COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

Multivariate Linear Regression

-- Vectorisation

Dr SHI Lei

Durham University

learning lab

- Vectorisation of Hypothesis Function
- Vectorisation of Cost Function
- Vectorisation of Gradient Descent

Vectorisation of Hypothesis Function

Vectorisation of Hypothesis Function

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \cdots + \theta_n \cdot x_n$$

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 \cdot 1 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \cdots + \theta_n \cdot x_n$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \cdots \\ \theta_n \end{bmatrix} \in \mathbb{R}^{n+1} \qquad \boldsymbol{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n+1}$$

parameter vector

variable vector

Vectorisation of Hypothesis Function

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \cdots + \theta_n \cdot x_n$$

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 \cdot x_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \cdots + \theta_n \cdot x_n \qquad (x_0 = 1, constant)$$

$$= [\theta_0, \theta_1, \theta_2, \dots, \theta_n] \cdot \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$

$= \boldsymbol{\theta}^T \boldsymbol{x}$ for single instance

Vectorisation of Hypothesis Function for the whole Training Set

EXAMPLE. Happiness predictor

features				label
income (k)	age	# of children	tea (cups/week)	happiness
38.63	46	1	7	2.314
49.79	37	0	15	3.433
40.34	52	3	20	4.03
21.18	25	0	3	1.45
...	...	...	...	...

EXAMPLE. Happiness predictor

x_1	x_2	x_3	x_4	y
income (k)	age	# of children	tea (cups/week)	happiness
38.63	46	1	7	2.314
49.79	37	0	15	3.433
40.34	52	3	20	4.03
21.18	25	0	3	1.45
...	...	...	...	...

EXAMPLE. Happiness predictor

x_1	x_2	x_3	x_4	...	x_n	y
income (k)	age	# of children	tea (cups/week)	...	the n^{th} feature	happiness
38.63	46	1	7	...	*	2.314
49.79	37	0	15	...	*	3.433
40.34	52	3	20	...	*	4.03
21.18	25	0	3	...	*	1.45
...	...	...	...	...	...	...

EXAMPLE. Happiness predictor

		income (k)	age	# of children	tea (cups/week)	...	the n^{th} feature	happiness	
$(x^{(1)})^T$	1	38.63	46	1	7	...	*	2.314	$y^{(1)}$
$(x^{(2)})^T$	1	49.79	37	0	15	...	*	3.433	$y^{(2)}$
$(x^{(3)})^T$	1	40.34	52	3	20	...	*	4.03	$y^{(3)}$
$(x^{(4)})^T$	1	21.18	25	0	3	...	*	1.45	$y^{(4)}$
...	...	...	...	...	...	...	...	...	...
$(x^{(m)})^T$	1	*	*	*	*	...	*	1.45	$y^{(m)}$

EXAMPLE. Happiness predictor

$$
\begin{array}{cccccccc}
 & x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} & \cdots & x_n^{(2)} \\
(x^{(2)})^T = & [\ 1 & 49.79 & 37 & 0 & 15 & \ldots & *\ ]
\end{array}
$$

EXAMPLE. Happiness predictor

$$x^{(2)} = \begin{array}{cc} \begin{bmatrix} 1 \\ 49.79 \\ 37 \\ 0 \\ 15 \\ \cdots \\ * \end{bmatrix} & \begin{array}{c} x_0^{(2)} \\ x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \\ x_4^{(2)} \\ \cdots \\ x_n^{(2)} \end{array} \end{array}$$

$$x^{(i)} = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \\ x_4^{(i)} \\ \cdots \\ x_n^{(i)} \end{bmatrix}$$

For the i-th instance

Vectorisation of Hypothesis Function

$$\boldsymbol{X} = \begin{bmatrix} (\boldsymbol{x}^{(1)})^T \\ (\boldsymbol{x}^{(2)})^T \\ (\boldsymbol{x}^{(3)})^T \\ (\boldsymbol{x}^{(4)})^T \\ \dots \\ (\boldsymbol{x}^{(m)})^T \end{bmatrix}$$

Vectorisation of Hypothesis Function

$$\boldsymbol{X} = \begin{bmatrix} (\boldsymbol{x}^{(1)})^T \\ (\boldsymbol{x}^{(2)})^T \\ (\boldsymbol{x}^{(3)})^T \\ (\boldsymbol{x}^{(4)})^T \\ \dots \\ (\boldsymbol{x}^{(m)})^T \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_j^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_j^{(2)} & \cdots & x_n^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} & \cdots & x_j^{(3)} & \cdots & x_n^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_1^{(i)} & x_2^{(i)} & \cdots & x_j^{(i)} & \cdots & x_n^{(i)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_j \\ \vdots \\ \theta_n \end{bmatrix}$$

Vectorisation of Hypothesis Function

$$h_{\boldsymbol{\theta}}\left(x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)}\right) = h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(1)})$$

$$\boldsymbol{X\theta} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_j^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_j^{(2)} & \cdots & x_n^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} & \cdots & x_j^{(3)} & \cdots & x_n^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_1^{(i)} & x_2^{(i)} & \cdots & x_j^{(i)} & \cdots & x_n^{(i)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \cdot \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_j \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} \boxed{\theta_0 + \theta_1 \cdot x_1^{(1)} + \theta_2 \cdot x_2^{(1)} + \cdots + \theta_n \cdot x_n^{(1)}} \\ \theta_0 + \theta_1 \cdot x_1^{(2)} + \theta_2 \cdot x_2^{(2)} + \cdots + \theta_n \cdot x_n^{(2)} \\ \theta_0 + \theta_1 \cdot x_1^{(3)} + \theta_2 \cdot x_2^{(3)} + \cdots + \theta_n \cdot x_n^{(3)} \\ \cdots \\ \theta_0 + \theta_1 \cdot x_1^{(i)} + \theta_2 \cdot x_2^{(i)} + \cdots + \theta_n \cdot x_n^{(i)} \\ \cdots \\ \theta_0 + \theta_1 \cdot x_1^{(m)} + \theta_2 \cdot x_2^{(m)} + \cdots + \theta_n \cdot x_n^{(m)} \end{bmatrix}$$

$\mathbb{R}^{m\times(n+1)}$ $\quad$ $\mathbb{R}^{(n+1)\times 1}$ $\quad$ $\mathbb{R}^{m\times 1}$

Vectorisation of Hypothesis Function

$$
\boldsymbol{X\theta} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_j^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_j^{(2)} & \cdots & x_n^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} & \cdots & x_j^{(3)} & \cdots & x_n^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_1^{(i)} & x_2^{(i)} & \cdots & x_j^{(i)} & \cdots & x_n^{(i)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \cdot \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_j \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(1)}) \\ h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(2)}) \\ h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(3)}) \\ \cdots \\ h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) \\ \cdots \\ h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(m)}) \end{bmatrix}
$$

$\mathbb{R}^{m\times(n+1)}$ $\qquad$ $\mathbb{R}^{(n+1)\times 1}$ $\qquad$ $\mathbb{R}^{m\times 1}$

Vectorisation of Hypothesis Function

$$
X\theta = \begin{bmatrix}
1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_j^{(1)} & \cdots & x_n^{(1)} \\
1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_j^{(2)} & \cdots & x_n^{(2)} \\
1 & x_1^{(3)} & x_2^{(3)} & \cdots & x_j^{(3)} & \cdots & x_n^{(3)} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
1 & x_1^{(i)} & x_2^{(i)} & \cdots & x_j^{(i)} & \cdots & x_n^{(i)} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_2^{(m)} & \cdots & x_n^{(m)}
\end{bmatrix} \cdot \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_j \\ \vdots \\ \theta_n \end{bmatrix} = \boxed{\begin{bmatrix} h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(1)}) \\ h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(2)}) \\ h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(3)}) \\ \dots \\ h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) \\ \dots \\ h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(m)}) \end{bmatrix}}
$$

Vectorised Hypothesis Function

$$\longrightarrow \quad \boldsymbol{h_{\theta}(x) = X\,\theta}$$

vector / matrix

A vector containing the results of the hypothesis function, i.e., the predicted value, for all the instances from the training set.

Vectorisation of Hypothesis Function

$h_\theta(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x}$ vs $\boldsymbol{h_\theta(x) = X\,\theta}$

takes single instance

outputs predicted label

takes all instances

outputs a vector of predicted labels

Vectorisation of Cost Function

Vectorisation of Cost Function

$$\boldsymbol{X\theta} - \boldsymbol{y} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} & \cdots & x_n^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \cdot \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \begin{bmatrix} h_\theta\left(x^{(1)}\right) - y^{(1)} \\ h_\theta\left(x^{(2)}\right) - y^{(2)} \\ h_\theta\left(x^{(3)}\right) - y^{(3)} \\ \vdots \\ h_\theta\left(x^{(m)}\right) - y^{(m)} \end{bmatrix}$$

Vectorisation of Cost Function

$$\boldsymbol{X\theta} - \boldsymbol{y} = \begin{bmatrix} h_\theta\left(x^{(1)}\right) - y^{(1)} \\ h_\theta\left(x^{(2)}\right) - y^{(2)} \\ h_\theta\left(x^{(3)}\right) - y^{(3)} \\ \vdots \\ h_\theta\left(x^{(m)}\right) - y^{(m)} \end{bmatrix}$$

Vectorisation of Cost Function

$$(\boldsymbol{X\theta} - \boldsymbol{y})^T \cdot (\boldsymbol{X\theta} - \boldsymbol{y})$$

$$= \left[h_\theta\left(x^{(1)}\right) - y^{(1)}, h_\theta\left(x^{(2)}\right) - y^{(2)}, h_\theta\left(x^{(3)}\right) - y^{(3)}, \ldots, h_\theta\left(x^{(m)}\right) - y^{(m)}\right] \cdot \begin{bmatrix} h_\theta\left(x^{(1)}\right) - y^{(1)} \\ h_\theta\left(x^{(2)}\right) - y^{(2)} \\ h_\theta\left(x^{(3)}\right) - y^{(3)} \\ \vdots \\ h_\theta\left(x^{(m)}\right) - y^{(m)} \end{bmatrix}$$

$$= (h_\theta(x^{(1)}) - y^{(1)})^2 + (h_\theta(x^{(2)}) - y^{(2)})^2 + (h_\theta(x^{(3)}) - y^{(3)})^2 + \ldots + (h_\theta(x^{(m)}) - y^{(m)})^2$$

$$= \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Vectorisation of Cost Function

$$(\boldsymbol{X\theta} - \boldsymbol{y})^T \cdot (\boldsymbol{X\theta} - \boldsymbol{y}) \rightleftarrows \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Vectorisation of Cost Function

$$\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 = (\boldsymbol{X\theta} - \boldsymbol{y})^T(\boldsymbol{X\theta} - \boldsymbol{y})$$

Vectorisation of Cost Function

$$\boxed{\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2} = \frac{1}{2m}\left(\boldsymbol{X\theta}-\boldsymbol{y}\right)^T\left(\boldsymbol{X\theta}-\boldsymbol{y}\right)$$

Vectorisation of Cost Function

$$J(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 = \frac{1}{2m}\left(\boldsymbol{X\theta} - \boldsymbol{y}\right)^T\left(\boldsymbol{X\theta} - \boldsymbol{y}\right)$$

Vectorisation of Cost Function

$$J(\boldsymbol{\theta}) = \frac{1}{2m}\,(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})^T\,(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})$$

Vectorisation of Gradient Descent

Vectorisation of Gradient Descent

Repeat until convergence {

$$\theta_0 := \theta_0 - \alpha \frac{\partial J}{\partial \theta_0}(\boldsymbol{\theta})$$

$$\theta_1 := \theta_1 - \alpha \frac{\partial J}{\partial \theta_1}(\boldsymbol{\theta})$$

...

$$\theta_n := \theta_1 - \alpha \frac{\partial J}{\partial \theta_n}(\boldsymbol{\theta})$$

}

$$\boxed{\frac{\partial J}{\partial \theta_k}(\boldsymbol{\theta})} = \frac{1}{m}\sum_{i=1}^{m} \boxed{\left(h_\theta(x)^{(i)} - y^{(i)}\right)}\boxed{x_k^{(i)}}$$

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

$$= \frac{1}{m}\boldsymbol{x}_k^T(\boldsymbol{X\theta} - \boldsymbol{y})$$

Vectorisation of Gradient Descent

Repeat until convergence {

$$\theta_0 := \theta_0 - \alpha \boxed{\frac{\partial J}{\partial \theta_0}(\boldsymbol{\theta})}$$

$$\theta_1 := \theta_1 - \alpha \boxed{\frac{\partial J}{\partial \theta_1}(\boldsymbol{\theta})}$$

...

$$\theta_n := \theta_1 - \alpha \boxed{\frac{\partial J}{\partial \theta_n}(\boldsymbol{\theta})}$$

}

$$\boxed{\frac{\partial J}{\partial \theta_k}(\boldsymbol{\theta})} = \boxed{\frac{1}{m}\boldsymbol{x}_k^T(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})}$$

Vectorisation of Gradient Descent

Repeat until convergence {

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \boldsymbol{x}_0^T (\boldsymbol{X\theta} - \boldsymbol{y})$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \boldsymbol{x}_1^T (\boldsymbol{X\theta} - \boldsymbol{y})$$

$$\vdots$$

$$\theta_n := \theta_1 - \alpha \frac{1}{m} \boldsymbol{x}_n^T (\boldsymbol{X\theta} - \boldsymbol{y})$$

}

Vectorisation of Gradient Descent

Repeat until convergence {

$$\boldsymbol{\theta} := \boldsymbol{\theta} - \alpha \frac{1}{m} \boldsymbol{X}^T (\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})$$

}

- Vectorisation of Hypothesis Function

$$h_\theta(x) = \boldsymbol{\theta}^T \boldsymbol{x} \qquad \boldsymbol{h_\theta}(\boldsymbol{x}) = \boldsymbol{X}\,\boldsymbol{\theta}$$

- Vectorisation of Cost Function

$$J(\boldsymbol{\theta}) = \frac{1}{2m}\,(\boldsymbol{X\theta} - \boldsymbol{y})^T\,(\boldsymbol{X\theta} - \boldsymbol{y})$$

- Vectorisation of Gradient Descent

Repeat until convergence {

$$\boldsymbol{\theta} := \boldsymbol{\theta} - \alpha\,\frac{1}{m}\boldsymbol{X}^T\,(\boldsymbol{X\theta} - \boldsymbol{y})$$

}

Comparing with un-vectorised

- ✓ Easier to implement
- ✓ Computationally more efficient