DS 598 DEEP LEARNING

Sungjoon Park (BUID: U38522578)

Feb. 26th, 2024

Knowledge Checks 6-9

1. (Problem 6.10)

$$m_0 = 0$$

$$m_1 = \beta m_0 + (1-\beta) \sum_{i \in \mathcal{B}_0} [\frac{\partial \ell_i[\phi_0]}{\partial \phi}] = (1-\beta) \sum_{i \in \mathcal{B}_0} [\frac{\partial \ell_i[\phi_0]}{\partial \phi}] = 0$$

$$m_2 = \beta m_1 + (1-\beta) \sum_{i \in \mathcal{B}_1} [\frac{\partial \ell_i[\phi_1]}{\partial \phi}] = (1-\beta) \sum_{i \in \mathcal{B}_1} [\frac{\partial \ell_i[\phi_1]}{\partial \phi}]$$

$$\begin{aligned} m_3 &= \beta m_2 + (1-\beta) \sum_{i \in \mathcal{B}_2} [\frac{\partial \ell_i[\phi_2]}{\partial \phi}] \\ &= \beta(1-\beta) \sum_{i \in \mathcal{B}_1} [\frac{\partial \ell_i[\phi_1]}{\partial \phi}] + (1-\beta) \sum_{i \in \mathcal{B}_2} [\frac{\partial \ell_i[\phi_2]}{\partial \phi}] \end{aligned}$$

Likewise,

$$m_4 = \beta^2(1-\beta) \sum_{i \in \mathcal{B}_1} [\frac{\partial \ell_i[\phi_1]}{\partial \phi}] + \beta(1-\beta) \sum_{i \in \mathcal{B}_2} [\frac{\partial \ell_i[\phi_2]}{\partial \phi}] + (1-\beta) \sum_{i \in \mathcal{B}_3} [\frac{\partial \ell_i[\phi_3]}{\partial \phi}]$$

Therefore,

$$m_t = \beta^{t-2}(1-\beta) \sum_{i \in \mathcal{B}_1} [\frac{\partial \ell_i[\phi_1]}{\partial \phi}] + \cdots + (1-\beta) \sum_{i \in \mathcal{B}_{t-1}} [\frac{\partial \ell_i[\phi_{t-1}]}{\partial \phi}]$$

Since all summations are gradients and $\beta^{t-2}(1-\beta) + \cdots + (1-\beta) = \frac{\beta}{1-(1-\beta)} = 1$ as t goes to infinity, m_t is an infinite weighted sum of the gradients (as t goes to infinity).

2. (Problem 7.3)

< The first term >

$$\frac{\partial h_1}{\partial f_0}$$

$\mathbb{R}^{D_1 \times D_1}$ (D_1: the number of hidden units in the first hidden layer)

< The second term >

$$\frac{\partial f_1}{\partial h_1}$$

$\mathbb{R}^{D_1 \times D_2}$ (D_2: the number of hidden units in the second hidden layer)

< The third term >

$$\frac{\partial h_2}{\partial f_1}$$

$\mathbb{R}^{D_2 \times D_2}$

< The fourth term >

$$\frac{\partial f_2}{\partial h_2}$$

$\mathbb{R}^{D_2 \times D_3}$ (D_3: the number of hidden units in the third hidden layer)

< The fifth term >

$$\frac{\partial h_3}{\partial f_2}$$

$\mathbb{R}^{D_3 \times D_3}$

< The sixth term >

$$\frac{\partial f_3}{\partial h_3}$$

$\mathbb{R}^{D_3 \times D_f}$ (D_f: the dimensionality of the model output f_3)

< The last term >

$$\frac{\partial \ell_i}{\partial f_3}$$

$D_{f_3} \times 1$

3. (Problem 8.2)

$< h_1 >$

weight: $(1-0)/(1-0) = 1$

bias: 0

$< h_2 >$

weight: $(1-1/3)/(2/3-0) = 1$

bias: $1 \cdot 1/3 + \text{bias} = 0 \rightarrow \text{bias} = -1/3$

$< h_3 > \leftarrow$ Likewise in $< h_2 >$

weight: $(1-2/3)/(1/3-0) = 1$

bias: $1 \cdot 2/3 + \text{bias} = 0 \rightarrow \text{bias} = -2/3$

4. (Problem 9.1)

$$\prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{x}_i,\phi)Pr(\phi) = \prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{x}_i,\phi)\prod_{j=1}^{J} Norm_{\phi_j}[0,\sigma_\phi^2]$$

The likelihood function is as follows.

$$\begin{aligned}
&\prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{x}_i,\phi)\prod_{j=1}^{J} Norm_{\phi_j}[0,\sigma_\phi^2] \\
&= \prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{x}_i,\phi)\prod_{j=1}^{J}[\frac{1}{\sqrt{2\pi\sigma_\phi^2}}\exp(-\frac{\phi_j^2}{2\sigma_\phi^2})] \\
&\propto \prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{x}_i,\phi)\prod_{j=1}^{J}\exp(-\frac{\phi_j^2}{2\sigma_\phi^2})
\end{aligned}$$

Thus, the simplified negative log-likelihood function is

$$\begin{aligned}
&-\sum_{i=1}^{I}\log(Pr(\mathbf{y}_i|\mathbf{x}_i,\phi)) - \sum_{j=1}^{J}(-\frac{\phi_j^2}{2\sigma_\phi^2}) \\
&= -\sum_{i=1}^{I}\log(Pr(\mathbf{y}_i|\mathbf{x}_i,\phi)) + \sum_{j=1}^{J}\frac{\phi_j^2}{2\sigma_\phi^2} \\
&= -\sum_{i=1}^{I}\log(Pr(\mathbf{y}_i|\mathbf{x}_i,\phi)) + \frac{1}{2\sigma_\phi^2}\sum_{j=1}^{J}\phi_j^2
\end{aligned}$$

Let $\lambda = \frac{1}{2\sigma_\phi^2}$

$$= -\sum_{i=1}^{I}\log(Pr(\mathbf{y}_i|\mathbf{x}_i,\phi)) + \lambda\sum_{j=1}^{J}\phi_j^2$$

This is nothing but imposing L2-norm regularization.