Assignment 1

Chen Luo

March 18, 2018

1 Softmax

1.1 a

$$softmax(\boldsymbol{x} + c) = \frac{e^{\boldsymbol{x+c}}}{\sum_{j=1}^{d} e^{x_j + c}} \tag{1}$$

$$= \frac{e^{\boldsymbol{x}} e^{c}}{e^{c} \sum_{j=1}^{d} e^{x_j}} = softmax(\boldsymbol{x}) \tag{2}$$

2 NN Basic

2.1 a

$$\frac{\partial \sigma(x)}{\partial x} = -1(1 + e^{-x})^{-2} \frac{\partial (1 + e^{-x})}{\partial x} \tag{3}$$

$$= \frac{e^{-x}}{(1 + e^{-x})^2} \tag{4}$$

$$= \frac{e^{-x} + 1 - 1}{(1 + e^{-x})^2} \tag{5}$$

$$= \sigma(x) - \sigma(x)^2 \tag{6}$$

2.2 b

$$CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_i y_i \log(\hat{y}_i) \tag{7}$$

$$\hat{\boldsymbol{y}} = softmax(\boldsymbol{\theta}) = \frac{e^{\boldsymbol{\theta}}}{\sum_j e^{\theta_j}} \tag{8}$$

$$\begin{aligned}
\frac{\partial CE(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \boldsymbol{\theta}} &= -\frac{\partial \sum_i 1_{[y_i==1]} \left\{ \log e^{\theta_i} - \log \sum_j e^{\theta_j} \right\}}{\partial \boldsymbol{\theta}} && (9)\\
&= -\frac{\partial \boldsymbol{y}^T \boldsymbol{\theta} - log \sum_j e^{\theta_j}}{\partial \boldsymbol{\theta}} && (10)\\
&= -\boldsymbol{y} + \frac{\log \mathbf{1}^T e^{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} && (11)\\
&= -\boldsymbol{y} + \frac{e^{\boldsymbol{\theta}}}{\mathbf{1}^T e^{\boldsymbol{\theta}}} && (12)\\
&= -\boldsymbol{y} + softmax(\boldsymbol{\theta}) && (13)\\
&= \hat{\boldsymbol{y}} - \boldsymbol{y} && (14)
\end{aligned}$$

2.3 c

$$\begin{aligned}
J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= CE(\boldsymbol{y}, softmax(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)) && (15)\\
&= CE(\boldsymbol{y}, softmax(\sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2)) && (16)
\end{aligned}$$

Actually,

$$J = -\boldsymbol{y}_{1\times D_y} \log(\boldsymbol{h}_{1\times H} \boldsymbol{W}_{H\times D_y} + \boldsymbol{b}_{1\times D_y})^T \tag{17}$$

$$\begin{aligned}
\frac{\partial J}{\partial \boldsymbol{x}} &= \frac{\partial J}{\partial(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)} \frac{\partial(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}} && (18)\\
&= (\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{W}_2^T \sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\{\mathbf{1} - \sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\}^T \boldsymbol{W}_1^T && (19)
\end{aligned}$$

, where $\frac{\partial(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)}{\partial \boldsymbol{h}}$ is Jaccobian matrix.

Or, [TODO]

$$\begin{aligned}
d(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2) &= d\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{h}d\boldsymbol{W}_2 + d\boldsymbol{b}_2 && (20)\\
vec(d(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)) &= vec(d\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{h}d\boldsymbol{W}_2 + d\boldsymbol{b}_2) && (21)\\
&= vec(\mathbf{1}d\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{h}d\boldsymbol{W}_2\boldsymbol{I} + \mathbf{1}d\boldsymbol{b}_2\boldsymbol{I}) \to preparation for && (22)\\
vec(AXB) &= (B^T \otimes A)vec(X) && (23)\\
&= \boldsymbol{W}_2^T \otimes 1 vec(d\boldsymbol{h}) + \boldsymbol{I} \otimes \boldsymbol{h} vec(d\boldsymbol{W}_2) + \boldsymbol{I} \otimes 1 vec(d\boldsymbol{b}_2) && (24)
\end{aligned}$$

2.4 d

$$\boldsymbol{W}_1 : D_x \times H \tag{25}$$
$$\boldsymbol{b}_1 : 1 \times H \tag{26}$$
$$\boldsymbol{W}_2 : H \times D_y \tag{27}$$
$$\boldsymbol{b}_2 : 1 \times D_y \tag{28}$$
$$H(D_x + D_y + 1) + D_y \tag{29}$$

2.5 g

$$dJ = (\hat{\boldsymbol{y}} - \boldsymbol{y})^T d(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2) = (\hat{\boldsymbol{y}} - \boldsymbol{y})^T (d\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{h}d\boldsymbol{W}_2 + d\boldsymbol{b}_2) \tag{30}$$
$$= (\hat{\boldsymbol{y}} - \boldsymbol{y})^T d\boldsymbol{h}\boldsymbol{W}_2 + (\hat{\boldsymbol{y}} - \boldsymbol{y})^T \boldsymbol{h}d\boldsymbol{W}_2 + (\hat{\boldsymbol{y}} - \boldsymbol{y})^T d\boldsymbol{b}_2) \tag{31}$$
$$= trace((\hat{\boldsymbol{y}} - \boldsymbol{y})^T d\boldsymbol{h}\boldsymbol{W}_2 + (\hat{\boldsymbol{y}} - \boldsymbol{y})^T \boldsymbol{h}d\boldsymbol{W}_2 + (\hat{\boldsymbol{y}} - \boldsymbol{y})^T d\boldsymbol{b}_2)) \tag{32}$$
$$\frac{\partial J}{\partial \boldsymbol{W}_2} = (\boldsymbol{W}_2(\hat{\boldsymbol{y}} - \boldsymbol{y})^T)^T \tag{33}$$
$$= \boldsymbol{h}^T(\hat{\boldsymbol{y}} - \boldsymbol{y}) \tag{34}$$
$$\frac{\partial J}{\partial \boldsymbol{b}_2} = \hat{\boldsymbol{y}} - \boldsymbol{y} \tag{35}$$
$$d\boldsymbol{h} = d(\sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)) = \sigma'(\boldsymbol{\cdot}) \odot (\boldsymbol{x}d\boldsymbol{W}_1 + d\boldsymbol{b}_1) \tag{36}$$
$$dJ = trace((\hat{\boldsymbol{y}} - \boldsymbol{y})^T \sigma'(\cdot) \odot (\boldsymbol{x}d\boldsymbol{W}_1 + d\boldsymbol{b}_1)\boldsymbol{W}_2) \tag{37}$$
$$= trace(\boldsymbol{W}_2(\hat{\boldsymbol{y}} - \boldsymbol{y})^T \sigma'(\cdot) \odot (\boldsymbol{x}d\boldsymbol{W}_1 + d\boldsymbol{b}_1)) \tag{38}$$
$$Apply\ law\ of\ trace\ and\ element - wise\ product: \tag{39}$$
$$trace(A^T(B \odot C)) = trace((A \odot B)^T C) \tag{40}$$
$$= trace([(\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{W}_2^T \odot \sigma'(\cdot)]^T(\boldsymbol{x}d\boldsymbol{W}_1 + d\boldsymbol{b}_1)) \tag{41}$$
$$\frac{\partial J}{\partial \boldsymbol{W}_1} = \{[(\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{W}_2^T \odot \sigma'(\cdot)]^T \boldsymbol{x}\}^T \tag{42}$$
$$= \boldsymbol{x}^T[(\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{W}_2^T \odot \sigma'(\cdot)] \tag{43}$$
$$\frac{\partial J}{\partial \boldsymbol{b}_1} = [(\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{W}_2^T \odot \sigma'(\cdot)] \tag{44}$$
$$\tag{45}$$