Error functions - classification

Kristian Wichmann

April 25, 2017

1 Continous-valued predictors

Most predictors for classification will be based on a continuous-valued score for each of the two outcomes. The prediction is then made based on these values (usually, the one with the highest score is chosen).

Given the values from such a predictor, we would like to have some quantitative measure of how good the prediction is.

1.1 Terminology

Consider classification between n cases. If the true classification for a sample x is k, then we define the n-dimensional vector t by $t_i = \delta_{ik}$. Let the values made by the predictor be the vector h, also a n-dimensional vector.

2 Error functions

2.1 Squared error

In this scheme we use the same error as used for OLS regression - the squared error:

$$J(h) = \frac{1}{2}||h - t||^2 = \frac{1}{2}(h - t)^T(h - t) = \frac{1}{2}\sum_{i=1}^{n}(h_i - t_i)^2 \tag{2.1}$$

2.1.1 Derivative

The derivative with respect to the h_j is:

$$\frac{\partial}{\partial h_j}J(h) = \frac{1}{2}\sum_{i=1}^{n}\delta_{ij}2(h_i - t_i) = h_j - t_j \tag{2.2}$$

2.2 Cross-entropy error: Bernoulli

This error is based on the sum of log-likelihoods of n Bernoulli experiments:

$$J(h) = \sum_{i=1}^{n} \left[-t_i \log(h_i) - (1 - t_i) \log(1 - h_i) \right] \tag{2.3}$$

2.2.1 Derivative

The derivative can be shown to be:

$$\frac{\partial}{\partial h_j} J(h) = \frac{h_j - t_j}{h_j(1 - h_j)} \tag{2.4}$$

2.3 Cross-entropy error

This error is similar to the one above, but here we consider h to be a vector of probabilities. These must sum to one. The cross-entropy between t and h is then:

$$J(h) = -\sum_{i=1}^{n} t_i \log(h_i) = -\log(h_k) \tag{2.5}$$

In the last step it's been used that $t_i = \delta_{ik}$.

2.3.1 Derivative

Once again, we're interested in the derivative:

$$\frac{\partial}{\partial h_j} J(h) = -\delta_{jk} \frac{1}{h_k} \tag{2.6}$$

3 Forms of the hypothesis h

The hypothesis h is in general the function of some m-dimensional input feature vector x, so $h = h(x)$. Here we will consider only hypotheses in which h_i depends on linearly weighted combinations of the vector x. The resulting vector z can therefore be written $z = w^T x$ in matrix form. Or element-wise $z_i = w_i x$. In other words, we only look at $h_i(x) = f_i(z) = f_i(w^T x)$. The functions f are known as the *activation functions*.

In general the derivative with respect to a weight w_{jk} will then be:

$$\frac{\partial}{\partial w_{jk}} h_i = \sum_l \frac{\partial f_i}{\partial z_l} \frac{\partial z_l}{\partial w_{jk}} \tag{3.1}$$

Since the last partial derivative is $\delta_{jl}x_k$ this means:

$$\frac{\partial}{\partial w_{jk}}h_i = \frac{\partial f_i}{\partial z_j}x_k \tag{3.2}$$

Often, the activation function f_i depends only on z_i, and so only derivatives with respect to w_{ij} are non-zero:

$$\frac{\partial}{\partial w_{ij}}h_i = \frac{\partial f_i}{\partial z_i}x_j \tag{3.3}$$

3.1 Linear hypothesis

Here, the hypothesis is simply equal to $w_i^T x$, so that:

$$h_i(x) = w_i^T x \tag{3.4}$$

3.2 Derivative

Since we're using the identity function as f, equation 3.3 is simply:

$$\frac{\partial}{\partial w_{ij}}h_i = x_j \tag{3.5}$$

3.3 Logistic hypothesis

Here, the hypothesis is the logistic function of $w_i^T x$. The logistic function is:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{3.6}$$

And so, the hypothesis is:

$$h_i(x) = \sigma(w_i^T x) \tag{3.7}$$

3.3.1 Derivative

We seek the derivative with respect to the weight w_{ij}. According to equation 3.3:

$$\frac{\partial}{\partial w_{ij}}h_i = \frac{\partial \sigma}{\partial z_i}x_j \tag{3.8}$$

The derivative of the logistic function can be shown to be $\sigma'(z) = \sigma(z)(1 - \sigma(z))$. So:

$$\frac{\partial}{\partial w_{ij}}h_i = \sigma(z_i)(1-\sigma(z_i))x_j \tag{3.9}$$

3.4 Softmax hypothesis

Here, the softmax function is used to normalize the hypothesis, so that they sum to 1. The softmax is defined by:

$$s_i(z) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{3.10}$$

And the hypothesis:

$$h_i(x) = s_i(w^T x) = \frac{e^{w_i^T x}}{N(w, x)} \tag{3.11}$$

Here, w is a matrix in which the columns are the weights w_i, and $N(w, x)$ is the normalization constant $\sum_{j=1}^{n} e^{z_j}$.

3.4.1 Derivative

It can be shown that the derivatives of the softmax is:

$$\frac{\partial s_i(z)}{\partial z_j} = s_i(z)(\delta_{ij} - s_j(z)) \tag{3.12}$$

So according to equation 3.2:

$$\frac{\partial}{\partial w_{jk}} h_i = s_i(z)(\delta_{ij} - s_j(z))x_k \tag{3.13}$$

4 Combining error and hypothesis types

4.1 Squared error with logistic hypothesis

Combining equations 2.1 and 3.7 we get the error function:

$$J(w, x) = \frac{1}{2} \sum_{i=1}^{n} (\sigma(w_i^T x) - t_i)^2 \tag{4.1}$$

4.2 Cross-entropy Bernoulli error with logistic hypothesis

Combining equations 2.3 and 3.7 we arrive at the error function:

$$J(w, x) = \sum_{i=1}^{n} \left[-t_i \log(\sigma(w_i^T x)) - (1 - t_i) \log(1 - \sigma(w_i^T x)) \right] \tag{4.2}$$

4.3 Cross-entropy error with softmax hypothesis

Combining equations 2.5 and 3.11 we arrive at the error function:

$$J(w, x) = -\log \frac{e^{w_k^T x}}{N(w, x)} = \log N(w, x) - w_k^T x \tag{4.3}$$