Politecnico di Bari

Logistic Regression

Classification

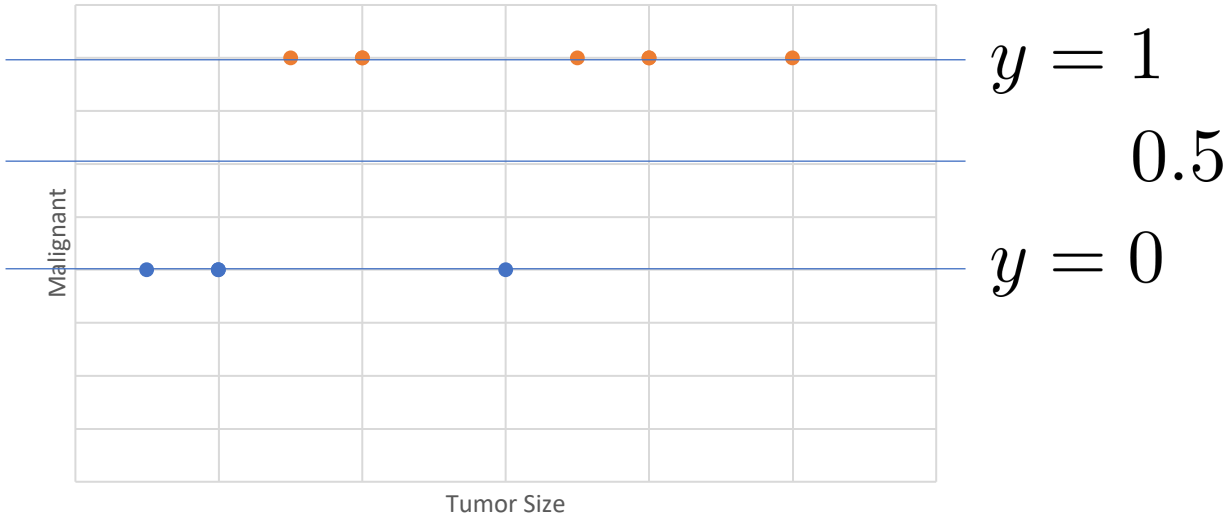
Classification problem

- Email: Spam/NotSpam?
- Online Transactions: Fraudulent (Yes/No)?
- Tumor: Malignant/Benign?

$y \in \{0, 1\}$ 0:“Negative Class”(e.g.,benign tumor)
1:“Positive Class”(e.g.,malignant tumor)

$y \in \{0, 1, 2, 3, 4, 5\}$ Multiple classes

Classifier threshold



Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict «y=1»

If $h_\theta(x) \leq 0.5$, predict «y=0»

Bounding output

Classification: y = 0 or 1

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

In the Logistic regression we want the hypothesis output to be bound: $0 \leq h_\theta(x) \leq 1$

Politecnico di Bari

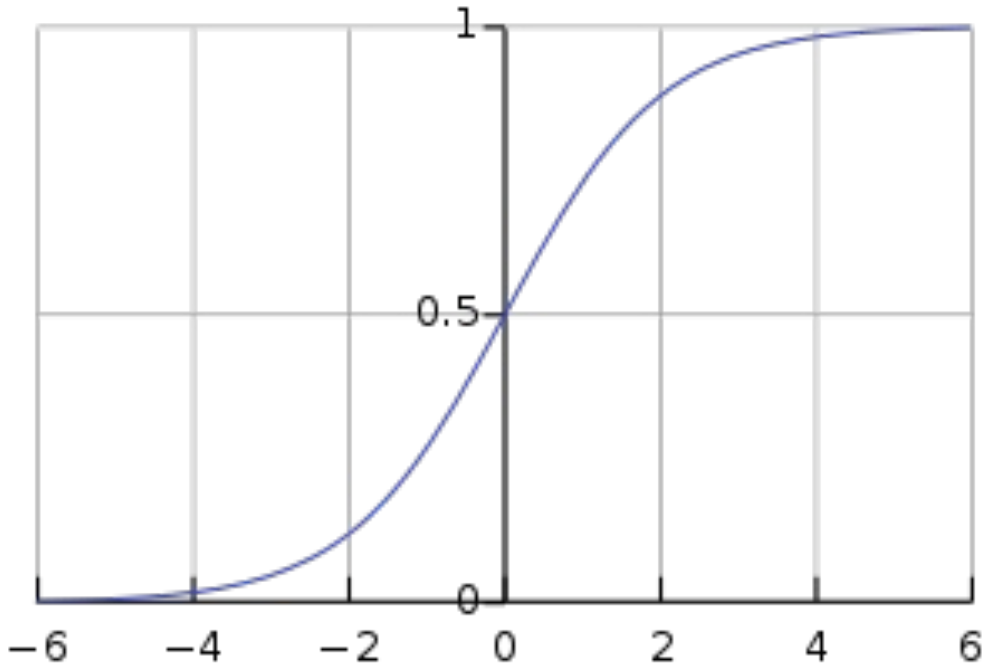
Logistic Regression

Hypothesis Representation

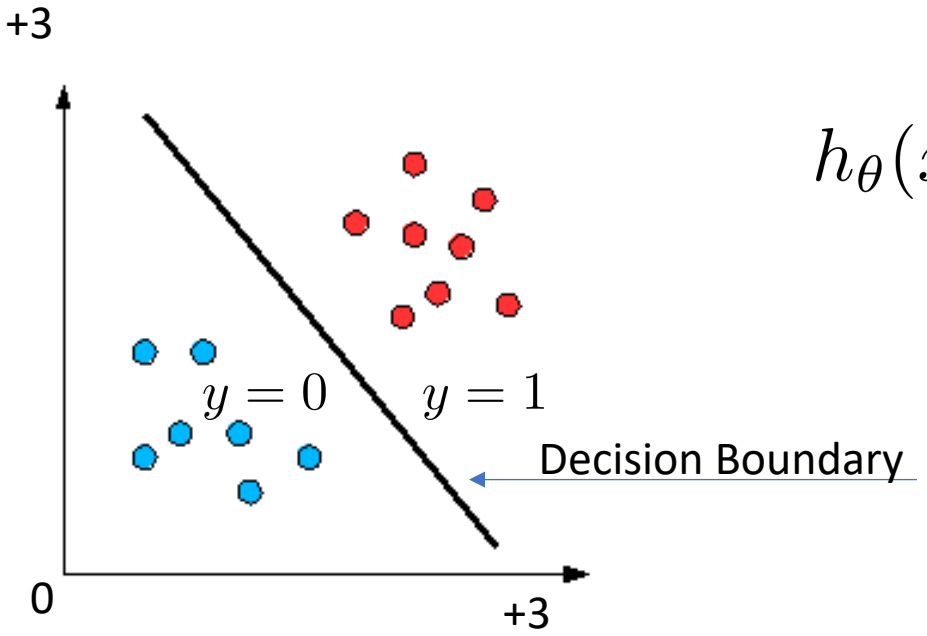
Logistic regression model

Sigmoid (Logistic) function as hypothesis:

$$
\begin{aligned}
h_\theta(x) &= \theta^T x \\
&= g(\theta^T x) \\
&= g(z) \\
&= \frac{1}{1+e^{-z}} \\
h_\theta(x) &= \frac{1}{1+e^{-\theta^T x}}
\end{aligned}
$$



Decision boundary intuition

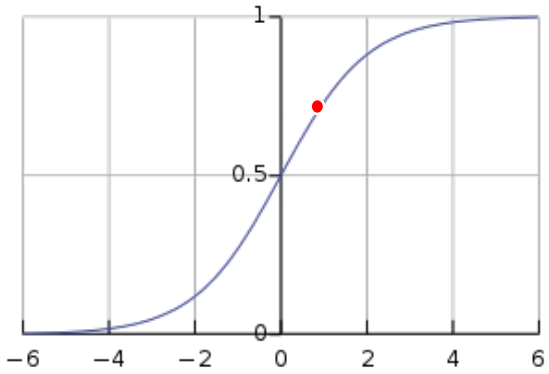


$$h_\theta(x) = g(\underbrace{\theta_0}_{-3} + \underbrace{\theta_1}_{+1} x_1 + \underbrace{\theta_2}_{+1} x_2)$$

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

Probabilistic Interpretation

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



If $h_\theta(x) = 0.7$, it is estimated a probability of 0.7 that the tumor is malignant. It represents the probability that y is 1, given x, parametrized by theta.

$$h_\theta(x) = p(y = 1|x; \theta)$$

$$y = 0 \quad or \quad 1$$

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

$$p(y = 0|x; \theta) = 1 - p(y = 1|x; \theta)$$

Cost Function

As before, we want that solving the Cost function minimization problem would be equal to solve the determine the hypothesis using the maximum likelihood criterion.

We can recap the formula, under the same assumptions (the training samples are independent and with identical distribution). In this case we have only two possible outputs, 0 and 1, so we are facing a Bernoulli distribution.

$$L(\theta) = L(\theta; X; \mathbf{y}) = p(\mathbf{y}|X;\theta) = \prod_{i=1}^{m} p(y^{(i)}|\mathbf{x}^{(\mathbf{i})};\theta). \quad (1)$$

Cost Function (cont.d)

As seen before

$$\begin{cases} p(y^{(i)} = 1|\mathbf{x^{(i)}}; \theta) = h_\theta(\mathbf{x^{(i)}}) \\ p(y^{(i)} = 0|\mathbf{x^{(i)}}; \theta) = 1 - h_\theta(\mathbf{x^{(i)}}) \end{cases}$$

And we can express it in a more compact form:

$$p(y^{(i)}|\mathbf{x^{(i)}}; \theta) = h_\theta(\mathbf{x^{(i)}})^{y^{(i)}} \cdot (1 - h_\theta(\mathbf{x^{(i)}}))^{1-y^{(i)}}.$$

We can substitute the latter in (1):

$$L(\theta) = \prod_{i=1}^{m} h_\theta(\mathbf{x^{(i)}})^{y^{(i)}} \cdot (1 - h_\theta(\mathbf{x^{(i)}}))^{1-y^{(i)}}$$

Cost Function (cont.d)

Now we can move to logarithimic form:

$$l(\theta) = \log \prod_{i=1}^{m} h_\theta(\mathbf{x^{(i)}})^{y^{(i)}} \cdot (1 - h_\theta(\mathbf{x^{(i)}}))^{1-y^{(i)}}$$

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

We can move from maximization to the minimization problem

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

Cross Entropy Error Function

Finally we can set the whole problem as a paremeters fitting problem

$$\hat{\theta} = argmin_\theta J(\theta).$$

Errors

The logistic function cannot be easily studied analytically, but we can get an intuition using the error contribution:

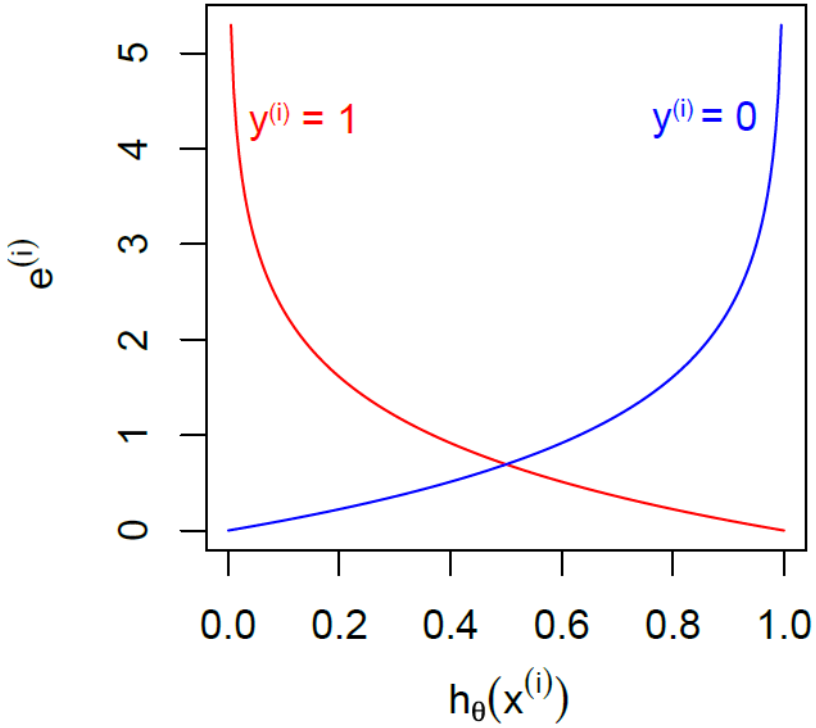
$$e^{(i)} = -y^{(i)} \log h_\theta(\mathbf{x^{(i)}}) - (1 - y^{(i)}) \log (1 - h_\theta(\mathbf{x^{(i)}}))$$

If $y^{(i)} = 1$ the error has value $e^{(i)} = -\log h_\theta(\mathbf{x^{(i)}})$

- if h(x) is low the error is high: $h_\theta(\mathbf{x^{(i)}}) = 0 \Rightarrow e^{(i)} \to \infty$
- if h(x) is high the error is low: $h_\theta(\mathbf{x^{(i)}}) = 1 \Rightarrow e^{(i)} \to 0$

If $y^{(i)} = 0$ the error has value $e^{(i)} = -\log (1 - h_\theta(\mathbf{x^{(i)}}))$

- if h(x) is low the error is low: $h_\theta(\mathbf{x^{(i)}}) = 0 \Rightarrow e^{(i)} \to 0$
- if h(x) is high the error is high: $h_\theta(\mathbf{x^{(i)}}) = 1 \Rightarrow e^{(i)} \to \infty$



Politecnico di Bari

Logistic Regression

Gradient descent

Gradient descent

Goal: $\hat{\theta} = argmin_{\theta} J(\theta).$

Weights update: $\theta_k = \theta_k - \alpha \frac{\partial J(\theta)}{\partial \theta_k}$

We can derive the Logistic function

$$g'(z) = \frac{\partial}{\partial z}\left(\frac{1}{1+e^{-z}}\right) = \frac{e^{-z}}{(1+e^{-z})^2}$$

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

Weight update derivation

From $g'(z) = g(z)(1 - g(z))$

We can compute the derivative of $h_\theta(x)$

$$\frac{\partial h_\theta(\mathbf{x})}{\partial \theta_k} = h_\theta(\mathbf{x})(1 - h_\theta(\mathbf{x}))\frac{\partial(\theta^T \mathbf{x})}{\partial \theta_k}$$

In linear regression gradient descent we saw that $\frac{\partial(\theta^T \mathbf{x})}{\partial \theta_k} = x_k$

Hence $\frac{\partial h_\theta(\mathbf{x})}{\partial \theta_k} = h_\theta(\mathbf{x})(1 - h_\theta(\mathbf{x}))x_k$

Weight update derivation (cont.d)

We can compute the the partial derivative of

$$\frac{\partial J(\theta)}{\partial \theta_k} = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\underbrace{\frac{\partial}{\partial \theta_k}(\log h_\theta(\mathbf{x^{(i)}}))} + (1-y^{(i)})\underbrace{\frac{\partial}{\partial \theta_k}(\log(1-h_\theta(\mathbf{x^{(i)}})))}\right) \quad (1)$$

We can deal with the two derivative separately

$$\frac{\partial}{\partial \theta_k}\log h_\theta(\mathbf{x^{(i)}}) = \frac{1}{h_\theta(\mathbf{x^{(i)}})}h_\theta(\mathbf{x^{(i)}})(1-h_\theta(\mathbf{x^{(i)}}))x_k^{(i)} = (1-h_\theta(\mathbf{x^{(i)}}))x_k^{(i)}$$

$$\frac{\partial}{\partial \theta_k}\log(1-h_\theta(\mathbf{x^{(i)}})) = \frac{1}{1-h_\theta(\mathbf{x^{(i)}})}(-h_\theta(\mathbf{x^{(i)}}))(1-h_\theta(\mathbf{x^{(i)}}))x_k^{(i)} = -h_\theta(\mathbf{x^{(i)}})x_k^{(i)}$$

We substitute the latters in (1)

Weight update derivation (cont.d)

And finally we reach the last derivative step:

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta_k} &= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}(1-h_\theta(\mathbf{x^{(i)}}))x_k^{(i)} + (1-y^{(i)})(-h_\theta(\mathbf{x^{(i)}}))x_k^{(i)}\right) \\
&= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}(1-h_\theta(\mathbf{x^{(i)}})) + (1-y^{(i)})(-h_\theta(\mathbf{x^{(i)}}))\right)x_k^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(\mathbf{x^{(i)}}) - y^{(i)}\right)x_k^{(i)}
\end{aligned}$$

Politecnico di Bari

Logistic Regression

Multi-class classification
One vs all

Multi-class classification

Email tagging: Work, Friends, Family, Hobby

Medical diagrams: Not ill, Cold, Flu

Weather: Sunny, Cloudy, Rain, Snow

Multi-class classification

Email tagging: Work, Friends, Family, Hobby

$y = [\ 1\ ,\ 2\ ,\ 3\ ,\ 4\ ]$

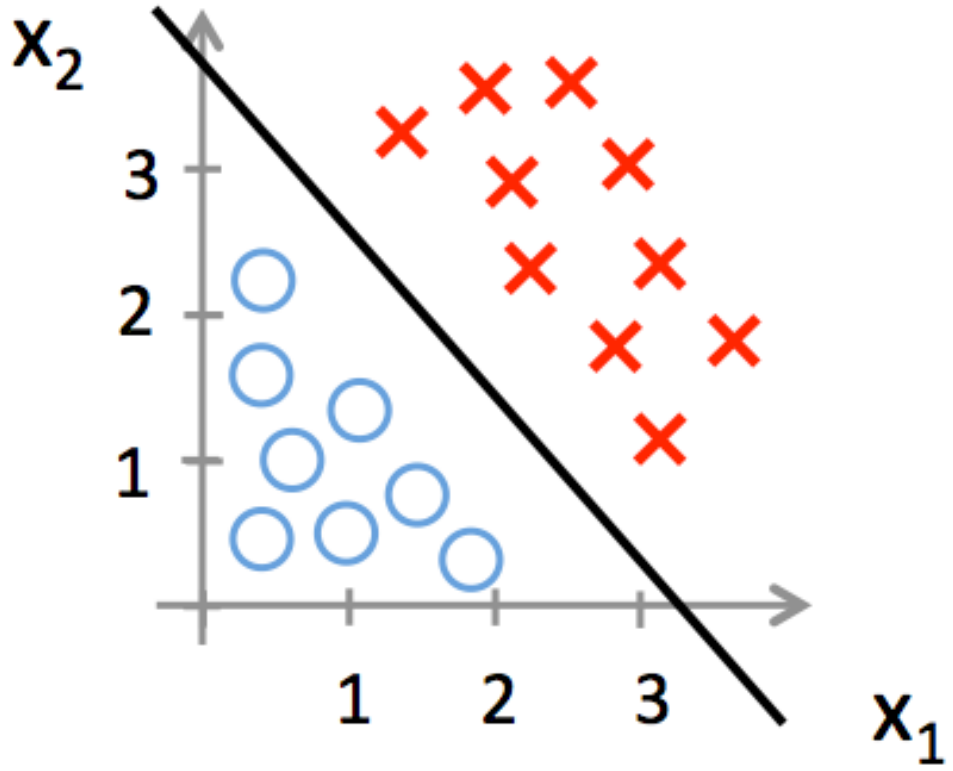
Medical diagrams: Not ill, Cold, Flu

$y = [\ 1\ ,\ 2\ ,\ 3\ ]$

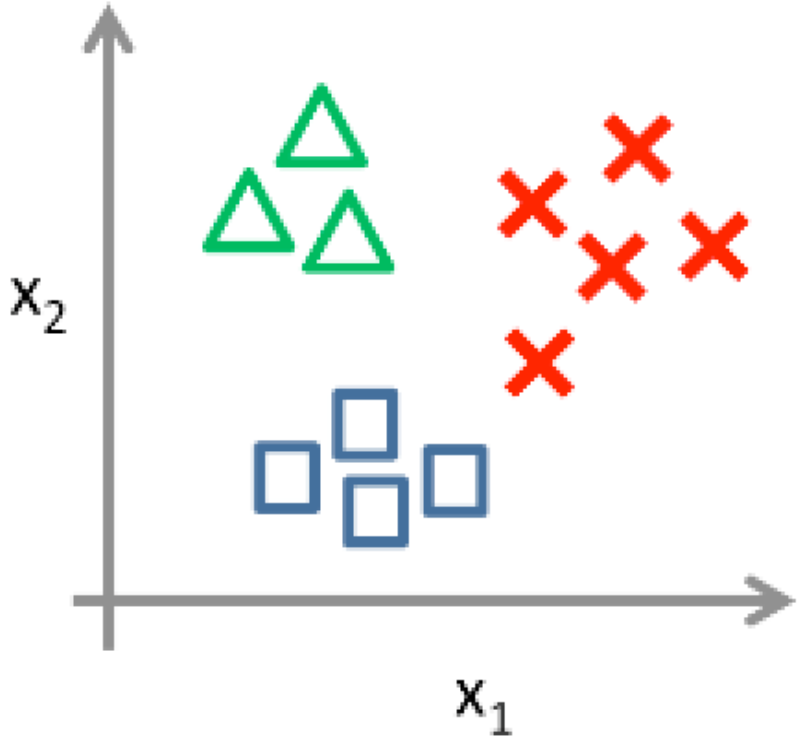
Weather: Sunny, Cloudy, Rain, Snow

$y = [\ 1\ ,\ 2\ ,\ 3\ ,\ 4\ ]$

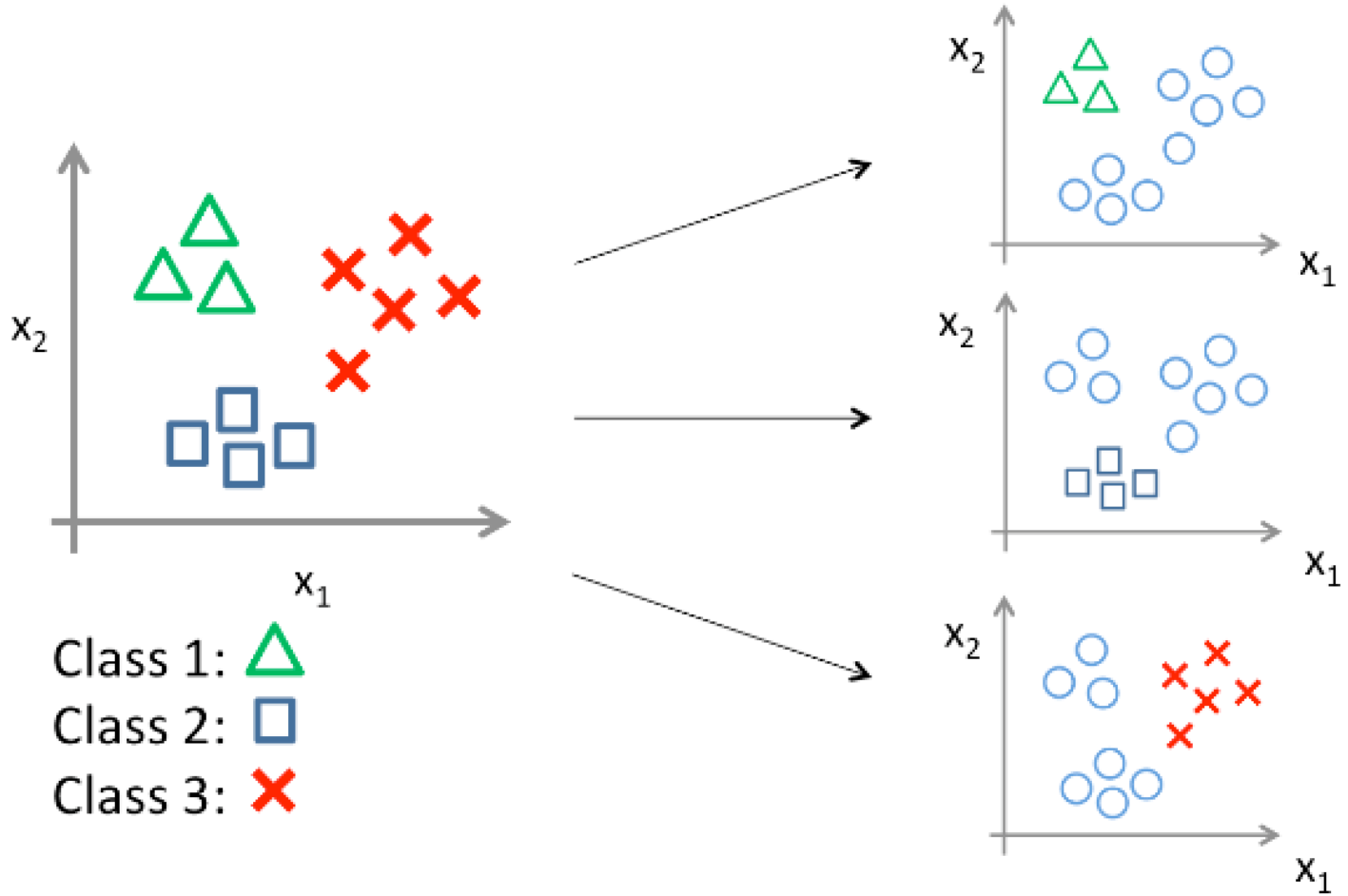
Binary classification



Multi-class classification



One vs all



$$h_\theta^{(i)}(x) = p(y = i|x; \theta) \quad (i = 1, 2, 3)$$