CSC345/M45: Big Data & Machine Learning (logistic regression)

Prof. Xianghua Xie

x.xie@swansea.ac.uk
http://csvision.swan.ac.uk
224 Computational Foundry, Bay Campus

Recap: Linear Regression

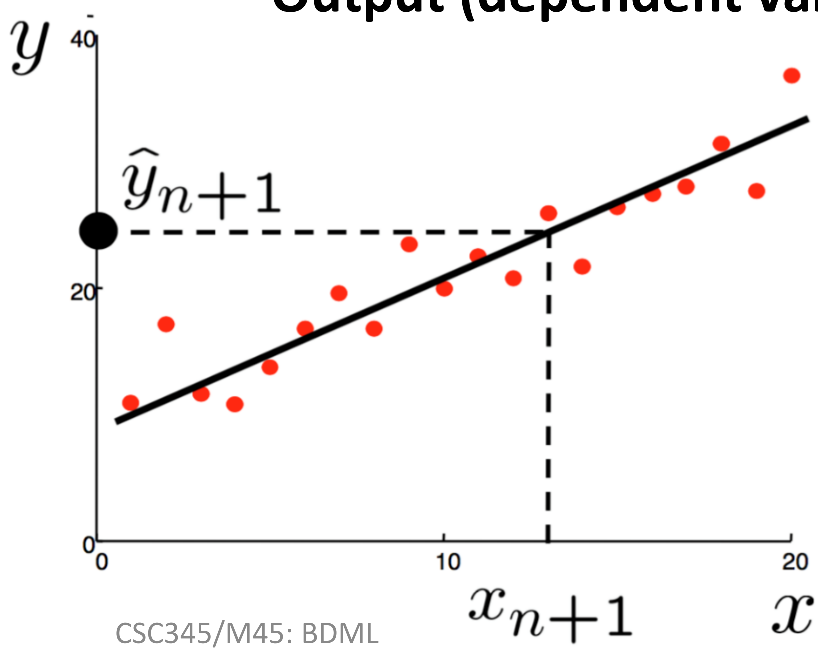
- We wish to estimate $\hat{y}$ by a linear function of data x:

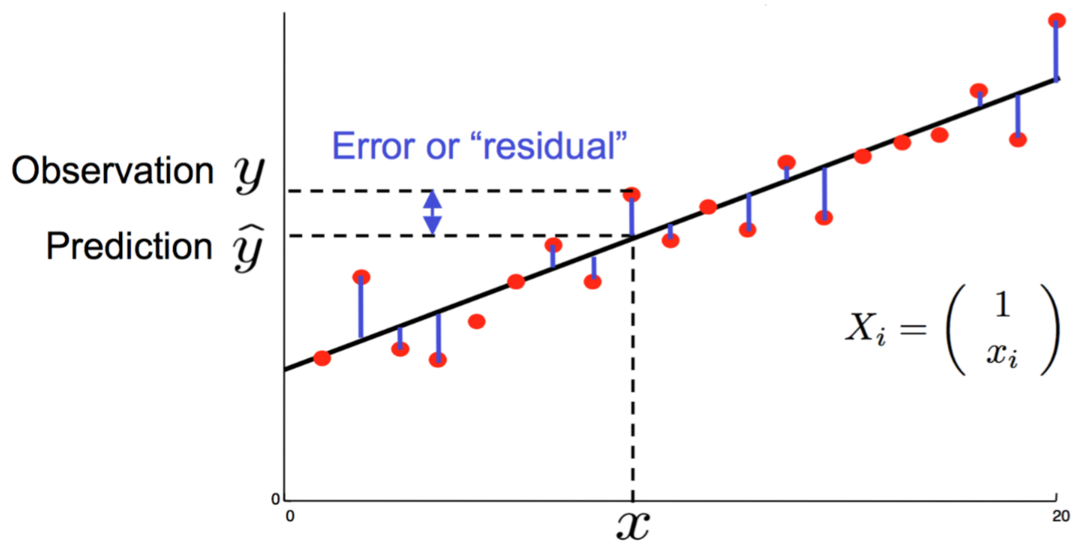
$$\hat{y}_{n+1} = \omega^T x_{n+1}$$

  - where $\omega = (\omega_0, \omega_1, \omega_2, \ldots)^T$, $x_{n+1} = (1, x_{n+1,1}, x_{n+1,2}, \ldots)^T$
  - Cost function: Least Mean Squares (LMS)

$$E = \frac{1}{2}\sum_{i=1}^{n}(w^\top x_i - y_i)^2 = \sum_{i=1}^{n} E_i$$

  - **Output (dependent variable): continuous**



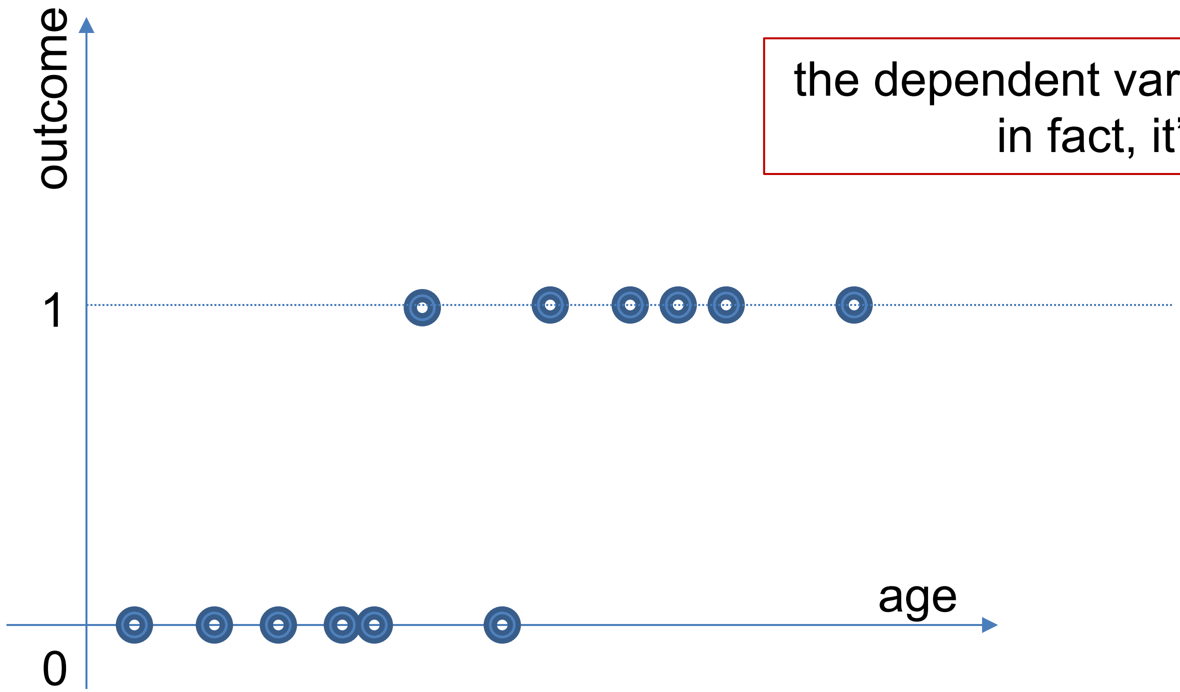


Toy Problem

- Consider the following problem
  - Since 2001, European clinicians pioneered a novel clinical procedure to treat patients with aortic valve disease by inserting artificial valves through catheter (known as TAVI), rather than using open heart surgery. The procedure was introduced to UK in 2007.
  - Among many others factors, age is one key factor that impacts on the outcome of the procedure.
  - Clinicians try to determine the “safe age” to adopt this operation on patients
    - That is below this age, the benefit of this operation outweights the risk compared to open heart surgery
  - The following data has been collected

Toy Problem

- The outcome of TAVI are classified as "successful" or "unsuccessful"
  - "1" denotes successful outcome (benefit over-weights risk), y=1
  - "0" denotes unsuccessful outcome (risk over-weights benefit), y=0

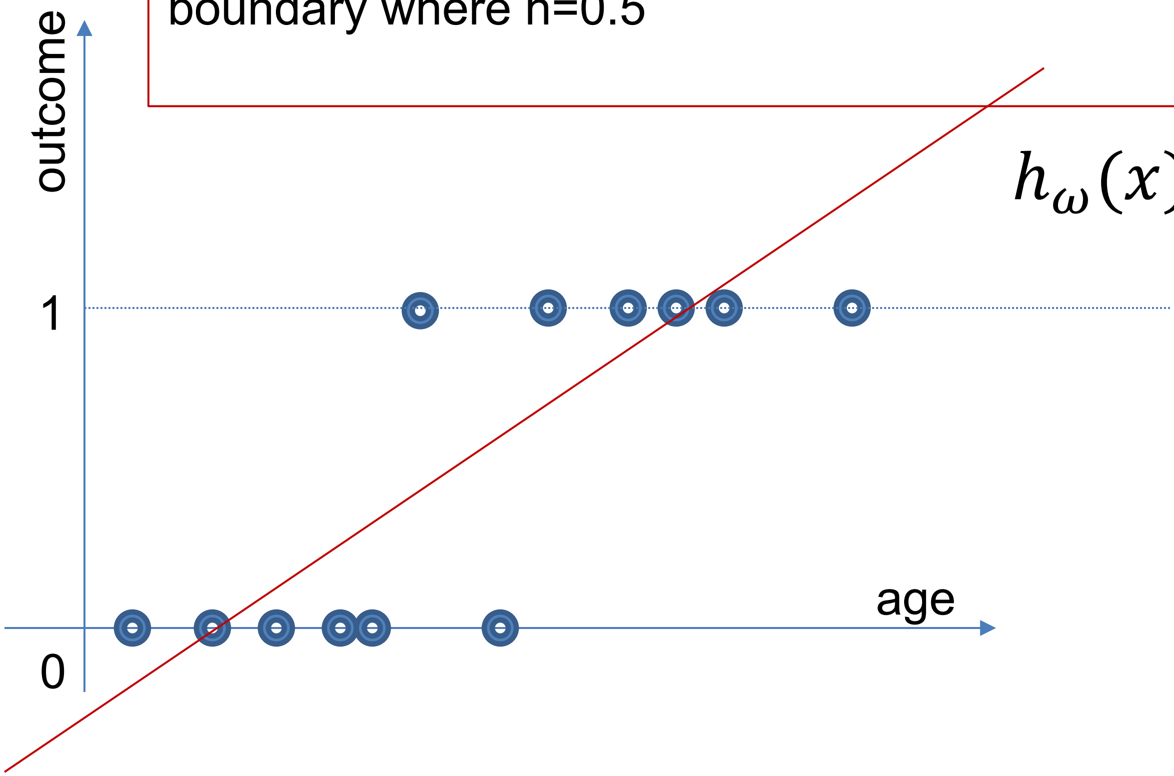


the dependent variable is not continuous; in fact, it's binary {0,1}

Toy Problem

- Apply linear regression to our problem
  - The red line depicts the linear regression result

Two problems with this:
1. Dependent variable is not in binary form and can be less than 0 and larger than 1; not so serious since we can still draw a decision boundary where h=0.5



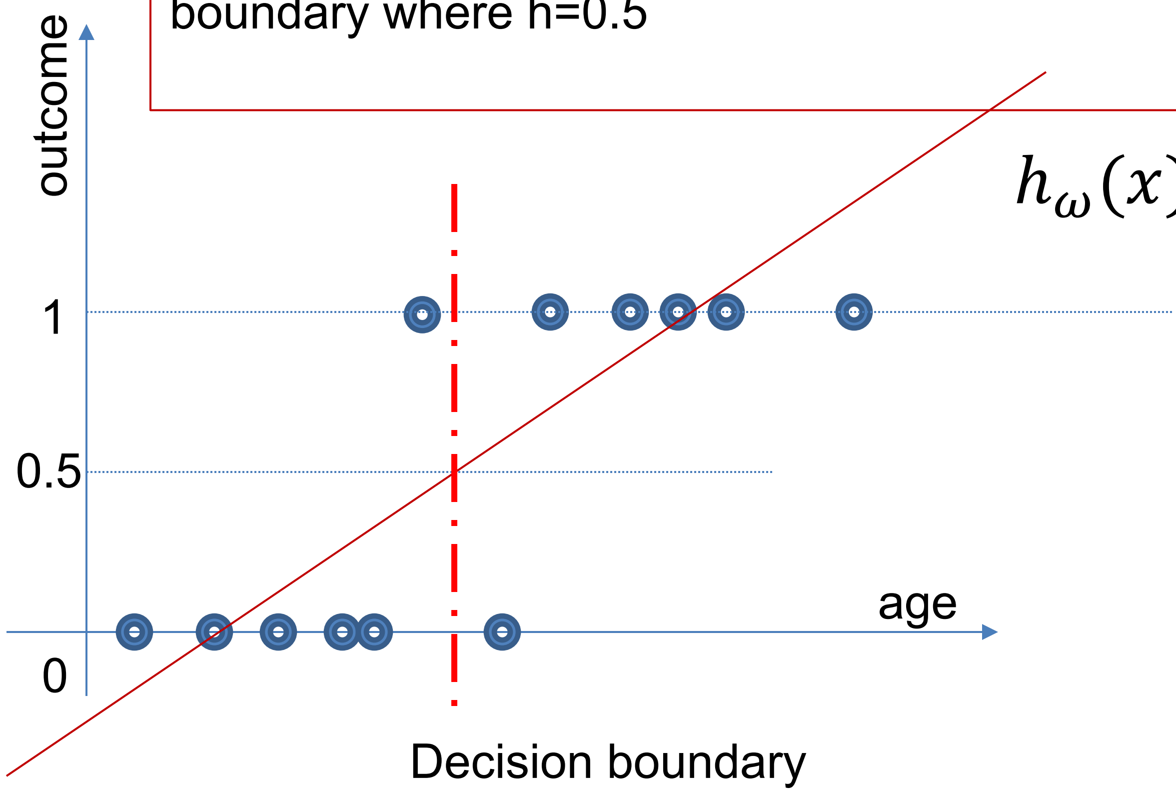
$$h_\omega(x) = \omega^T x$$

Toy Problem

- Apply linear regression to our problem
  - The red line depicts the linear regression result

Two problems with this:
1. Dependent variable is not in binary form and can be less than 0 and larger than 1; not so serious since we can still draw a decision boundary where h=0.5

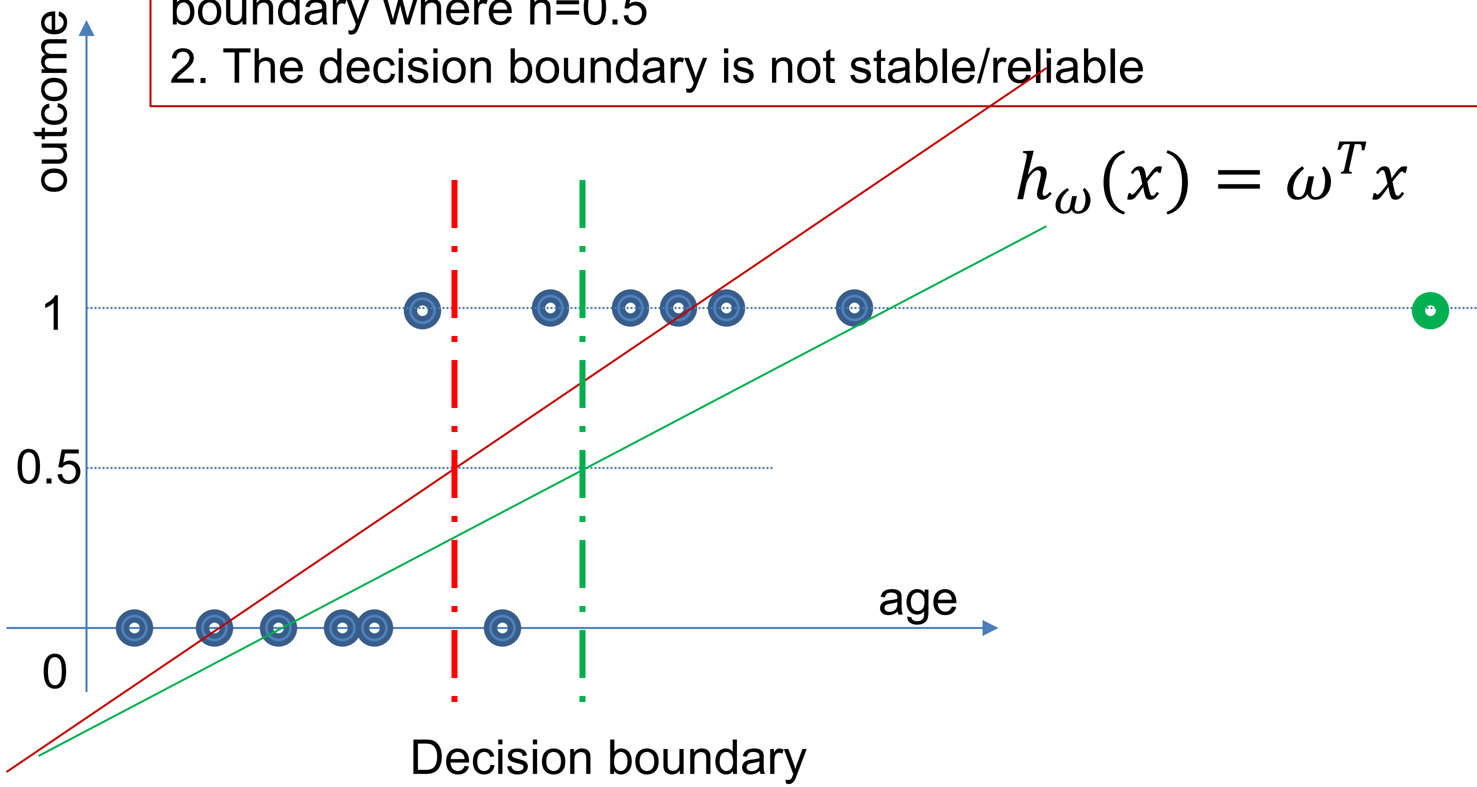
$$h_\omega(x) = \omega^T x$$



Toy Problem

- Apply linear regression to our problem
  - The red line depicts the linear regression result

Two problems with this:
1. Dependent variable is not in binary form and can be less than 0 and larger than 1; not so serious since we can still draw a decision boundary where h=0.5
2. The decision boundary is not stable/reliable



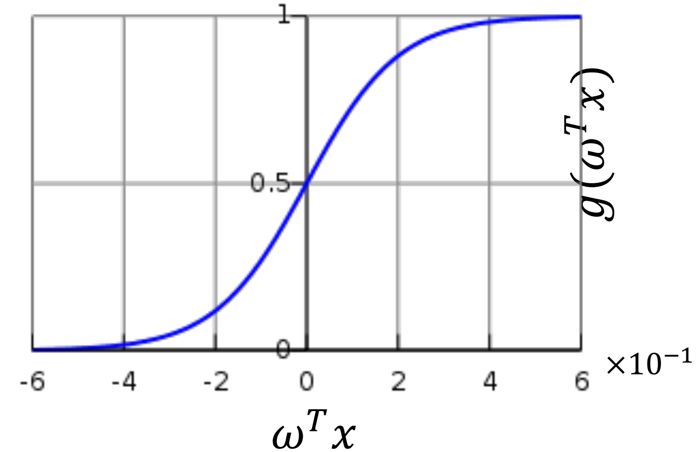
Logistic Regression

- The dependent variable needs fall in [0,1]
  - We commonly denote h as “hypothesis”

$$0 \leq h_\omega(x) \leq 1$$

- Apply **logistic function**
  - Also known as **sigmoid function**

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



  - The regression (originally $h_\omega(x) = \omega^T x$) now takes the following form:

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

Probability Interpolation

- $h_\omega(x)$ can be interpolated as probability that $y = 1$ on input x.

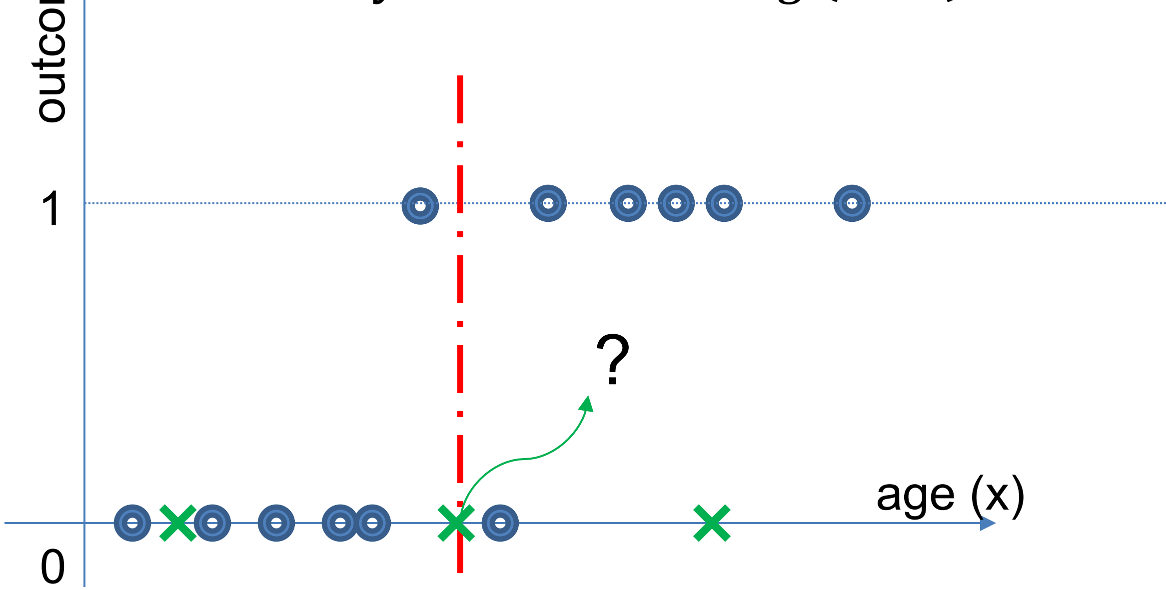
$$h_\omega(x) = g(\omega^T x) = p(y = 1|x; \omega)$$

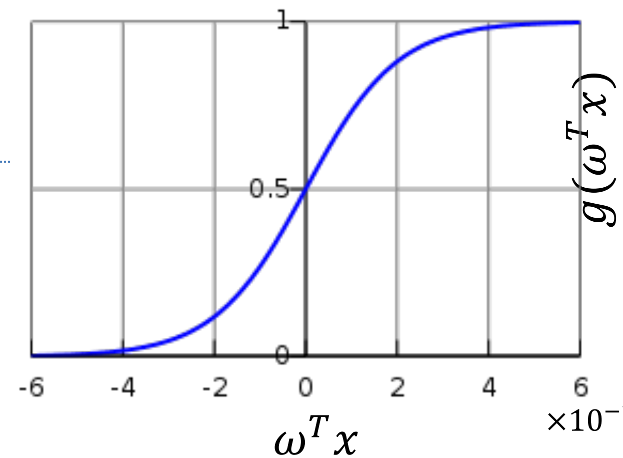
$$p(y = 0|x; \omega) + p(y = 1|x; \omega) = 1$$

  – E.g. **one-dimensional feature**

    Predict $y = 1$: $\omega^T x \geq 0, g(\omega^T x) \geq 0.5$

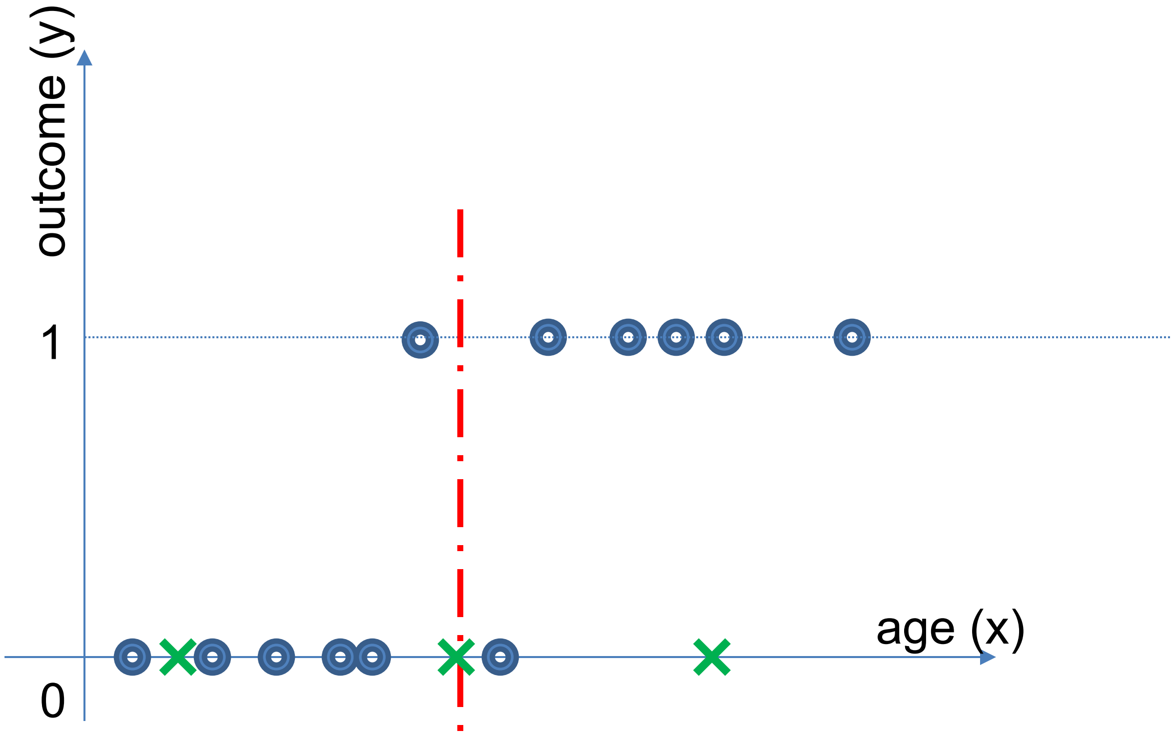
    Predict $y = 0$: $\omega^T x < 0, g(\omega^T x) < 0.5$

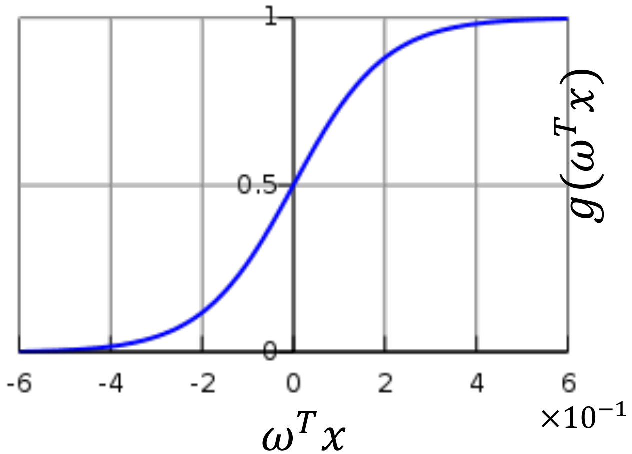




Decision Boundary

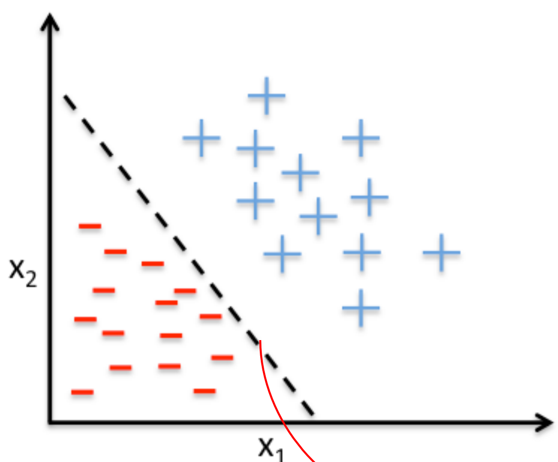
- Decision boundary for one-dimensional feature
  - Is simply a threshold value
  - Or a vertical line in the illustrated example



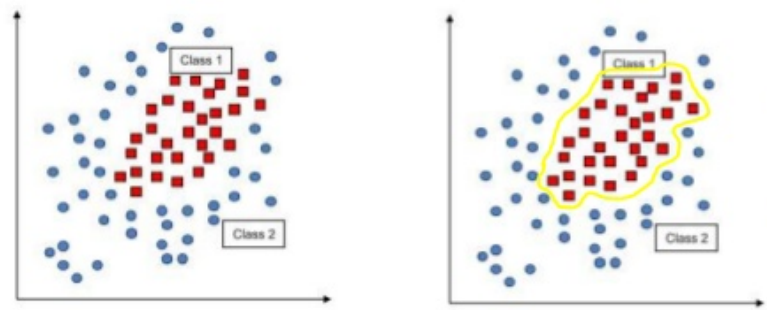


Decision Boundary

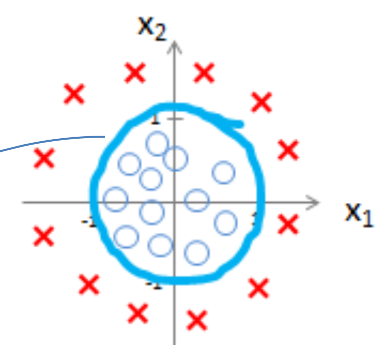
- In the case of more than one-dimensional feature
  - E.g. two-dimensional feature space:



Example of a linear decision boundary for binary classification.



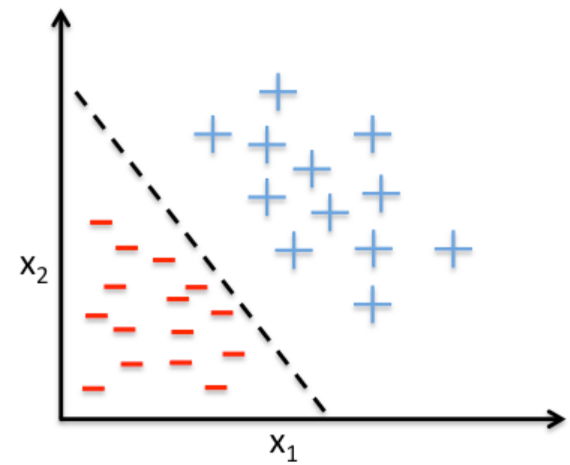
Non-linear decision boundaries



How to represent these decision boundaries?

Linear Decision Boundary

- Take the two-dimensional feature space as an example:
  - The decision boundary is a straight line:
    - $\omega^T x = 0$
    - positive samples >0 (but can be much higher than 0)
    - negative samples <0 (can be much lower than 0)
  - Wrap the decision boundary (line) equation with logistic function:
    - $h_\omega(x) = g(\omega^T x) \in [0,1]$
    - Output is now bounded
    - Predict $\mathrm{y} = 1$: $\omega^T x \geq 0, g(\omega^T x) \geq 0.5$
    - Predict $\mathrm{y} = 0$: $\omega^T x < 0, g(\omega^T x) < 0.5$
    - e.g. ω^T=(-2,2,1)

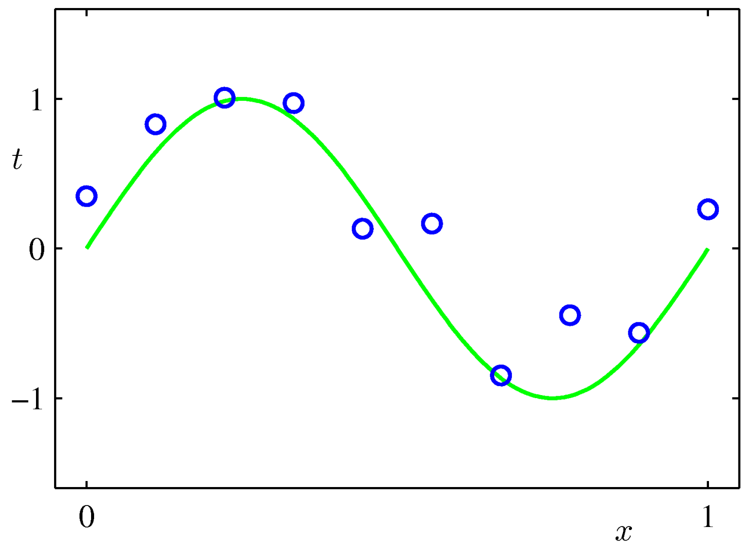


Example of a linear decision boundary for binary classification.

Nonlinear Decision Boundary

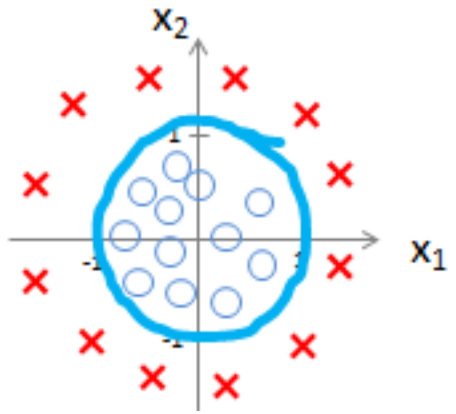
- Similar to the linear case, we try to fit the decision boundary
  - But with higher order polynomial functions

Recap: polynomial fitting
(e.g. one-dimensional independent variable)



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Non-linear decision boundaries

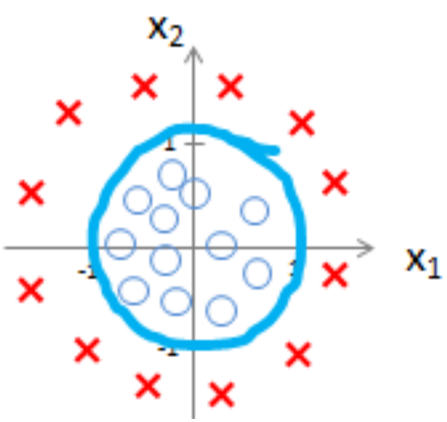


In the example above: two-dimensional independent variable

Nonlinear Decision Boundary

- Example: 2-dimensional feature space
  - the decision boundary is $\omega^T \phi(x) = 0$
  - except it involves nonlinear combinations of the features, **e.g.**

$$h_\omega(x) = g(\omega^T \phi(x)) = g(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \omega_4 x_2^2)$$



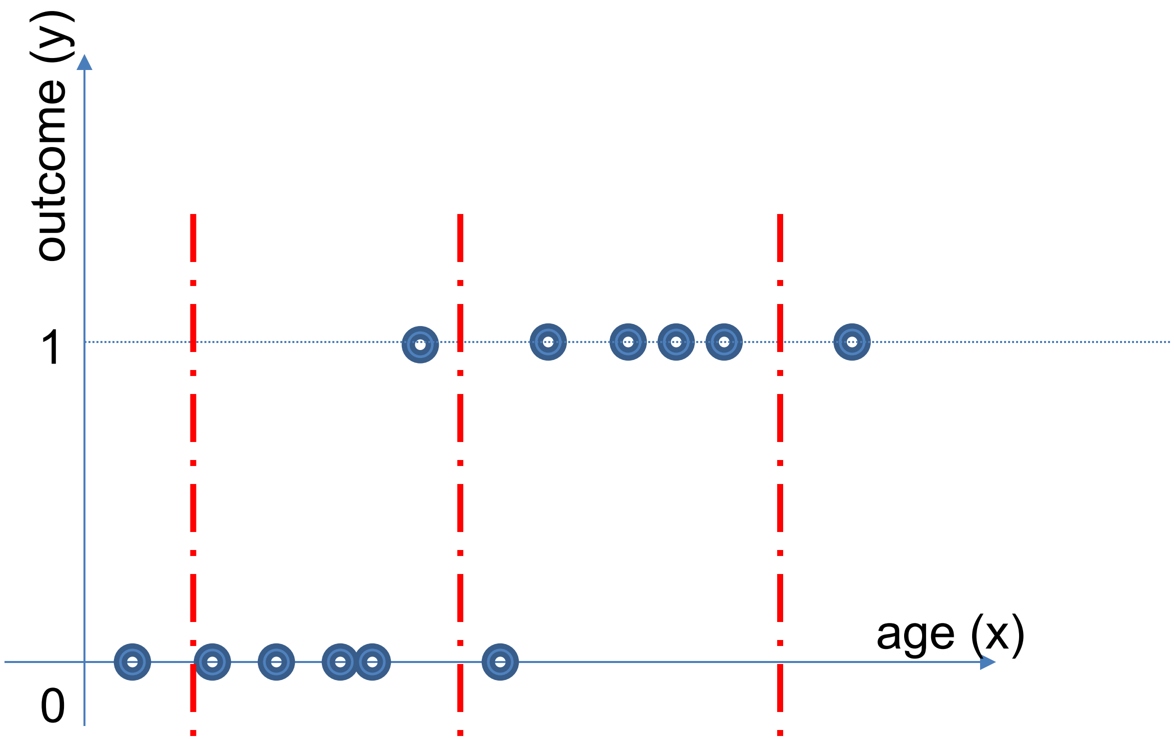
Suppose $\omega^T = (-1,0,0,1,1)$,
thus $h_\omega(x) = g(-1 + x_1^2 + x_2^2)$

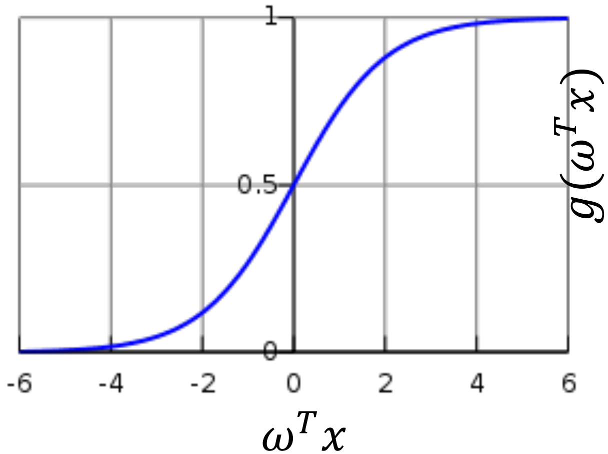
- Predict y=1, if:
- Predict y=0, if:

In practice, more complex polynomials can be used to represent complex decision boundaries.

What is a Good Decision Boundary?

- When y=1, for a given decision boundary:
  - Project all the positive samples (labelled as y=1) to the sigmoid/logistic function g
  - The projected values should be as close to 1 as possible

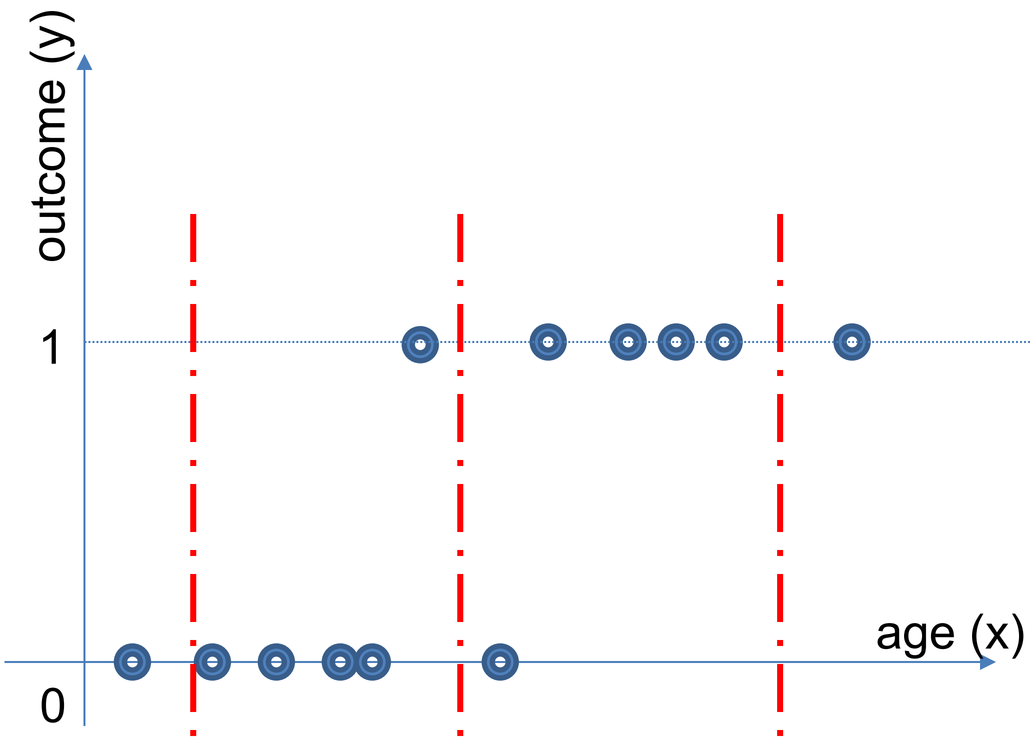


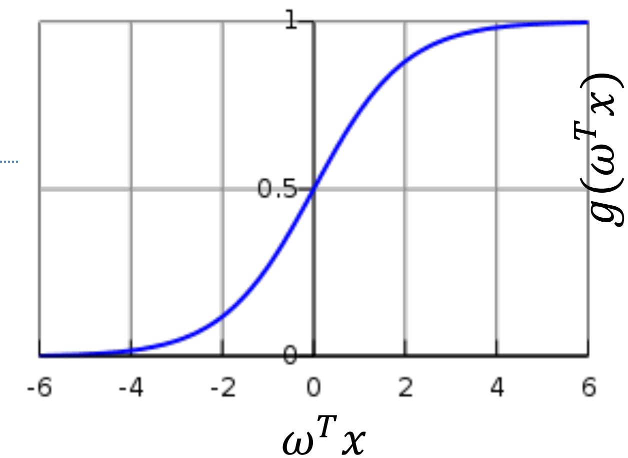


What is a Good Decision Boundary?

- When y=1, for a given decision boundary, i.e. ω:
  - the following mean should as close to 1 as possible:
    - the mean is bounded between 0 and 1

$$\frac{1}{m}\sum_{i=1}^{m} g(\omega^T x_i) = \frac{1}{m}\sum_{i=1}^{m} h_\omega(x_i)$$



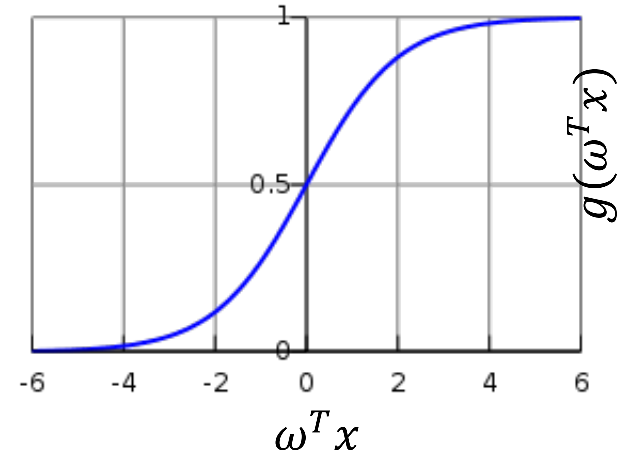


What is a Good Decision Boundary?

- When y=1, for a given decision boundary, i.e. ω:
  - the following mean should as close to 1 as possible:
    - the mean is bounded between 0 and 1

$$\frac{1}{m}\sum_{i=1}^{m} g(\omega^T x_i) = \frac{1}{m}\sum_{i=1}^{m} h_\omega(x_i)$$

  - however, function g is nonlinear and difficult to find the optimal solution

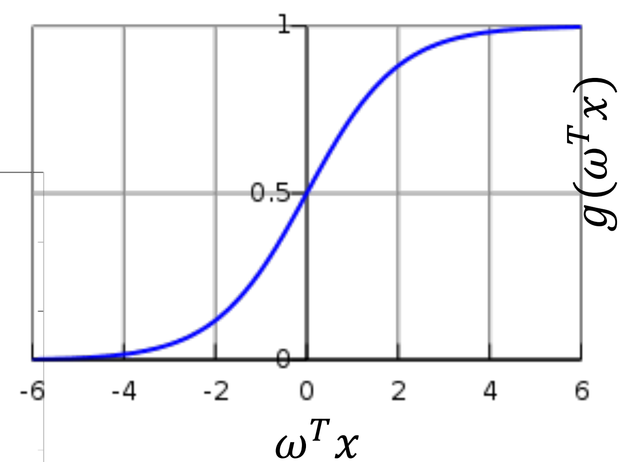


What is a Good Decision Boundary?

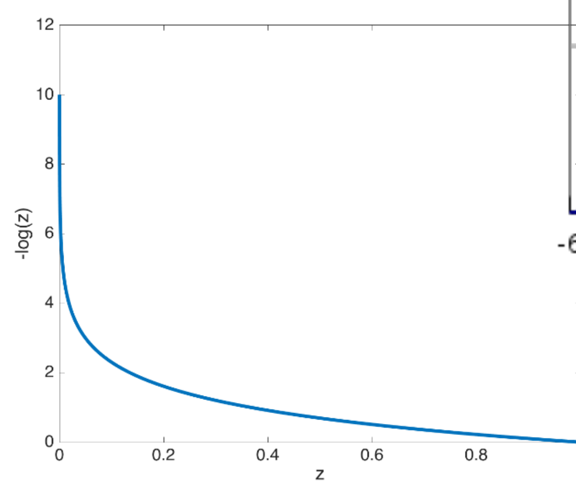
- When y=1, for a given decision boundary, i.e. ω:
  - the following mean should as close to 1 as possible:
    - the mean is bounded between 0 and 1

$$\frac{1}{m}\sum_{i=1}^{m} g(\omega^T x_i) = \frac{1}{m}\sum_{i=1}^{m} h_\omega(x_i)$$

  - however, function g is nonlinear and difficult to find the optimal solution
  - we apply negative log transform:



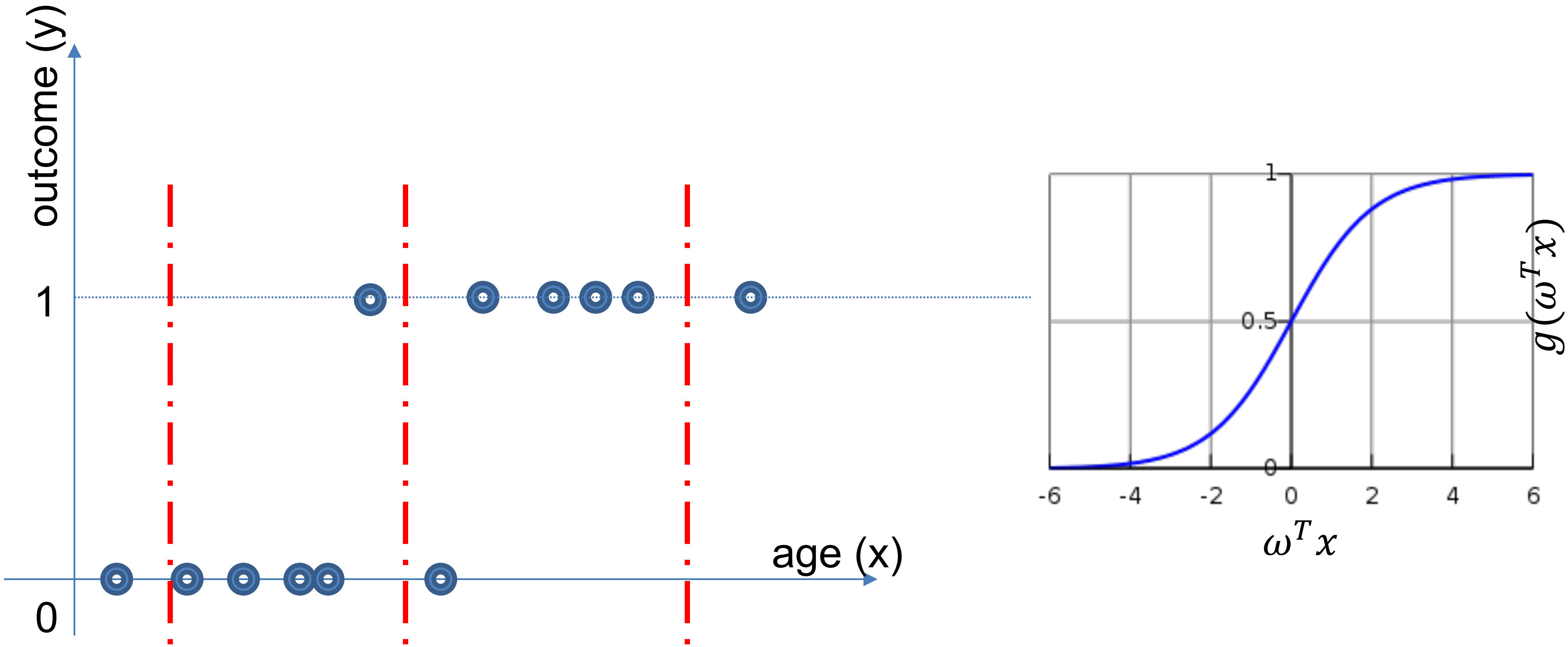
$$-\frac{1}{m}\sum_{i=1}^{m} \log h_\omega(x_i)$$



This value needs to be lower or higher?

What is a Good Decision Boundary?

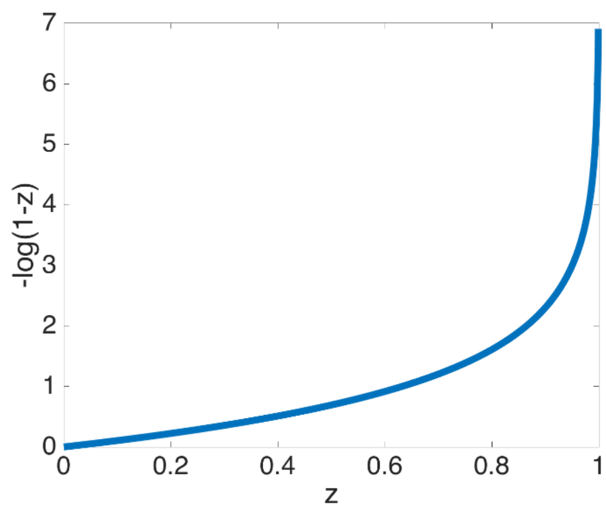
- When y=0:
  - Project all the negative samples (labelled as y=0) to the sigmoid/logistic function g
  - The projected values should be as close to 0 as possible

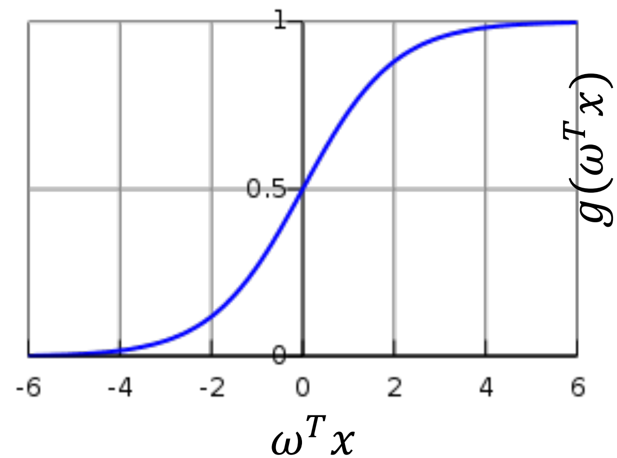


What is a Good Decision Boundary?

- When y=0:
  - Project all the negative samples (labelled as y=0) to the sigmoid/logistic function g
  - The projected values should be as close to 0 as possible
  - After applying the same negative log transform, we have the following:

$$-\frac{1}{n}\sum_{i=1}^{n}\log(1-h_\omega(x_i))$$





Cost Function

- To find the decision boundary, we ought to minimise:

$$-\frac{1}{m}\sum_{i=1}^{m}\log h_\omega(x_i) \qquad if\ y_i = 1$$

$$-\frac{1}{n}\sum_{i=1}^{n}\log(1 - h_\omega(x_i)) \qquad if\ y_i = 0$$

  - m: number of positive samples
  - n: number of negative samples

Cost Function

- To find the decision boundary, we ought to minimise:

$$-\frac{1}{m}\sum_{i=1}^{m}\log h_\omega(x_i) \qquad\qquad if\ y_i\ = 1$$

$$-\frac{1}{n}\sum_{i=1}^{n}\log(1-h_\omega(x_i)) \qquad\qquad if\ y_i = 0$$

  - m: number of positive samples
  - n: number of negative samples

- These two terms can be combined together
  - N: number of total samples

$$E(\omega) = -\frac{1}{N}\left[\sum_{i=1}^{N} y_i \log h_\omega(x_i) + \sum_{i=1}^{N}(1-y_i)\log(1-h_\omega(x_i))\right]$$

Cost Function

- **Regularised** logistic regression cost function:

$$E(\omega) = -\frac{1}{N}\left[\sum_{i=1}^{N} y_i \log h_\omega(x_i) + \sum_{i=1}^{N}(1-y_i)\log(1-h_\omega(x_i))\right] + \frac{\lambda}{2N}\omega^T\omega$$

$$= -\frac{1}{N}\sum_{i=1}^{N}[y_i \log h_\omega(x_i) + (1-y_i)\log(1-h_\omega(x_i))] + \frac{\lambda}{2N}\omega^T\omega$$

Cost Function

- Regularised logistic regression cost function:

$$E(\omega) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log h_\omega(x_i) + (1 - y_i)\log(1 - h_\omega(x_i))] + \frac{\lambda}{2N}\omega^T\omega$$

- To find the parameters: $\min_{\omega} E(\omega)$
  - Again, use gradient descent

$$\omega_j := \omega_j - \alpha\frac{\partial}{\partial\omega_j}E(\omega)$$

Cost Function

- Regularised logistic regression cost function:

$$E(\omega) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log h_\omega(x_i) + (1 - y_i)\log(1 - h_\omega(x_i))] + \frac{\lambda}{2N}\omega^T\omega$$

- To find the parameters: $\min_{\omega} E(\omega)$
  - Again, use gradient descent

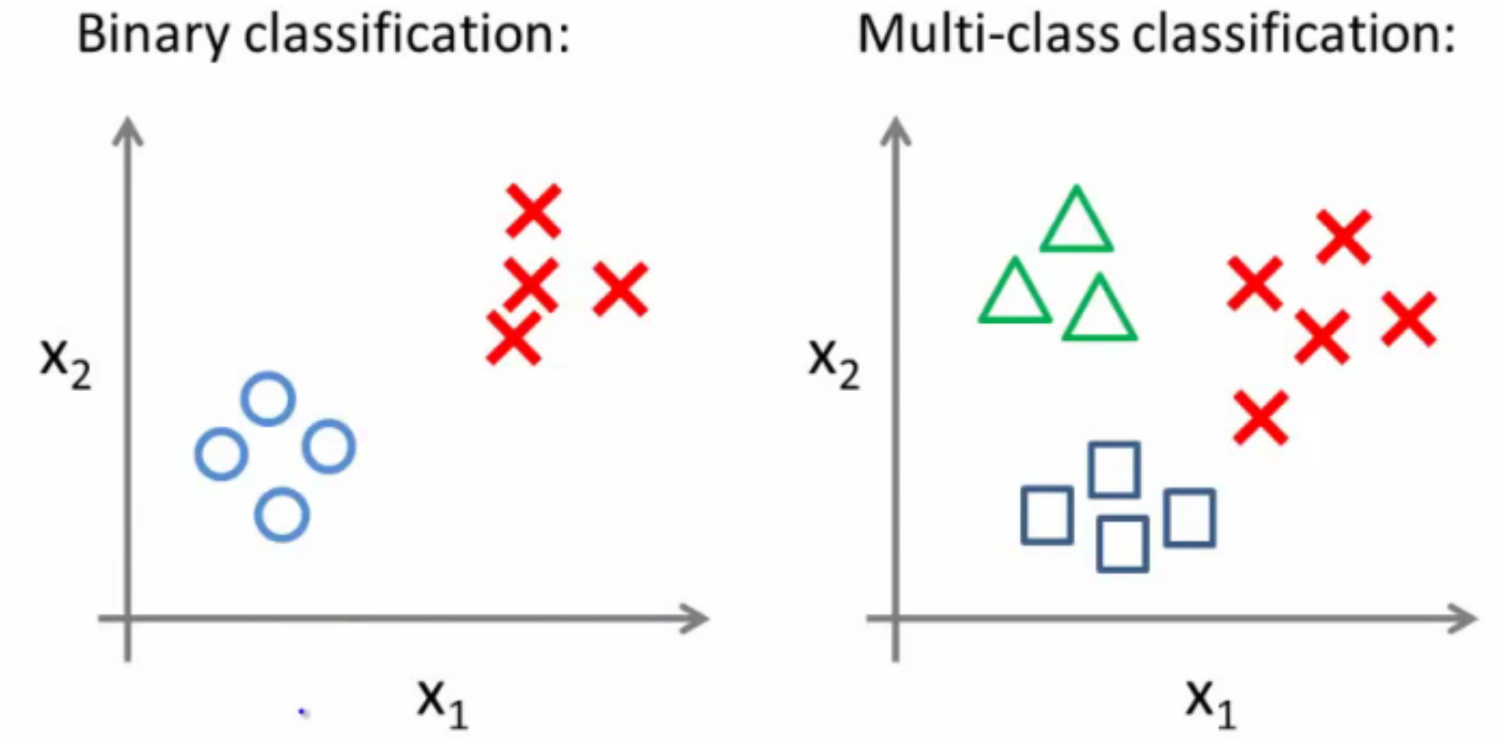
$$\omega_j := \omega_j - \alpha \frac{\partial}{\partial \omega_j} E(\omega)$$

- To make a prediction for a new data x_{N+1}:

Calculate $h_\omega(x_{N+1}) = \frac{1}{1+e^{-\omega^T x_{N+1}}}$

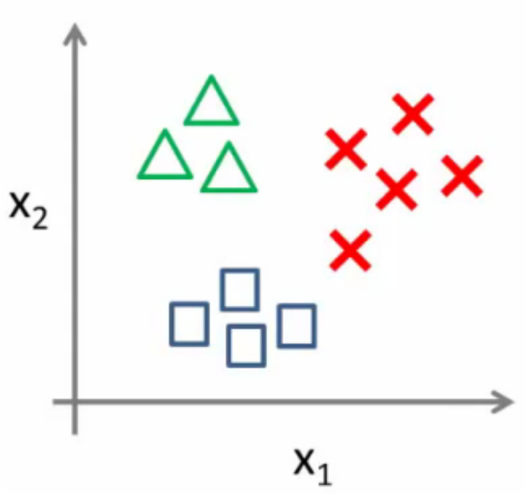
Multi-class Classification

- E.g. weather forecast: sunny, cloudy, rain, snow …



One vs. All

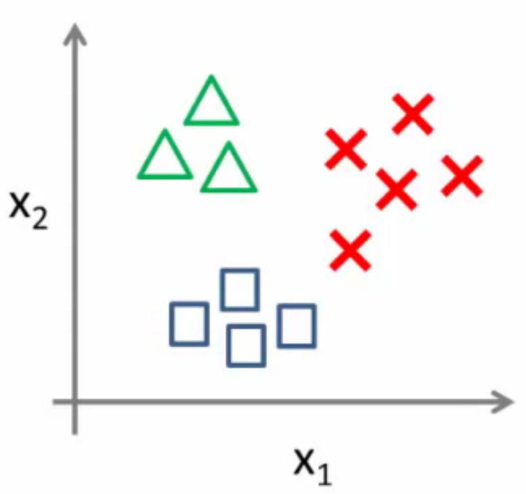
- Apply one-vs-all training strategy



$$h_\omega^{(i)}(x) = P(y = i|x;\omega), \qquad i = 1,2,3$$

One vs. All

- Apply one-vs-all training strategy



$$h_\omega^{(i)}(x) = P(y = i|x; \omega), \qquad i = 1,2,3$$

- At the testing stage: for new data x, pick the class i that maximises:

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

Example

Case Summaries[a]

		Cured?	Intervention	Number of Days with Problem before Treatment	Predicted probability	Predicted group
1		Not Cured	No Treatment	7	.42857	Not Cured
2		Not Cured	No Treatment	7	.42857	Not Cured
3		Not Cured	No Treatment	6	.42857	Not Cured
4		Cured	No Treatment	8	.42857	Not Cured
5		Cured	Intervention	7	.71930	Cured
6		Cured	No Treatment	6	.42857	Not Cured
7		Not Cured	Intervention	7	.71930	Cured
8		Cured	Intervention	7	.71930	Cured
9		Cured	No Treatment	8	.42857	Not Cured
10		Not Cured	No Treatment	7	.42857	Not Cured
11		Cured	Intervention	7	.71930	Cured
12		Cured	No Treatment	7	.42857	Not Cured
13		Cured	No Treatment	5	.42857	Not Cured
14		Not Cured	Intervention	9	.71930	Cured
15		Not Cured	No Treatment	6	.42857	Not Cured
Total	N	15	15	15	15	15

a. Limited to first 15 cases.