Logistic regression

Kristian Wichmann

August 18, 2017

1 Definitions and setup

1.1 Logistic model

A logistic model on $\mathbb{R}^n$ is a function:

$$x \mapsto a = \sigma(w^t x + b) = \sigma(z) \tag{1.1}$$

Here, σ is the logistic sigma function, $w \in \mathbb{R}^n$ is the weight vector, and $b \in \mathbb{R}$ the bias. The result a is usually interpreted as the probability of a given condition being true; it is a binary classification model.

1.2 Training set

The training set consists of m labelled data points. I.e. we have m points in $\mathbb{R}^n$ along with a labelling of whether the condition is question is actually met for the data point, one-hot encoded. So m pairs $(x^{(i)}, y^{(i)})$. Corresponding a's and z's are defined through:

$$a^{(i)} = \sigma(w^t x^{(i)} + b) = \sigma(z^{(i)}) \tag{1.2}$$

1.3 Cost and loss functions

To train the model, we need to specify a function to optimize. Here, for a single data point $x^{(i)}$ we will use the cross-entropy:

$$C^{(i)} = -\left[y^{(i)} \log a^{(i)} + (1 - y^{(i)}) \log(1 - a^{(i)})\right] \tag{1.3}$$

This is the cost function. The loss function is the average of the cost for the entire training set:

$$J = \frac{1}{m} \sum_{i=1}^{m} C^{(i)} \tag{1.4}$$

We wish to find the values of w and b which minimizes J.

2 Finding derivatives

To minimize, we seek the derivatives:

$$\frac{\partial J}{\partial w}, \quad \frac{\partial J}{\partial b} \tag{2.1}$$

Both can be found using the chain rule:

$$\frac{\partial J}{\partial w} = \sum_{i=1}^{m} \frac{\partial J}{\partial a^{(i)}} \frac{\partial a^{(i)}}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial w}, \quad \frac{\partial J}{\partial b} = \sum_{i=1}^{m} \frac{\partial J}{\partial a^{(i)}} \frac{\partial a^{(i)}}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial b} \tag{2.2}$$

The first two terms in each formula are the same. The first one:

$$\frac{\partial J}{\partial a^{(i)}} = -\frac{1}{m}\left[\frac{y^{(i)}}{a^{(i)}} - \frac{1-y^{(i)}}{1-a^{(i)}}\right] = \tag{2.3}$$

$$-\frac{1}{m}\frac{y^{(i)}(1-a^{(i)}) - (1-y^{(i)})a^{(i)}}{a^{(i)}(1-a^{(i)})} = \tag{2.4}$$

$$\frac{1}{m}\frac{a^{(i)} - y^{(i)}}{a^{(i)}(1-a^{(i)})} \tag{2.5}$$

Here we've used that we only get a non-zero result when the index matches. The second comes from a standard result for the logistic sigmoid:

$$\frac{\partial a^{(i)}}{\partial z^{(i)}} = a^{(i)}(1-a^{(i)}) \tag{2.6}$$

So when combined, the denominator cancels:

$$\frac{\partial J}{\partial a^{(i)}} \frac{\partial a^{(i)}}{\partial z^{(i)}} = \frac{1}{m} a^{(i)} - y^{(i)} = \frac{1}{m}\delta^{(i)} \tag{2.7}$$

Here we've introduced the output error $\delta^{(i)} = a^{(i)} - y^{(i)}$.

2.1 Derivative for w

In this case the final term is:

$$\frac{\partial z^{(i)}}{\partial w} = \frac{\partial}{\partial w}\left(w^t x^{(i)} + b\right) = x^{(i)} \tag{2.8}$$

So the derivative is:

$$\frac{\partial J}{\partial w} = \frac{1}{m}\sum_{i=1}^{m} \delta^{(i)} x^{(i)} \tag{2.9}$$

2.2 Derivative for b

Here the final term is:

$$\frac{\partial z^{(i)}}{\partial b} = \frac{\partial}{\partial b}\left(w^t x^{(i)} + b\right) = 1 \tag{2.10}$$

So the derivative is:

$$\frac{\partial J}{\partial b} = \frac{1}{m}\sum_{i=1}^{m} \delta^{(i)} \tag{2.11}$$

3 Vectoriation

For vectorization purposes, we will collect the data in a matrix X:

$$X = \begin{pmatrix} | & \cdots & | \\ x^{(i)} & \cdots & x^{(m)} \\ | & \cdots & | \end{pmatrix} \in \mathbb{R}^{n\times m} \tag{3.1}$$

The labels are collected into a row vector:

$$Y = \begin{pmatrix} y^{(1)} & \cdots & y^{(m)} \end{pmatrix} \in \mathbb{R}^{1\times m} \tag{3.2}$$

We can now find the z values by matrix multiplication:

$$Z = w^t X \in \mathbb{R}^{1\times m} \tag{3.3}$$

The a's are then found by applying σ elementwise:

$$A = \sigma(Z) \in \mathbb{R}^{1\times m} \tag{3.4}$$

The errors are then:

$$\Delta = A - Y \in \mathbb{R}^{1\times m} \tag{3.5}$$

And finally, we can get the derivatives:

$$\frac{\partial J}{\partial w} = \frac{1}{m} X\Delta^t \in \mathbb{R}^{n\times 1}, \quad \frac{\partial J}{\partial b} = \frac{1}{m} J_m \Delta^t \in \mathbb{R}^{1\times 1} \tag{3.6}$$

Here J_m is the $1 \times m$ row vector of all ones.