Inclass 11: Logistic Regression

[SCS4049] Machine Learning and Data Science

Seongsik Park (s.park@dgu.edu)

Department of Artificial Intelligence, Dongguk University

Logistic regression

We begin our treatment of generalized linear models by considering the problem of two-class classification. Under rather general assumptions, the posterior probability of class $\mathcal{C}_1$ can be written as a logistic sigmoid acting on a linear function of the feature vector $\boldsymbol{\phi}$ so that

$$p(\mathcal{C}_1 \mid \boldsymbol{\phi}) = y(\boldsymbol{\phi}) = \sigma\left(\mathbf{w}^T \boldsymbol{\phi}\right) \tag{1}$$

with $p(\mathcal{C}_2 \mid \boldsymbol{\phi}) = 1 - p(\mathcal{C}_1 \mid \boldsymbol{\phi})$. Here, $\sigma(\cdot)$ is the *logistic sigmoid* function defined by

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{2}$$

In the terminology of statistics, this model is known as *logistic regression*, although it should be emphasized that this is a model for classification rather than regression.

Logistic regression

For a data set $\{\boldsymbol{\phi}_n, t_n\}$, where $t_n \in \{0, 1\}$ and $= (\mathbf{x}_n)$, with $n = 1, ..., N$, the likelihood function can be written

$$p(\mathbf{t} \mid \mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1 - y_n\}^{1-t_n} \tag{3}$$

where $\mathbf{t} = (t_1, ..., t_N)^T$ and $y_n = p(\mathcal{C}_1 \mid \phi_n)$. As usual, we can define an error function by taking the negative logarithm of the likelihood, which gives the *cross-entropy* error function in the form

$$E(\mathbf{w}) = -\log p(\mathbf{t} \mid \mathbf{w}) = -\sum_{n=1}^{N} \{t_n \log y_n + (1 - t_n) \log(1 - y_n)\} \tag{4}$$

where $y_n = \sigma(a_n)$ and $a_n = \mathbf{w}^T \boldsymbol{\phi}_n$. Taking the gradient of the error function with respect to $\mathbf{w}$, we obtain

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n) \boldsymbol{\phi}_n. \tag{5}$$

Logistic regression

The contribution to the gradient from data point n is given by the error $y_n - t_n$ between the target value and the prediction of the model, times the basis function vector $\boldsymbol{\phi}_n$. This takes precisely the same form as the gradient of the sum-of-squares error function for the linear regression model.