

Multimedia Appendix 3. Constructing a predictive model in the recurrent neural network (RNN).

The learning sample set \mathcal{S} is consisted of ordered pairs (\mathbf{x}, \mathbf{y}) with inputs and correct answers. Input \mathbf{x} is an element of input set \mathcal{X} , whereas correct answer \mathbf{y} is an element of correct answers set \mathcal{Y} . Here, \mathbf{x} is consisted of serial data in the form of an individual's information. The T in input $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_t, \dots, \mathbf{x}_T)$ represents the length of a sample (i.e. the number of events) and varies individually. A health examination record is consisted of an event; and, all events $\mathbf{x}_1, \dots, \mathbf{x}_T$ for a person are in chronological order. \mathbf{x}_t represents an event (examination record) at a specific time and is used as a vector with 9 features. The correct answer $\mathbf{y} = (y_1, \dots, y_K)$ has a Boolean value indicating whether or not dementia occurred in the past. The correct answer is one-hot encoded to be used as cross entropy in a loss function. K —the number of classes—is set as 2. Assuming the output of prediction model is $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_K)$, cross entropy as shown in Equation 1 is used for our loss function.

$$E(y, \hat{y}) = - \sum_{i=1}^K [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (\text{Equation 1})$$

The output sequence for Long Short-Term Memory (LSTM) was $\mathbf{o} = (\mathbf{o}_1, \dots, \mathbf{o}_t, \dots, \mathbf{o}_T)$ where T was the event length being the same as the event length of \mathbf{x} . The prediction result of an input sample provided the probability of dementia in the near future taken the occurrence of past event into account.

Only the last output \mathbf{o}_T among $\mathbf{o}_1, \dots, \mathbf{o}_t, \dots, \mathbf{o}_T$ was used and reflected to an output z as shown in Equation 2. Here, $W \in \mathbb{R}^{K \times H}$ was the parameter to be optimized and H was the number of hidden nodes in the last hidden layer. To calculate the probability of \hat{y}_i from z , the Softmax function was used as shown in Equation 3.

$$z = W\mathbf{o}_T \quad (\text{Equation 2})$$

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad \text{for } i = 1, \dots, K \quad (\text{Equation 3})$$

LSTM had a memory cell with input, forget and output gates. Each LSTM unit uses the equations in Equation 4 which are commonly used in LSTM.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci} \circ c_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf} \circ c_{t-1} + b_f)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co} \circ c_t + b_o)$$

$$h_t = o_t \circ \tanh(c_t) \tag{Equation 4}$$

σ was the logistic sigmoid function and i , f , o and c were respectively the *input gate*, *forget gate*, *output gate*, and *cell*. The subscripts in weight matrix above have an obvious meaning. For example, W_{hi} is the hidden-input gate matrix while W_{xo} being the input-output gate matrix. The b s are bias terms which are added for i , f , o and c equations. Let N be the number of LSTM blocks and M the number of inputs, $W_i, W_f, W_c, W_o \in \mathbb{R}^{N \times M}$