Assignment 1

Chen Luo

March 14, 2018

1 Softmax

1.1 a

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

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

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

$$
\begin{aligned}
CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= -\sum_i y_i \log(\hat{y}_i) && (7)\\
\hat{\boldsymbol{y}} &= softmax(\boldsymbol{\theta}) = \frac{e^{\boldsymbol{\theta}}}{\sum_j e^{\theta_j}} && (8)\\
\frac{\partial CE(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \boldsymbol{\theta}} &= -\frac{\partial \sum_i 1_{[y_i==1]} \left\{ \log e^{\theta_i} - \log \sum_j e^{\theta_j} \right\}}{\partial \boldsymbol{\theta}} && (9)\\
&= -\frac{\partial \boldsymbol{y}^T \boldsymbol{\theta} - log \sum_j e^{\theta_j}}{\partial \boldsymbol{\theta}} && (10)\\
&= -\boldsymbol{y} + \frac{\log \mathbf{1}^T e^{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} && (11)\\
&= -\boldsymbol{y} + \frac{e^{\boldsymbol{\theta}}}{\mathbf{1}^T e^{\boldsymbol{\theta}}} && (12)\\
&= -\boldsymbol{y} + softmax(\boldsymbol{\theta}) && (13)\\
&= \hat{\boldsymbol{y}} - \boldsymbol{y} && (14)
\end{aligned}
$$

2.3 c

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

$$
\begin{aligned}
\frac{\partial J}{\partial \boldsymbol{x}} &= \frac{\partial J}{\partial(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)} \frac{\partial(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}} && (17)\\
&= (\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{W}_2^T \sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\{\mathbf{1} - \sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\}^T \boldsymbol{W}_1^T && (18)
\end{aligned}
$$

2.4 d

$$
\begin{aligned}
\boldsymbol{W}_1 &: D_x \times H && (19)\\
\boldsymbol{b}_1 &: 1 \times H && (20)\\
\boldsymbol{W}_2 &: H \times D_y && (21)\\
\boldsymbol{b}_2 &: 1 \times D_y && (22)\\
& H(D_x + D_y + 1) + D_y && (23)
\end{aligned}
$$