Regularisation notes.

Robert K.

December 27, 2016

Abstract

This is just a document with notes regarding Regularisation in finding Θ values to prevent *overfitting*.

1 Underfit vs Overfit

Underfit - (a.k.a "high bias") is the prediction (learned hypothesis function) which does not fit the training or new input data very well. Too simplified hypothesis function.
Overfit - (a.k.a "high variance") is the prediction (learned hypothesis function) which fits well the training data, but fails to generalise to new input data. It can occur when we have too many *features* and small size of training set.

How to address overfitting problem:

1. Reduce the number of features, which can be done in two ways:
   - Manually select which features to keep
   - Use Model Selection Algorithm to automatically select features to keep
2. Use **regularisation** where we keep all features, but with different prioryty

2 Regularisation

Regularisation - is the technique of reducing overfitting by keeping all features but reducing the values of θ_j.
It works well wen we have a lot of features and each of them contributes a bit to predicting y.

Properties of regularisation:

- if θ_j is small then more likely we will not overfit. We want to keep θ_j small.

2.1 Regularisation of linear regression

2.1.1 Cost function and gradient

To use Regularisation we slightly modify cost function $J(\theta)$

$$J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \right] = \tag{1}$$

$$= \left[\frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \right] + \left[\frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2 \right]$$

If the θ_j is small, then the cost function $J(\theta)$ will be small too.
λ - is the **regularisation parameter** It determines how much the costs of our theta parameters are inflated.

Properties of new cost function:

- cost will be small if θ_j is small.
- cost will be small for the big **regularisation parameter** λ if θ_j will be super small
- If λ is chosen to be too large, then $\theta_1 ... \theta_n$ will be near zero to make cost small. It may cause underfitting.
- if $\lambda = 0$ or is too small then it is similar to case of not using regularisation at all - so we increase a chance of overfitting.

Gradient of regularised cost function $\frac{\delta}{\delta \theta_j} J(\theta)$

$$\frac{\delta}{\delta \theta_j} J(\theta) = \frac{\delta}{\delta \theta_j} \left[\frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \right] + \frac{\delta}{\delta \theta_j} \left[\frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2 \right] \tag{2}$$

where:

$$h_\theta(x^{(i)}) = \theta^T x^{(i)} = \sum_{j=0}^{n} \theta_j x_j^{(i)} \text{ with } x_0^{(i)} = 1$$

so gradient of cost function for linear regresion:

$$\frac{\delta}{\delta \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \text{ for } j = 0 \tag{3}$$

$$\frac{\delta}{\delta \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} \theta_j \text{ for } j \geq 0 \tag{4}$$

The **vectorised form of cost function**

$$J(\Theta) = J\left(\begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix}\right) = \frac{1}{2m}(X\Theta - y)^T(X\Theta - y) + \frac{\lambda}{2m}\begin{bmatrix}0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix}^T \begin{bmatrix}0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix} \tag{5}$$

The **vectorised form of cost function derivative**

$$\nabla J(\Theta) = \begin{bmatrix}\frac{\delta}{\delta\Theta_0}J(\Theta)\\ \frac{\delta}{\delta\Theta_1}J(\Theta)\\ \vdots\\ \frac{\delta}{\delta\Theta_n}J(\Theta)\end{bmatrix} = \frac{1}{m}\left[X^T(X\Theta - y)\right] + \frac{\lambda}{m}\begin{bmatrix}0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix} \tag{6}$$

2.1.2 Regularised normal equations

Regularisation using the non-iterative normal equation.

$$\Theta = \begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix} = \theta = \left(X^TX + \lambda\begin{bmatrix}0 & 0 & 0 & \dots & 0\\ 0 & 1 & 0 & \dots & 0\\ 0 & 0 & 1 & \dots & 0\\ \vdots & \vdots & \vdots & \ddots & 0\\ 0 & 0 & 0 & \dots & 1\end{bmatrix}\right)^{-1} X^Ty \tag{7}$$

2.2 Regularisation of logistic regression

2.2.1 Cost function and gradient

Regularised **cost function** has a form

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}log(h_\theta(x^{(i)})) + (1-y^{(i)})log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2 \tag{8}$$

where:

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

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

Derivative of the cost function for logistic regression looks like that:

$$\frac{\delta}{\delta\theta_j}J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} \text{ for } j = 0 \tag{9}$$

$$\frac{\delta}{\delta\theta_j}J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} + \frac{\lambda}{m}\theta_j \text{ for } j \geq 0 \tag{10}$$

where:

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

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

The **vectorised form of cost function**

$$J(\Theta) = J\left(\begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix}\right) = -\frac{1}{m}\left[y^T log(g(X\Theta)) + (1-y)^T log(1-g(X\Theta))\right] + \frac{\lambda}{2m}\begin{bmatrix}0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix}^T \begin{bmatrix}0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix} \tag{11}$$

where:

$$g(X\Theta) = \begin{bmatrix}g(\Theta^T x^{(1)})\\ g(\Theta^T x^{(2)})\\ \vdots\\ g(\Theta^T x^{(m)})\end{bmatrix}$$

$log(g(X\Theta))$ - is a column vector of size m. *Log* function is done on every element of $g(X\Theta)$ column vector.

The **vectorised form of cost function derivative**

$$\nabla J(\Theta) = \begin{bmatrix}\frac{\delta}{\delta\Theta_0}J(\Theta)\\ \frac{\delta}{\delta\Theta_1}J(\Theta)\\ \vdots\\ \frac{\delta}{\delta\Theta_n}J(\Theta)\end{bmatrix} = \frac{1}{m}\left[X^T(g(X\Theta) - y)\right] + \frac{\lambda}{m}\begin{bmatrix}0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix} \tag{12}$$

where:

$$g(X\Theta) = \begin{bmatrix}g(\Theta^T x^{(1)})\\ g(\Theta^T x^{(2)})\\ \vdots\\ g(\Theta^T x^{(m)})\end{bmatrix}$$