Least-square closed-form solution

Author: Zhihan Yang @ Carleton College

Date: Sunday, March 22, 2020 (during spring break)

The least-square method is a way to solve for the weights of a *generalized linear discriminant*, which is a form of single-layer networks.

Generalized linear discriminant

The generalized linear discriminant is given by:

$$y_o(\mathbf{x}) = \sum_{i=0}^{I} w_{o,i}\phi_i(\mathbf{x}) \tag{1}$$

- $\phi_i(\mathbf{x})$: a feature function that takes a vector $\mathbf{x}$ as input and outputs a scalar.

Loss function

The loss function E of a weight vector $\mathbf{w}$ is defined as follows:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\sum_{o=1}^{O}\{t_{n,o} - \sum_{i=0}^{I} w_{o,i}\phi_{n,i}\}^2 \tag{2}$$

- N: the number of training examples
- O: the number of output nodes / the number of dependent variables to model
- I: the number of input notes / the number of independent variables to model
- $t_{n,o}$: the label for the o-th output of the n-th training example
- $w_{o,i}$: the weight connecting from the i-th input node to the o-th output node
- $\phi_{n,i}$: the value of i-th feature of the n-th training example

Derivative of $E(\mathbf{w})$ with respect to $w_{o,i}$

Expand the summation over O:

$$\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\sum_{n=1}^{N}\sum_{o=1}^{O}\{t_{n,o} - \sum_{i=0}^{I} w_{o,i}\phi_{n,i}\}^2 & (3)\\
&= \frac{1}{2}\sum_{n=1}^{N}[(t_{n,1} - \sum_{i=0}^{I} w_{1,i}\phi_{n,i})^2 + \cdots + & (4)\\
&\quad \color{red}{(t_{n,o} - \sum_{i=0}^{I} w_{o,i}\phi_{n,i})^2} + \cdots + & (5)\\
&\quad (t_{n,O} - \sum_{i=0}^{I} w_{O,i}\phi_{n,i})^2] & (6)
\end{aligned}$$

We see that only the term highlighted with red involves the term $w_{o,i}$. For now, let us only consider the derivative of this part with respect to $w_{o,i}$, since the derivative of a sum (over N) is equivalent to a sum of the derivatives of its consisting elements. Before we derive its derivative, let us further expand it:

$$\begin{aligned}
E_{n,o}(\mathbf{w}) &= (t_{n,o} - \sum_{i=0}^{I} w_{o,i}\phi_{n,i})^2 & (7)\\
&= (\sum_{i=0}^{I} w_{o,i}\phi_{n,i} - t_{n,o})^2 & (8)\\
&= (w_{o,0}\phi_{n,0} + \cdots + \color{orange}{w_{o,i}}\color{blue}{\phi_{n,i}} + \cdots + w_{o,I}\phi_{n,I} - t_{n,o})^2 & (9)
\end{aligned}$$

The terms that won't be zero after differentiation are those that directly involve $w_{o,i}$:

$$2w_{o,0}\color{orange}{w_{o,i}}\phi_{n,0}\color{blue}{\phi_{n,i}} + \cdots + w_{o,i}^{\color{orange}{2}}\phi_{n,i}^{\color{blue}{2}} + \cdots + 2w_{o,I}\color{orange}{w_{o,i}}\phi_{n,I}\color{blue}{\phi_{n,i}} - 2\color{orange}{w_{o,i}}\color{blue}{\phi_{n,i}}t_{n,o} \tag{10}$$

Taking the derivative of the long expression above with respect to $w_{o,i}$, obtain:

$$\frac{\partial E_{n,o}(\mathbf{w})}{\partial w_{o,i}} = 2w_{o,0}\phi_{n,0}\phi_{n,i} + \cdots + 2w_{o,i}\phi_{n,i}^2 + \cdots + 2w_{o,I}\phi_{n,I}\phi_{n,i} - 2\phi_{n,i}t_{n,o} \tag{11}$$

Factor, where i' and o' serve the same purpose as i and o:

$$= 2[(\sum_{i'=0}^{I} w_{o',i'}\phi_{n,i'}) - t_{n,o}]\phi_{n,i} \tag{12}$$

Therefore,

$$\frac{\partial E(\mathbf{w})}{\partial w_{o,i}} = \sum_{n=1}^{N}\{[(\sum_{i'=0}^{I} w_{o',i'}\phi_{n,i'}) - t_{n,o}]\phi_{n,i}\} \tag{13}$$

We set the RHS to zero the solve for the value of $w_{o,i}$ that leads to zero loss:

$$\sum_{n=1}^{N}\{[(\sum_{i'=0}^{I} w_{o',i'}\phi_{n,i'}) - t_{n,o}]\phi_{n,i}\} = 0 \tag{14}$$

In order to find a solution to the equation above it is convenient to write it in a matrix notation. The challenge is how?

Matrix notation

A matrix multiplication between two matrices can be denoted by:

$$(AB)_{i,j} = \sum_{n} A_{i,n}B_{n,j} \tag{15}$$

A matrix multiplication between three matrices can be denoted by:

$$(ABC)_{i,j} = \sum_{n}(AB)_{i,n}C_{n,j} \tag{16}$$

$$= \sum_{n}(\sum_{n'} A_{i,n'}B_{n,'j})C_{n,j} \tag{17}$$

Notice the similarity between $\frac{\partial E(\mathbf{w})}{\partial w_{o,i}}$ and the expansion of a triple-matrix multiplication as summations. Therefore, we can convert $\frac{\partial E(\mathbf{w})}{\partial w_{o,i}}$ into matrix notation involving a triple-matrix multiplication:

$$\sum_{n=1}^{N}\{[(\sum_{i'=0}^{I} w_{o',i'}\phi_{n,i'}) - t_{n,o}]\phi_{n,i}\} = 0 \tag{18}$$

$$\sum_{n=1}^{N}\{(\sum_{i'=0}^{I} w_{o',i'}\phi_{n,i'})\phi_{n,i} - t_{n,o}\phi_{n,i}\} = 0 \tag{19}$$

$$\sum_{n=1}^{N}(\sum_{i'=0}^{I} w_{o',i'}\phi_{n,i'})\phi_{n,i} - \sum_{n=1}^{N} t_{n,o}\phi_{n,i} = 0 \tag{20}$$

To get the dimensions of the matrices right, align the axis alphabets:

$$\sum_{n=1}^{N}(\sum_{i'=0}^{I} w_{o',i'}\phi_{i',n})\phi_{n,i} = \sum_{n=1}^{N} t_{o,n}\phi_{n,i} \tag{21}$$

$$W^T\boldsymbol{\phi}^T\boldsymbol{\phi} = T^T\boldsymbol{\phi} \tag{22}$$

where the following shapes have been assumed:

- W: I, O
- $\boldsymbol{\phi}$: N, I
- T: N, O
- so that $\boldsymbol{\phi}W$ would have the same shape as T.

To have the equation in terms of W instead of W_T, transpose both sides:

$$(\boldsymbol{\phi}^T\boldsymbol{\phi})W = \boldsymbol{\phi}^T T \tag{23}$$

$$W = (\boldsymbol{\phi}^T\boldsymbol{\phi})^{-1}\boldsymbol{\phi}^T T \tag{24}$$

$$= \boldsymbol{\phi}^\star T \tag{25}$$

where $\boldsymbol{\phi}^* = (\boldsymbol{\phi}^T\boldsymbol{\phi})^{-1}\boldsymbol{\phi}^T$ and is called the pseudo-inverse of $\boldsymbol{\phi}$.