Logistic Regression

Chang Lu

1 Model

Let $\mathbf{x} \in \mathbb{R}^{s \times n}$ be the inputs, where n is the number of samples; and s is the dimension of each feature vector. The logistic regression model is:

$$\hat{\mathbf{y}} = \sigma\left(\mathbf{x}^\top \mathbf{w}\right) \tag{1}$$

$$= \frac{1}{1 + e^{-\mathbf{x}^\top \mathbf{w}}}. \tag{2}$$

Here, $\mathbf{w} \in \mathbb{R}^s$ is the weight. For an input sample $\mathbf{x}_i$, let y_i be the ground-truth for $\mathbf{x}_i$. The prediction $0 < \hat{y}_i < 1$ is the probability $p\left(y_i = 1 \mid \mathbf{x}_i\right)$.

2 Objective Function

Typically, the logistic regression is a binary classification. Therefore, we use a binary cross-entropy loss as the objective function. For an input sample $\mathbf{x}_i$, the binary cross-entropy loss $L(\mathbf{x}_i, y_i \mid \mathbf{w})$ is:

$$L(\mathbf{x}_i, y_i \mid \mathbf{w}) = -\left(y_i \log \hat{y}_i + (1 - y_i) \log\left(1 - \hat{y}_i\right)\right) \tag{3}$$

$$= -\left(y_i \log\left(\frac{1}{1 + e^{-\mathbf{x}_i^\top \mathbf{w}}}\right) + (1 - y_i) \log\left(\frac{e^{-\mathbf{x}_i^\top \mathbf{w}}}{1 + e^{-\mathbf{x}_i^\top \mathbf{w}}}\right)\right) \tag{4}$$

$$= \log\left(1 + e^{-\mathbf{x}_i^\top \mathbf{w}}\right) + \mathbf{x}_i^\top \mathbf{w}\left(1 - y_i\right) \tag{5}$$

For all input samples $\mathbf{x}$, the total loss is an average of the losses for all samples:

$$L(\mathbf{x}, \mathbf{y} \mid \mathbf{w}) = \frac{1}{N} \sum_{i=1}^{n} L(\mathbf{x}_i, y_i \mid \mathbf{w}) \tag{6}$$

$$= \frac{1}{N} \sum_{i=1}^{n} \left(\log\left(1 + e^{-\mathbf{x}_i^\top \mathbf{w}}\right) + \mathbf{x}_i^\top \mathbf{w}\left(1 - y_i\right)\right) \tag{7}$$

3 Back-propagation

$\mathbf{w}$ is the only variable in L. Therefore, we only need to calculate the gradient for $\mathbf{w}$:

$$\frac{\partial L}{\partial \mathbf{w}} = \frac{1}{N} \sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{w}} \log\left(1 + e^{-\mathbf{x}_i^\top \mathbf{w}}\right) + \frac{1}{N} \sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{w}}\left(\mathbf{x}_i^\top \mathbf{w}\left(1 - \mathbf{y}\right)\right) \tag{8}$$

$$= \frac{1}{N} \sum_{i=1}^{n} \frac{-\mathbf{x}_i e^{-\mathbf{x}_i^\top \mathbf{w}}}{1 + e^{-\mathbf{x}_i \mathbf{w}}} + \frac{1}{N} \sum_{i=1}^{n} \mathbf{x}_i^\top \left(1 - y_i\right) \tag{9}$$

$$= \frac{1}{N} \sum_{i=1}^{n} -\mathbf{x}_i \left(1 - \hat{y}_i\right) + \frac{1}{N} \sum_{i=1}^{n} \mathbf{x}_i^\top \left(1 - y_i\right) \tag{10}$$

$$= \frac{1}{N} \sum_{i=1}^{n} \mathbf{x}_i \left(\hat{y}_i - y_i\right) \tag{11}$$