HW1 for Intro to NLP

Fudi(Fred) Wang

October 2019

Part 1

1

Since f(.) is chosen as a linear function, then without loss of generality, let's suppose that f(x) = ax + b, for some constants a and b. Then,

$$\begin{aligned}
g_k(x) \equiv y_k &= f\big(\sum_j w_{kj} f(\sum_i w_{ji}x_i + w_{j0}) + w_{k0}\big) \\
&= a_2\big(\sum_j w_{kj}[a_1(\sum_i w_{ji}x_i + w_{j0}) + b_1] + w_{k0}\big) + b_2 \\
&= a_1a_2 \sum_j w_{kj} \sum_i w_{ji}x_i + a_2b_1 \sum_j w_{kj} + a_2b_2w_{k0}
\end{aligned}$$

for some constants a_1, a_2, b_1, b_2. Therefore, we can see that y_k eventually becomes the form of $\sum_i a_ix_i + b$, for some constants a_i and b, which is linear. Hence, nonlinearity will not be achieved if the activation function is chosen as a linear function.

2

(a) Since $Z = W^TX$, we have

$$\frac{\partial E}{\partial W} = \frac{\partial E}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial W} = -(g-y)y(1-y)x^T$$

where

$$y = \frac{1}{1+e^{-W^TX}}$$

Hence, the weight-updating formula follows:

$$W_{new} = W - \lambda\frac{\partial E}{\partial W} = W + (g-y)y(1-y)x^T$$

(b)

$$y = \frac{1}{1+e^{-W^TX}} = \frac{1}{1+e^{-3}} = 0.95$$

Hence, the new weight vector becomes:

$$W_{new} = W - \lambda\frac{\partial E}{\partial W} = [0.5, 1, 1]^T + 2.375 \times 10^{-4}[1, 2, 0.5]^T = [0.5002, 1.0004, 1.0001]^T$$