RNN

HSK

April 2024

1 Introduce

$$h_t = \Theta_h(W\ h_{t-1}\ +\ U\ x_t\ +\ b) \tag{1}$$
$$\Omega_t = V\ h_t\ +\ c \tag{2}$$
$$\hat{y}_t = \Theta_y(\Omega_t) \tag{3}$$
$$L_t = -\ y_t\ ln(\hat{y}_t) \tag{4}$$

Activation functions are Θ_h which represents *tanh* in (1), and Θ_y *softmax* in (3).

2 Through Time for Recurrent Neural Network

2.1 Softmax

$$\text{softmax}(\Omega_t) = \hat{y}_t = \frac{e^{\Omega_t}}{\sum_{k=1}^{|\mathcal{A}|} e^{\Omega_{t,k}}} \text{ for } t = 1, \ldots, k$$

Let us compute $\frac{\partial}{\partial \Omega_{t,j}}(\hat{y}_i)$ for some arbitrary i and j :

$$\frac{\partial \hat{y}_i}{\partial \Omega_{t,j}} = \frac{\partial}{\partial \Omega_{t,j}} \left(\frac{e^{\Omega_{t,i}}}{\sum_k e^{\Omega_{t,k}}} \right)$$

Since $\frac{\partial}{\partial \Omega_{t,j}} e^{\Omega_{t,k}} = 0$ for $k \neq j$, we have:

$$\frac{\partial}{\partial \Omega_{t,j}} \left(\sum e^{\Omega_{t,k}} \right) = \sum \left(\frac{\partial\ e^{\Omega_{t,k}}}{\partial \Omega_{t,j}} \right) = e^{\Omega_{t,j}}$$

Only meaningful derivatives is obtained for $i = j$ case in the above equation for our example presented in this chapter. Recall that in our example only one of values is a one

$$
\begin{aligned}
\frac{\partial}{\partial \Omega_{t,j}} \left(\frac{e^{\Omega_{t,i}}}{\sum e^{\Omega_{t,k}}} \right) &= \frac{e^{\Omega_{t,i}} \sum e^{\Omega_{t,k}} - e^{\Omega_{t,j}} e^{\Omega_{t,i}}}{\left(\sum e^{\Omega_{t,k}}\right)^2} \\
&= \frac{e^{\Omega_{t,i}} \left(\sum e^{\Omega_{t,k}} - e^{\Omega_{t,j}}\right)}{\left(\sum e^{\Omega_{t,k}}\right)^2} \\
&= \frac{e^{\Omega_{t,i}}}{\sum e^{\Omega_{t,k}}} \cdot \left(\frac{\sum e^{\Omega_{t,k}}}{\sum e^{\Omega_{t,k}}} - \frac{e^{\Omega_{t,j}}}{\sum e^{\Omega_{t,k}}} \right)
\end{aligned}
$$

$$
\hat{y}_{t,i} \left(1 - \hat{y}_{t,j}\right) \tag{5}
$$

2.2 Derivative of Loss Function w.r.t. Ω_t

Recall that cross-entropy loss is defined as:

$$
L = -\sum_{t=1}^{\mathcal{S}} y_t \ln(\hat{y}_t)
$$

Let us compute the partial derivative of L_t with respect to Ω_t at step t:

$$
\begin{aligned}
\frac{\partial L_t}{\partial \Omega} &= -\frac{\partial}{\partial \Omega} y_t \ln \hat{y}_t \ \ = -y_t \frac{\partial}{\partial \Omega_t} \log \hat{y}_t \\
&= -\sum y_t \cdot \frac{1}{\hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \\
&= -\hat{y}_{t,i} \left(1 - \hat{y}_{t,j}\right) \cdot \frac{y_t}{\hat{y}_t} \\
&= -\left(1 - \hat{y}_{t,j}\right) y_t \\
&= -\left(y_t - \hat{y}_{t,j} \hat{y}_t\right) \\
&= \hat{y}_{t,i} \hat{y}_t - y_t \\
&= \hat{y}_{t,j} \cdot \hat{y}_t \ (?) \ \ - y_t \\
&= \left(\hat{y}_t - y_t\right)
\end{aligned}
$$

$$
\frac{\partial L_t}{\partial \hat{\Omega}_t} = \left(\hat{y}_t - y_t\right) \tag{6}
$$

2.3 Derivative of V

The weight V is consistent across the entire time sequence, allowing us to perform differentiation at each time step and then aggregate the results.

$$
\begin{aligned}
\frac{\partial L}{\partial V} &= \sum_{t=1}^{S} \frac{\partial L_t}{\partial V} \\
&= \sum_{t=1}^{S} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial V} \\
&= \sum_{t=1}^{S} \frac{\partial L}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial V}
\end{aligned}
$$

We know that this formula $\frac{\partial \hat{y}_t}{\partial \Omega_t}$ from (**??**) and no other function exists between Omega and V, so simply taking the derivative coefficient of V yields h, thus the answer is h.

$$
= \sum_{t=1}^{S} (\hat{y}_t - y_t) \cdot h_t^{\top} \tag{7}
$$

2.4 Derivative of c

Similar to V, but its derivative is easier to calculate since it stands alone in the function.

$$
\begin{aligned}
\frac{\partial L}{\partial c} &= \sum_{t=1}^{T} \frac{\partial L_t}{\partial c} \\
&= \sum_{t=1}^{T} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial c} \\
&= \sum_{t=1}^{T} \frac{\partial L}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial c}
\end{aligned}
$$

In this case, The Analytical Derivatives of c becomes:

$$
= \sum_{t=1}^{T} (\hat{y}_t - y_t) \tag{8}
$$

2.5 Derivative of W

This function employs recursion, therefore, computing its derivative may take some time.

$$
\begin{aligned}
&h_t = tanh(W\ h_{t-1}\ +\ U\ x_t\ +\ b) \\
&h_1 = tanh(W\ h_0\ +\ U\ x_1\ +\ b) \quad h_2 = tanh(W\ h_1\ +\ U\ x_2\ +\ b) \\
&h_3 = tanh(W\ h_2\ +\ U\ x_3\ +\ b) \quad h_4 = tanh(W\ h_3\ +\ U\ x_4\ +\ b)
\end{aligned}
$$

By placing previous hidden layer terms into h_4, we get:

$h_4 = tanh(W\ tanh(W\ tanh(W\ tanh(Wh_0 + Ux_1 + b) + Ux_2 + b) + Ux_3 + b) + Ux_4 + b)$

We start from the first step go to the last step:

$$\frac{\partial L_1}{\partial W} = \frac{\partial L_1}{\partial \hat{y}_1}\frac{\partial \hat{y}_1}{\partial h_1}\frac{\partial h_1}{\partial W}$$
$$\frac{\partial L_2}{\partial W} = \frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial h_2}\frac{\partial h_2}{\partial W}$$
$$\frac{\partial L_3}{\partial W} = \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W} + \frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial h_3}\frac{\partial h_3}{\partial W}$$
$$\frac{\partial L_4}{\partial W} = \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial h_4}\frac{\partial h_4}{\partial W}$$

Let us group them under a $\sum\prod$ for step t:

$$\frac{\partial L_t}{\partial W} = \sum_{k=1}^{t} \ \frac{\partial L_t}{\partial \hat{y}_t}\frac{\partial \hat{y}_t}{\partial h_t}\left(\prod_{j=k}^{t-1}\frac{\partial h_{j+1}}{\partial h_j}\right)\frac{\partial h_k}{\partial W}$$

Let us now present the formula for $\mathcal{S}$ steps:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{\mathcal{S}}\left(\sum_{k=1}^{t} \ \frac{\partial L_t}{\partial \hat{y}_t}\frac{\partial \hat{y}_t}{\partial h_t}\prod_{j=k}^{t-1}\left(\frac{\partial h_{j+1}}{\partial h_j}\right)\frac{\partial h_k}{\partial W}\right) \tag{9}$$

Finally, we insert the individual partial derivatives to calculate our final gradients of L with respect to W, where:

$$\frac{\partial L_t}{\partial \hat{y}_t} = (y_t - \hat{y}_t)$$
$$\frac{\partial \hat{y}_t}{\partial h_t} = V^\top$$
$$\frac{\partial h_{j+1}}{\partial h_j} = W^\top\ (1 - h_{j+1}^2)$$
$$\frac{\partial h_k}{\partial W} = (1 - h_k^2)\ h_{k-1}$$

In this case, The Analytical Derivatives of Eq. (9) becomes:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{\mathcal{S}}\left(\sum_{k=1}^{t}\ (y_t - \hat{y}_t)\ V^\top\prod_{j=k}^{t-1}\left(\ W^\top\ (1 - h_{j+1}^2)\right)(1 - h_k^2)\ h_{k-1}\ \right) \tag{10}$$

2.6 Derivative of U

Now, let us compute the partial derivation of L with respect to U. Similar to the case of W in Eq. (9), for U we have:

$$\frac{\partial L}{\partial U} = \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \; \frac{\partial L_t}{\partial \hat{y}_t} \, \frac{\partial \hat{y}_t}{\partial h_t} \prod_{j=k}^{t-1} \left(\; \frac{\partial h_{j+1}}{\partial h_j} \right) \frac{\partial h_k}{\partial U} \; \right) \tag{11}$$

We insert the individual partial derivatives into Eq. (11) as follows:

$$\begin{aligned}
&\text{From Eq. (6):} \quad && \frac{\partial L_t}{\partial \hat{y}_t} = (y_t - \hat{y}_t) \\
&\text{From Eq. (2):} \quad && \frac{\partial \hat{y}_t}{\partial h_t} = V^\top \\
&\text{From Eq. (1):} \quad && \frac{\partial h_{j+1}}{\partial h_j} = W^\top \; (1 - h_{j+1}^2) \\
&\text{From Eq. (1):} \quad && \frac{\partial h_k}{\partial U} = (1 - h_k^2) \; x_k
\end{aligned}$$

Inserting the above derivatives into Eq. (11), we have:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \;\; (y_t - \hat{y}_t) \; V^\top \prod_{j=k}^{t-1} \left(\; W^\top \; (1 - h_{j+1}^2) \right) (1 - h_k^2) \; h_{k-1} \; \right) \tag{12}$$

2.7 Derivative of b

In the same manner, gradient of L with respect to b is calculated similar to Eq. (11) as follows:

$$\frac{\partial L}{\partial b} = \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \; \frac{\partial L_t}{\partial \hat{y}_t} \, \frac{\partial \hat{y}_t}{\partial h_t} \prod_{j=k}^{t-1} \left(\; \frac{\partial h_{j+1}}{\partial h_j} \right) \frac{\partial h_k}{\partial b} \; \right) \tag{13}$$

Recall that the derivatives used in Eq. (13) are:

$$\begin{aligned}
\frac{\partial L_t}{\partial \hat{y}_t} &= (y_t - \hat{y}_t) \\
\frac{\partial \hat{y}_t}{\partial h_t} &= V^\top \\
\frac{\partial h_{j+1}}{\partial h_j} &= W^\top \; (1 - h_{j+1}^2) \\
\frac{\partial h_k}{\partial b} &= (1 - h_k^2)
\end{aligned}$$

In this case, Eq. (13) becomes:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \;\; (y_t - \hat{y}_t) \; V^\top \prod_{j=k}^{t-1} \left(\; W^\top \; (1 - h_{j+1}^2) \right) (1 - h_k^2) \; \right) \tag{14}$$