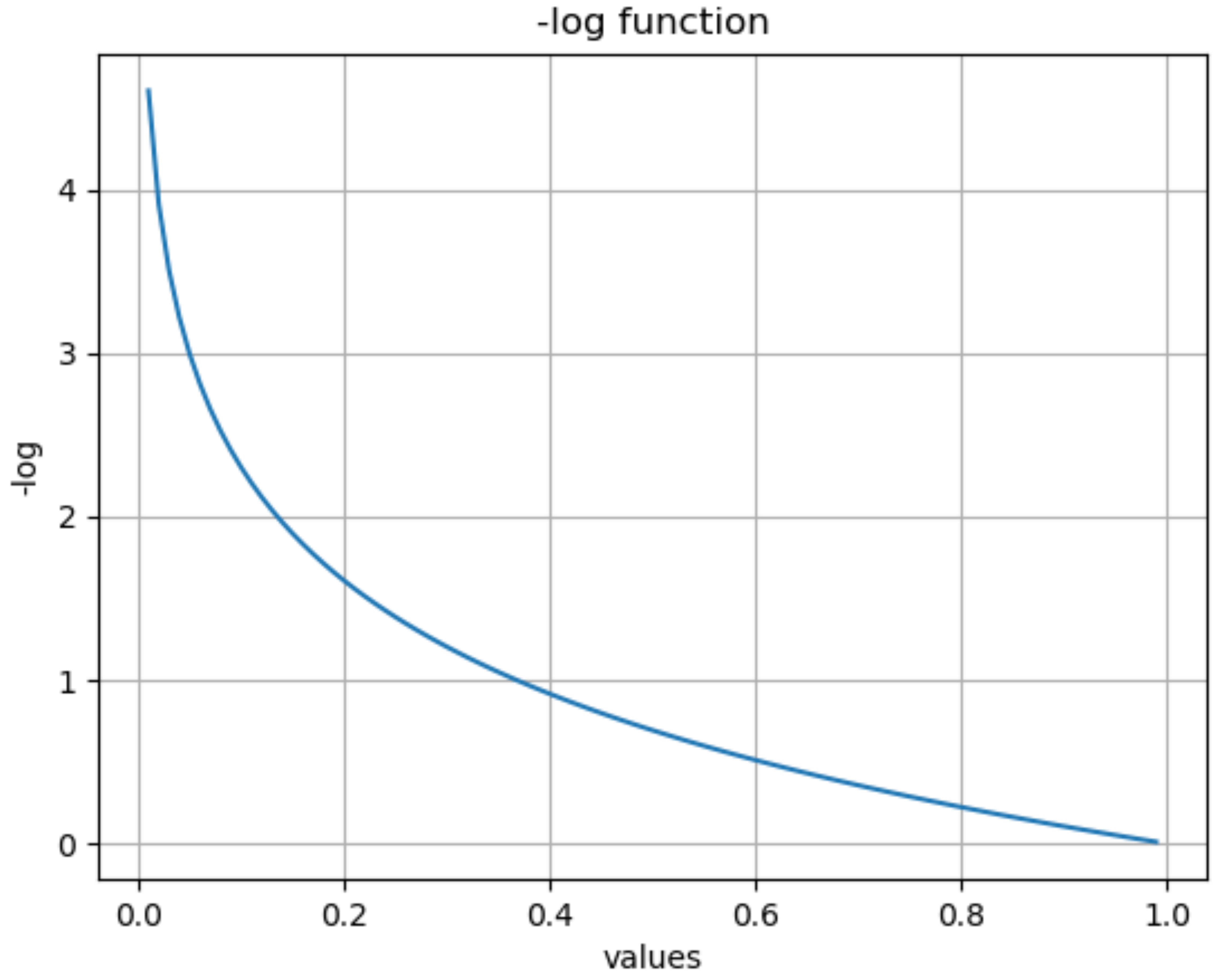
Cross-Entropy cost function

- cost $(h_w(x), y) = \begin{cases} -\log(h_w(x)) & if\ y = 1 \\ -\log(1 - h_w(x)) & if\ y = 0 \end{cases}$

Logistic Regression

Cross-Entropy cost function

- cost $(h_w(x), y) =$
$$\begin{cases} -\log(h_w(x)) & if\ y = 1 \\ -\log(1 - h_w(x)) & if\ y = 0 \end{cases}$$
- $h_w(x) = 1$, cost is 0
- $h_w(x) = 0$, penalization with large cost



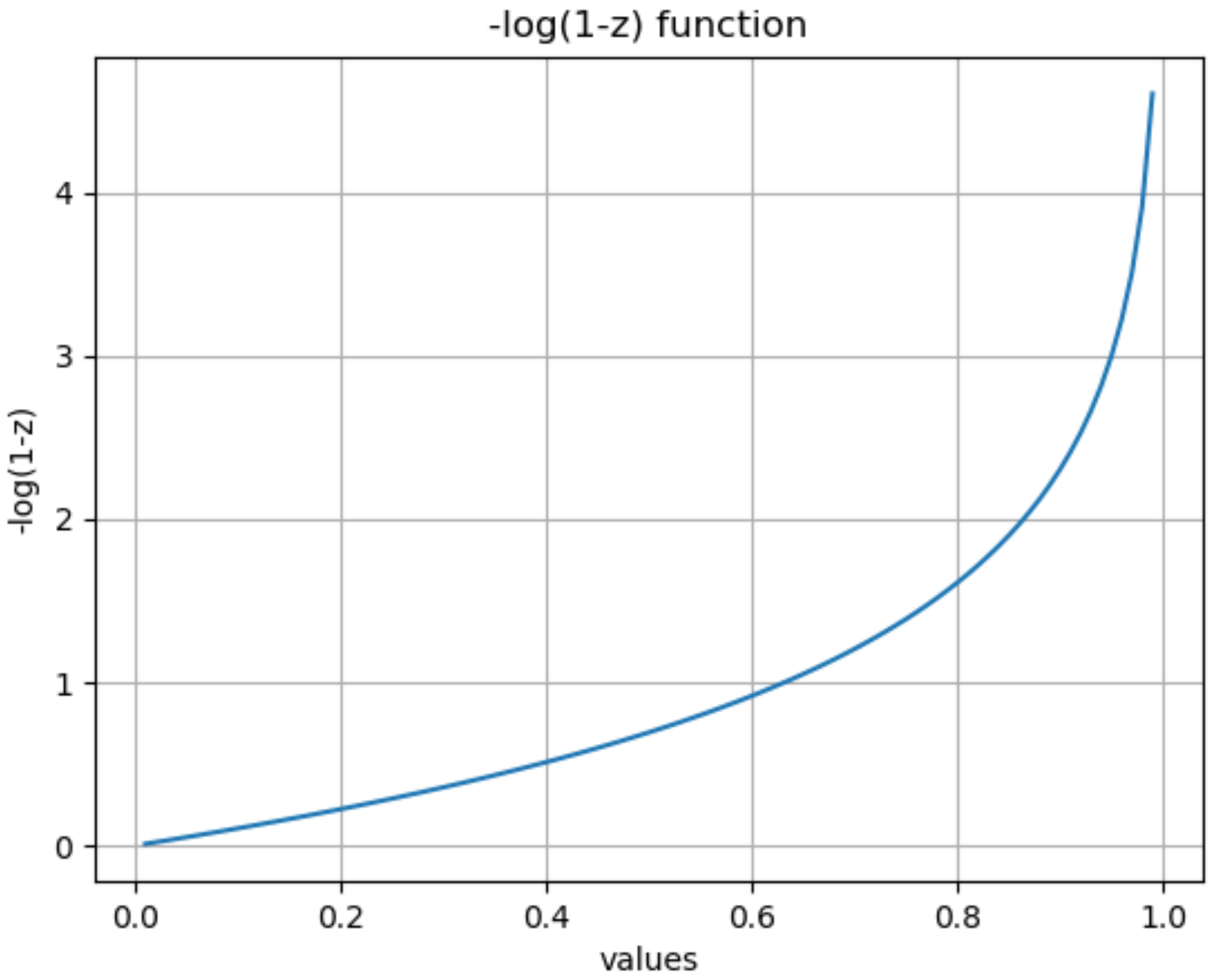
Logistic Regression

Cross-Entropy cost function

- cost $(h_w(x), y) =$
$$\begin{cases} -\log(h_w(x)) & if\ y = 1 \\ -\log(1 - h_w(x)) & if\ y = 0 \end{cases}$$

- $h_w(x) = 0$, cost is 0

- $h_w(x) = 1$, penalization with large cost



Logistic Regression

Cross-Entropy cost function - Putting things together

- cost $(h_w(x), y) = -y \log(h_w(x)) - (1-y)\log(1-h_w(x))$
- $J(w) = -y \log(h_w(x)) - (1-y)\log(1-h_w(x))$
- At $y = 1,\ J(w) =$?

Logistic Regression

Cross-Entropy cost function - Putting things together

- cost $(h_w(x), y) = -y \log(h_w(x)) - (1-y)\log(1-h_w(x))$
- $J(w) = -y \log(h_w(x)) - (1-y)\log(1-h_w(x))$
- At $y = 0$, $J(w)$ = ?

Logistic Regression

Cross-Entropy cost function - Minimization

- cost $(h_w(x), y) = \ -y\ \log(h_w(x)) - (1 - y)\log(1 - h_w(x))$
- $J(w) = -\ y\ \log(h_w(x)) - (1 - y)\log(1 - h_w(x))$
- min $J(w)$

Logistic Regression

Gradient descent!

- $J(w) = \frac{1}{m}\sum_{i=1}^{m} - y^{(i)} \log(h_w(x^{(i)})) - (1 - y^{(i)})\log(1 - h_w(x^{(i)}))$
- min $J(w)$

Logistic Regression

Gradient descent!

- $J(w) = \frac{1}{m}\sum_{i=1}^{m} - y^{(i)} \log(h_w(x^{(i)})) - (1 - y^{(i)})\log(1 - h_w(x^{(i)}))$
- $\frac{\partial J}{\partial w} = \frac{\partial J}{\partial h} \cdot \frac{\partial h}{\partial w}$

Logistic Regression

Gradient descent!

- $J(w) = \dfrac{1}{m} \sum_{i=1}^{m} - y^{(i)} \log(h_w(x^{(i)})) - (1 - y^{(i)})\log(1 - h_w(x^{(i)}))$
- $\dfrac{\partial J}{\partial h} = ?$

Logistic Regression

Gradient descent!

- $J(w) = \dfrac{1}{m}\sum_{i=1}^{m} - y^{(i)} \log(h_w(x^{(i)})) - (1 - y^{(i)})\log(1 - h_w(x^{(i)}))$
- $\dfrac{\partial J}{\partial h} = \dfrac{-y}{h} - \dfrac{1-y}{1-h}(-1)$
- $\dfrac{\partial J}{\partial h} = \dfrac{h-y}{h(1-h)}$

Logistic Regression

Gradient descent!

- $J(w) = \frac{1}{m}\sum_{i=1}^{m} - y^{(i)} \log(h_w(x^{(i)})) - (1 - y^{(i)})\log(1 - h_w(x^{(i)}))$
- $\frac{\partial h}{\partial w} = ?$

Logistic Regression

Gradient descent!

- $\sigma(z) = \dfrac{1}{1 + e^{-z}}$
- $\dfrac{\partial h}{\partial w} = ?$
- $\dfrac{\partial \sigma}{\partial z} = ?$

Logistic Regression

Gradient descent!

- $J(w) = \frac{1}{m}\sum_{i=1}^{m} - y^{(i)} \log(h_w(x^{(i)})) - (1 - y^{(i)})\log(1 - h_w(x^{(i)}))$
- $\frac{\partial h}{\partial w} = \sigma(1 - \sigma)x$
- $\frac{\partial J}{\partial h} = \frac{h - y}{h(1 - h)}$
- $\frac{\partial J}{\partial w} = \frac{\partial J}{\partial h} . \frac{\partial h}{\partial w}$

Logistic Regression

Gradient descent!

- $J(w) = \dfrac{1}{m}\sum_{i=1}^{m} - y^{(i)} \log(h_w(x^{(i)})) - (1 - y^{(i)})\log(1 - h_w(x^{(i)}))$
- $\dfrac{\partial h}{\partial w} = \sigma(1 - \sigma)x$
- $\dfrac{\partial J}{\partial w} = (h - y)x$

Logistic Regression

Gradient descent!

- $J(w) = \dfrac{1}{m}\sum_{i=1}^{m} - y^{(i)} \log(h_w(x^{(i)})) - (1 - y^{(i)})\log(1 - h_w(x^{(i)}))$
- $\dfrac{\partial J}{\partial w_j} = \dfrac{1}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})x_j^{(i)}$

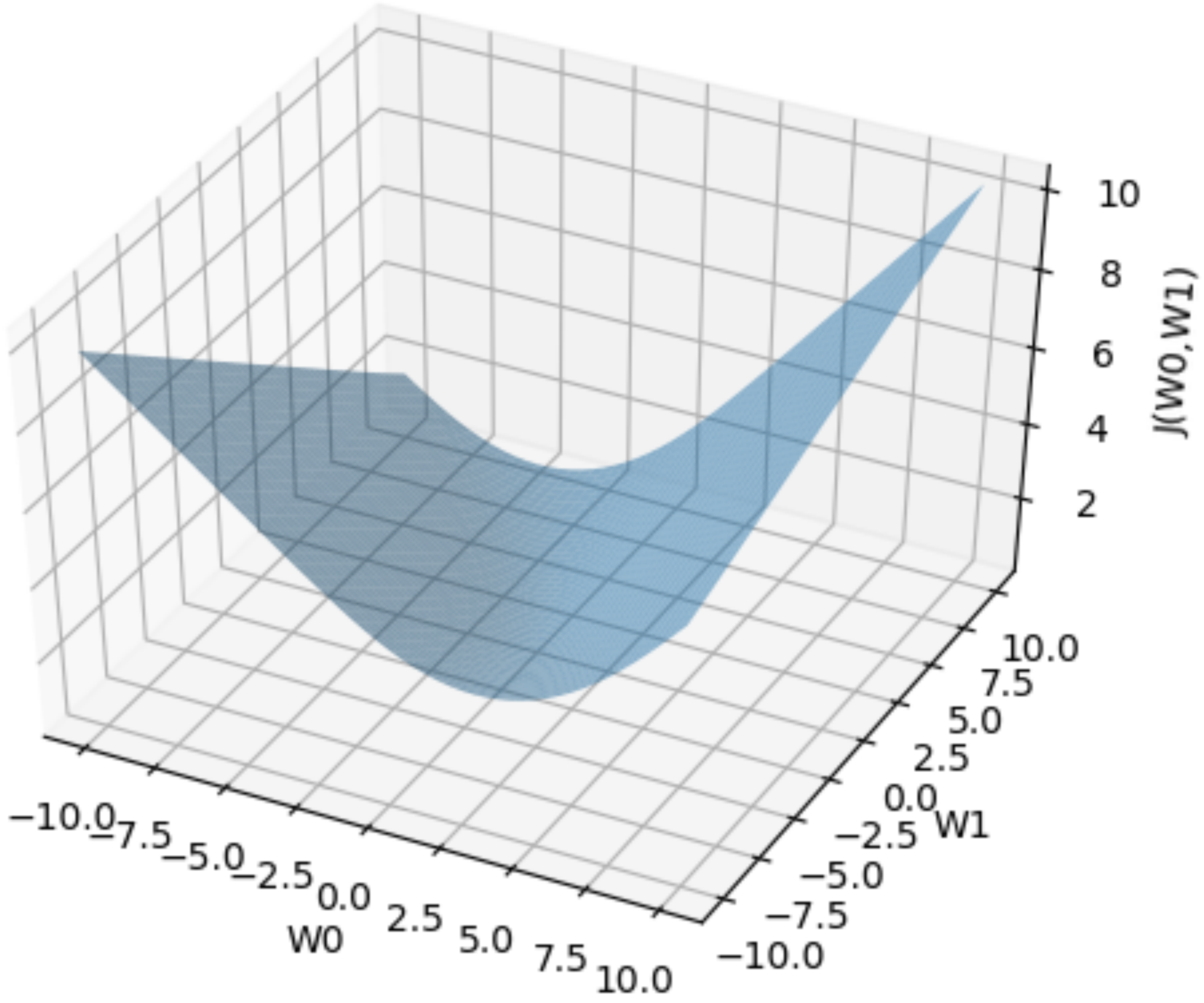
Logistic Regression

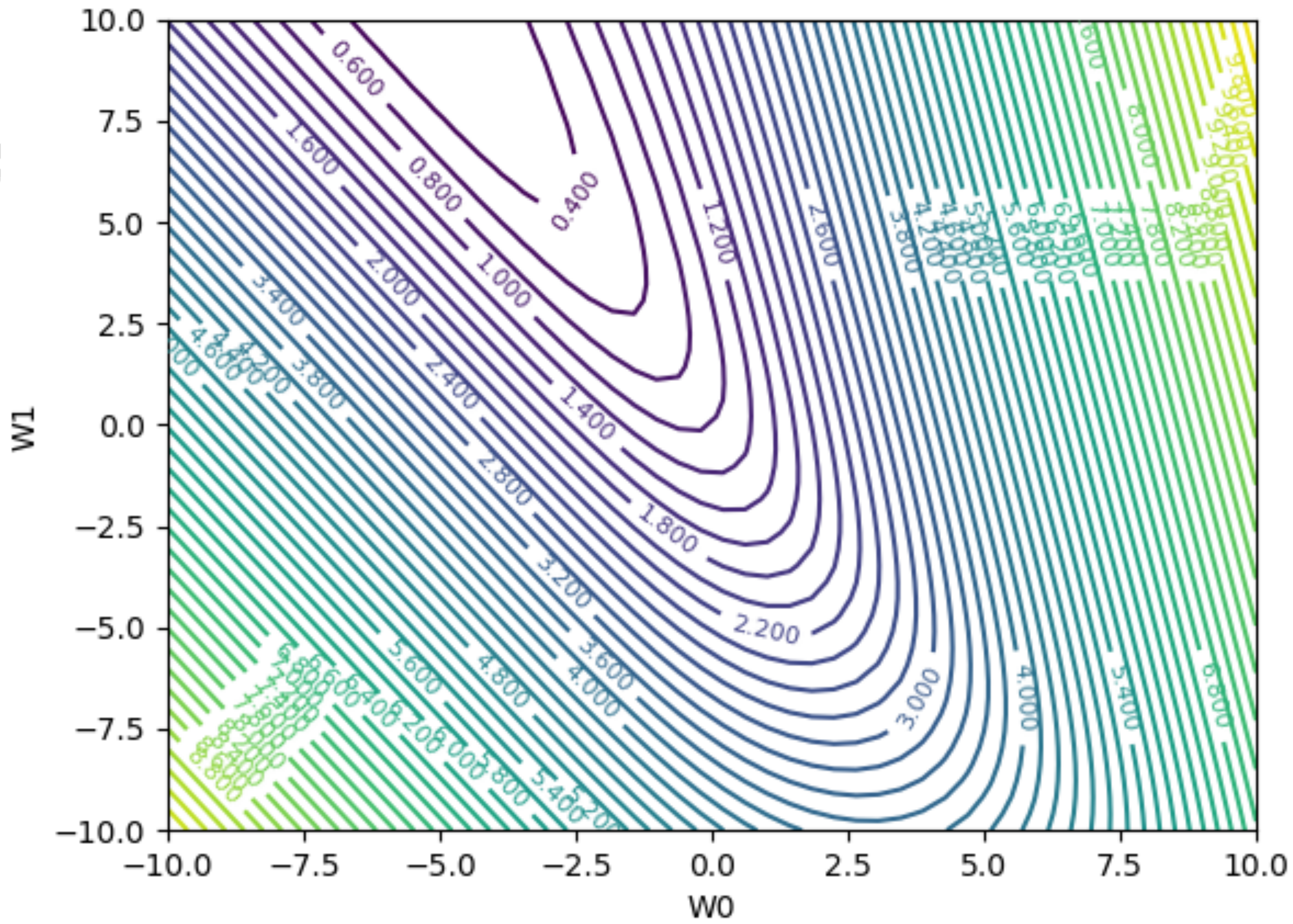
Gradient descent update

- $J(w) = \frac{1}{m}\sum_{i=1}^{m} - y^{(i)} \log(h_w(x^{(i)})) - (1 - y^{(i)})\log(1 - h_w(x^{(i)}))$
- $\frac{\partial J}{\partial w_j} = \frac{1}{m}\sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})x_j^{(i)}$
- $w_j^{k+1} = w_j^k - \alpha_k \frac{\partial J}{\partial w_j}$

Logistic Regression

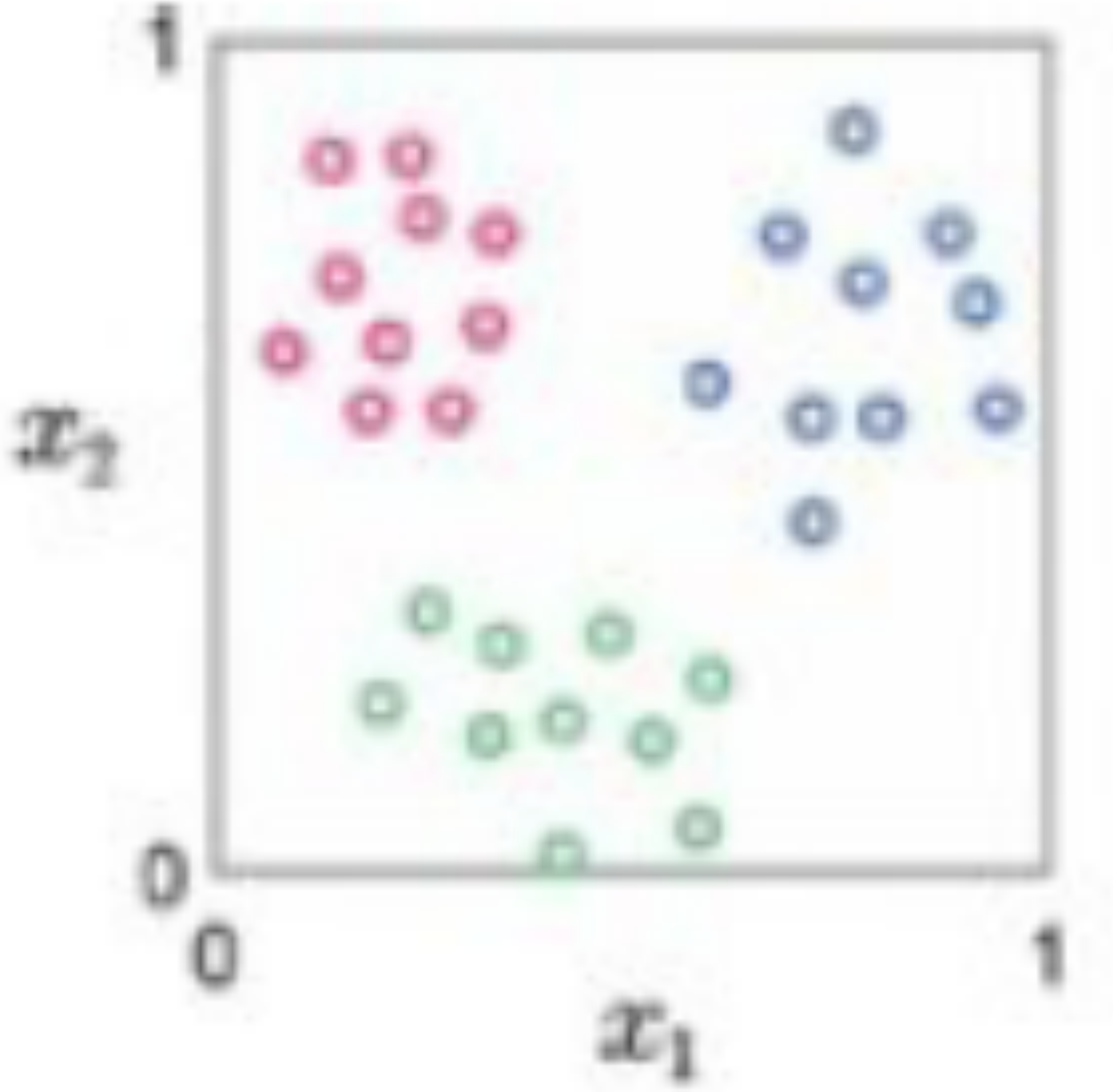
Plot the cost function $J(w)$





One-vs-all (OvA) multi-class classification

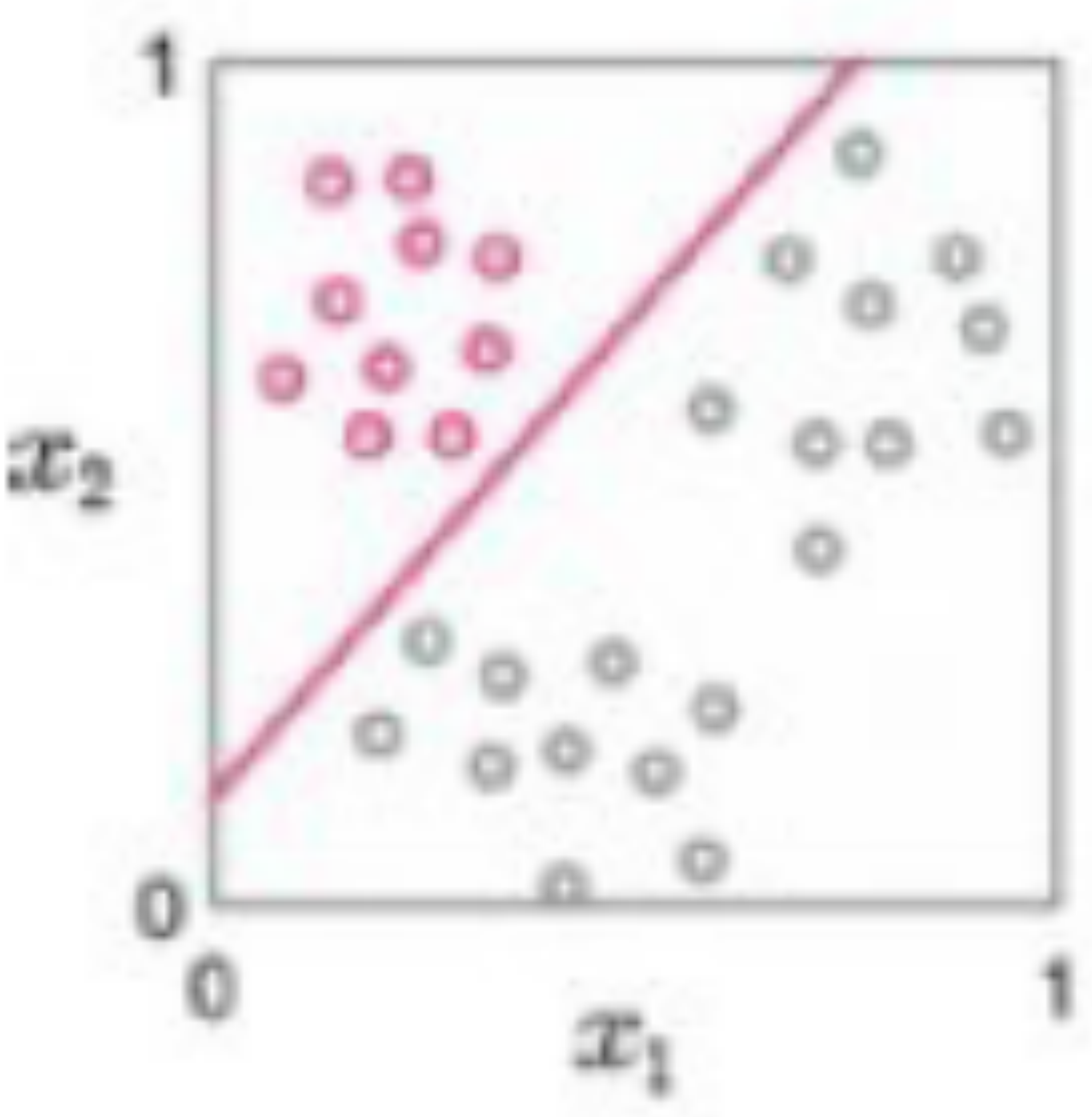
OvA



One-vs-all (OvA) multi-class classification

OvA

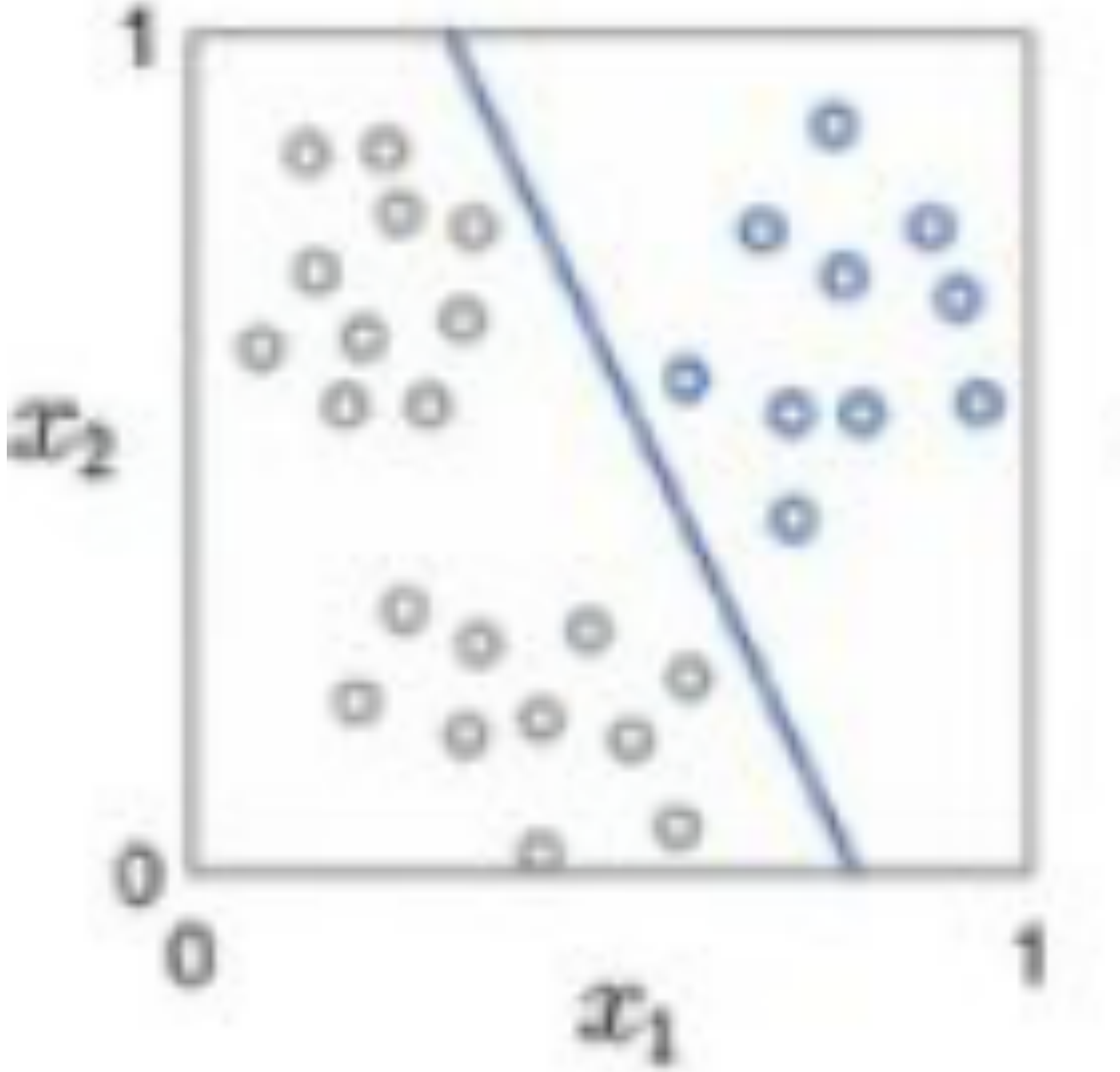
$h_w^{(1)}(x)$



One-vs-all (OvA) multi-class classification

OvA

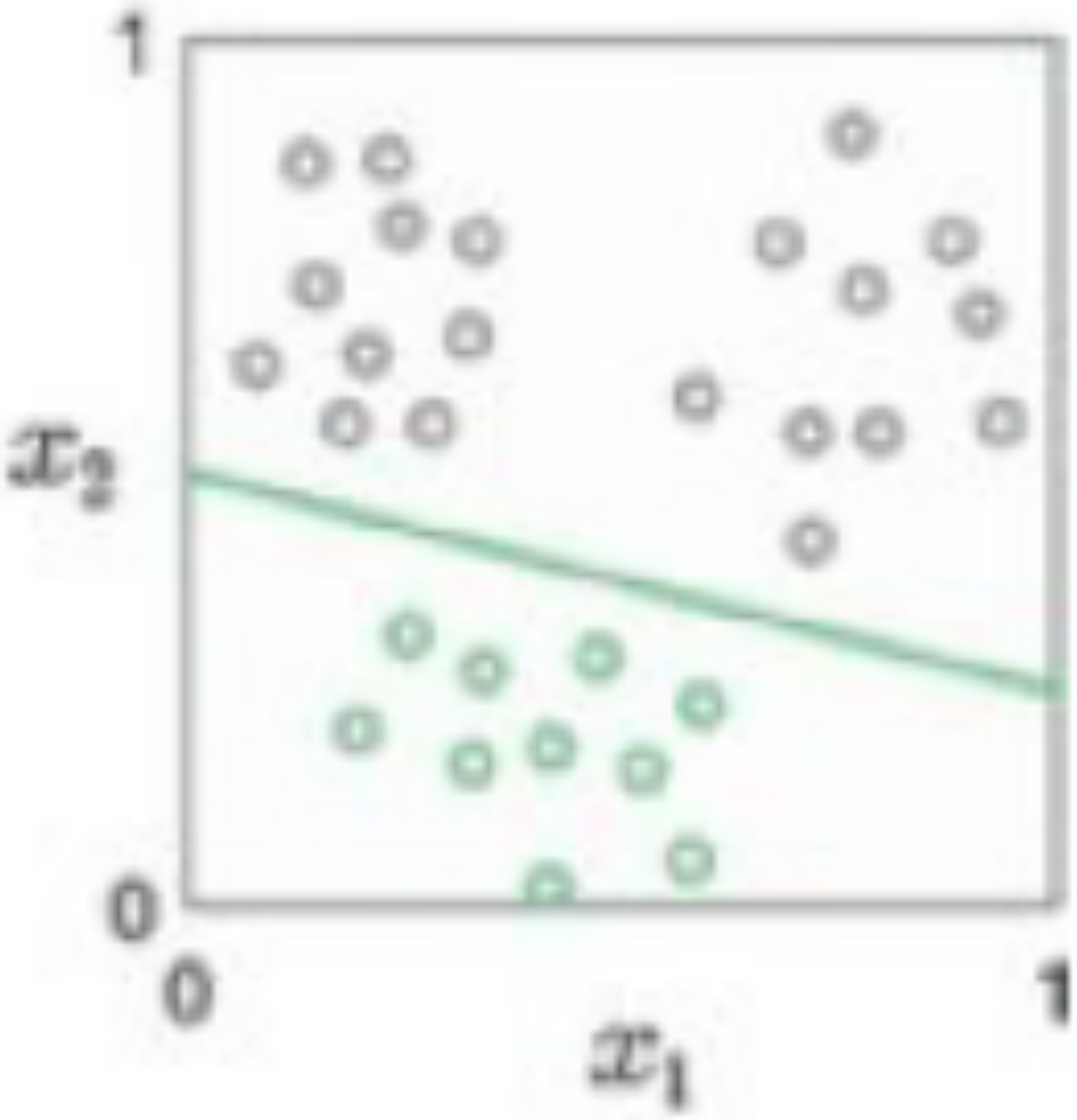
$h_w^{(2)}(x)$



One-vs-all (OvA) multi-class classification

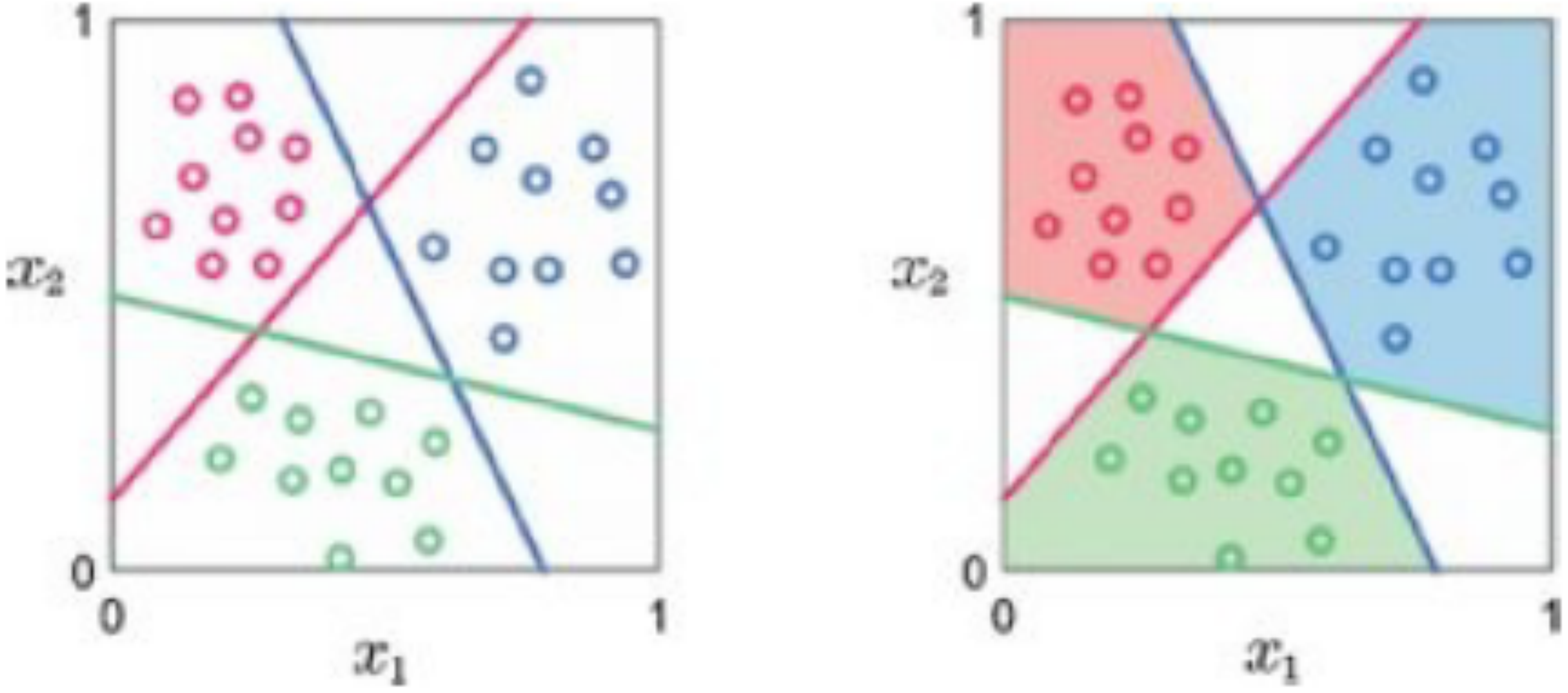
OvA

$h_w^{(3)}(x)$



One-vs-all (OvA) multi-class classification

OvA



One-vs-all (OvA) multi-class classification

OvA - Fusion rule

$$\max_i h_w^{(i)}(x)$$