Machine Learning II, Sheet2

Fynn Beuttenmller, Jan Lammel

May 19, 2016

1 Theory

1.1 Linear Activation Function

$$\begin{aligned}
Z_l =& \Phi_l(\tilde{Z}_l) = \Phi_l(B_l Z_{l-1}) \overbrace{=}^{\Phi_l linear} B_l \Phi_l(\Phi_{l-1}(\tilde{Z}_{l-1})) && (1)\\
=& B_l \Phi_l(\Phi_{l-1}(B_{l-1} Z_{l-2})) \underbrace{=}_{\Phi_{l-1} linear} B_l B_{l-1} \Phi_l(\Phi_{l-1}(Z_{l-2})) && (2)\\
=& ... = \underbrace{\prod_{l=L}^{1} B_l}_{=:\bar{B}} \cdot \underbrace{\Phi_L \circ .. \circ \Phi_1}_{=:\bar{\Phi}}(Z_0) && (3)\\
=& \bar{B}\bar{\Phi}(Z_0) && (4)
\end{aligned}$$

$\Rightarrow$ 1 Layer form

1.2 Weight Decayle

1.2.1 Part 1

$$Loss(\omega) = L_0(\omega) + L_{reg}(\omega) = L_0(\omega) + \frac{\lambda}{2N}\omega^T \omega \tag{5}$$

$$\frac{\partial Loss}{\partial \omega} = \frac{\partial L_0}{\partial \omega} + \frac{\partial L_{reg}}{\partial \omega} = \frac{\partial L_0}{\partial \omega} + \frac{\lambda}{N}\omega \tag{6}$$

$$\begin{aligned}
\omega^{(t+1)} =& \omega^{(t)} - \tau \frac{\partial Loss}{\partial \omega} = \omega^t - \tau \left(\frac{\partial L_0}{\partial \omega} + \frac{\lambda}{N}\omega^{(t)} \right) && (7)\\
=& \left(1 - \underbrace{\frac{\lambda}{N}\tau}_{=:\epsilon} \right) \omega^{(t)} - \tau \frac{\partial L_0}{\partial \omega} && (8)\\
=& (1-\epsilon)\,\omega^{(t)} - \tau \frac{\partial L_0}{\partial \omega} && (9)
\end{aligned}$$

1.2.2 Part 2

$(1-\epsilon)\omega^{(t)}$ reduces magnitude of $||\omega||$, which prevents one weight to become dominant over the others.

1.2.3 Part 3

$$L_{reg} = \frac{\lambda}{2N}||\omega||_1 \tag{10}$$

$$\frac{\partial L_r eg}{\partial \omega} = \frac{\lambda}{2N} sign(\omega) \tag{11}$$

$$\Rightarrow \omega^{(t+1)} = \omega^{(t)} - \tau\left(\frac{\partial L_0}{\partial w} + \frac{\lambda}{2N} sign(\omega^{(t)})\right) \tag{12}$$

1.2.4 Part 4

Due to the weight decay the bias weight would become more and more dominant against the others, so there is no benefit on doing that.