

Assignment 1

Chen Luo

March 20, 2018

1 Softmax

1.1 a

$$\text{softmax}(\mathbf{x} + c) = \frac{e^{\mathbf{x}+c}}{\sum_{j=1}^d e^{x_j+c}} \quad (1)$$

$$= \frac{e^{\mathbf{x}} e^c}{e^c \sum_{j=1}^d e^{x_j}} = \text{softmax}(\mathbf{x}) \quad (2)$$

2 NN Basic

2.1 a

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

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

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

$$= \sigma(x) - \sigma(x)^2 \quad (6)$$

2.2 b

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

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

$$\frac{\partial CE(\mathbf{y}, \hat{\mathbf{y}})}{\partial \boldsymbol{\theta}} = - \frac{\partial \sum_i 1_{[y_i=1]} \left\{ \log e^{\theta_i} - \log \sum_j e^{\theta_j} \right\}}{\partial \boldsymbol{\theta}} \quad (9)$$

$$= - \frac{\partial \mathbf{y}^T \boldsymbol{\theta} - \log \sum_j e^{\theta_j}}{\partial \boldsymbol{\theta}} \quad (10)$$

$$= -\mathbf{y} + \frac{\log \mathbf{1}^T e^{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \quad (11)$$

$$= -\mathbf{y} + \frac{e^{\boldsymbol{\theta}}}{\mathbf{1}^T e^{\boldsymbol{\theta}}} \quad (12)$$

$$= -\mathbf{y} + \text{softmax}(\boldsymbol{\theta}) \quad (13)$$

$$= \hat{\mathbf{y}} - \mathbf{y} \quad (14)$$

2.3 c

$$J = CE(\mathbf{y}, \hat{\mathbf{y}}) = CE(\mathbf{y}, \text{softmax}(\mathbf{h}\mathbf{W}_2 + \mathbf{b}_2)) \quad (15)$$

$$= CE(\mathbf{y}, \text{softmax}(\sigma(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2)) \quad (16)$$

Actually,

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

$$\frac{\partial J}{\partial \mathbf{x}} = \frac{\partial J}{\partial (\mathbf{h}\mathbf{W}_2 + \mathbf{b}_2)} \frac{\partial (\mathbf{h}\mathbf{W}_2 + \mathbf{b}_2)}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \quad (18)$$

$$= (\hat{\mathbf{y}} - \mathbf{y}) \mathbf{W}_2^T \sigma(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1) \{ \mathbf{1} - \sigma(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1) \}^T \mathbf{W}_1^T \quad (19)$$

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

Or, [TODO]

$$d(\mathbf{h}\mathbf{W}_2 + \mathbf{b}_2) = d\mathbf{h}\mathbf{W}_2 + \mathbf{h}d\mathbf{W}_2 + d\mathbf{b}_2 \quad (20)$$

$$\text{vec}(d(\mathbf{h}\mathbf{W}_2 + \mathbf{b}_2)) = \text{vec}(d\mathbf{h}\mathbf{W}_2 + \mathbf{h}d\mathbf{W}_2 + d\mathbf{b}_2) \quad (21)$$

$$= \text{vec}(\mathbf{1}d\mathbf{h}\mathbf{W}_2 + \mathbf{h}d\mathbf{W}_2\mathbf{I} + \mathbf{1}d\mathbf{b}_2\mathbf{I}) \rightarrow \text{preparation for} \quad (22)$$

$$\text{vec}(AXB) = (B^T \otimes A)\text{vec}(X) \quad (23)$$

$$= \mathbf{W}_2^T \otimes \mathbf{1} \text{vec}(d\mathbf{h}) + \mathbf{I} \otimes \mathbf{h} \text{vec}(d\mathbf{W}_2) + \mathbf{I} \otimes \mathbf{1} \text{vec}(d\mathbf{b}_2) \quad (24)$$

2.4 d

$$\mathbf{W}_1 : D_x \times H \quad (25)$$

$$\mathbf{b}_1 : 1 \times H \quad (26)$$

$$\mathbf{W}_2 : H \times D_y \quad (27)$$

$$\mathbf{b}_2 : 1 \times D_y \quad (28)$$

$$H(D_x + D_y + 1) + D_y \quad (29)$$

2.5 g

$$dJ = (\hat{\mathbf{y}} - \mathbf{y})^T d(\mathbf{h}\mathbf{W}_2 + \mathbf{b}_2) = (\hat{\mathbf{y}} - \mathbf{y})^T (d\mathbf{h}\mathbf{W}_2 + \mathbf{h}d\mathbf{W}_2 + d\mathbf{b}_2) \quad (30)$$

$$= (\hat{\mathbf{y}} - \mathbf{y})^T d\mathbf{h}\mathbf{W}_2 + (\hat{\mathbf{y}} - \mathbf{y})^T \mathbf{h}d\mathbf{W}_2 + (\hat{\mathbf{y}} - \mathbf{y})^T d\mathbf{b}_2 \quad (31)$$

$$= \text{trace}((\hat{\mathbf{y}} - \mathbf{y})^T d\mathbf{h}\mathbf{W}_2 + (\hat{\mathbf{y}} - \mathbf{y})^T \mathbf{h}d\mathbf{W}_2 + (\hat{\mathbf{y}} - \mathbf{y})^T d\mathbf{b}_2)) \quad (32)$$

$$\frac{\partial J}{\partial \mathbf{W}_2} = (\mathbf{W}_2(\hat{\mathbf{y}} - \mathbf{y})^T)^T \quad (33)$$

$$= \mathbf{h}^T (\hat{\mathbf{y}} - \mathbf{y}) \quad (34)$$

$$\frac{\partial J}{\partial \mathbf{b}_2} = \hat{\mathbf{y}} - \mathbf{y} \quad (35)$$

$$d\mathbf{h} = d(\sigma(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)) = \sigma'(\cdot) \odot (\mathbf{x}d\mathbf{W}_1 + d\mathbf{b}_1) \quad (36)$$

$$dJ = \text{trace}((\hat{\mathbf{y}} - \mathbf{y})^T \sigma'(\cdot) \odot (\mathbf{x}d\mathbf{W}_1 + d\mathbf{b}_1)\mathbf{W}_2) \quad (37)$$

$$= \text{trace}(\mathbf{W}_2(\hat{\mathbf{y}} - \mathbf{y})^T \sigma'(\cdot) \odot (\mathbf{x}d\mathbf{W}_1 + d\mathbf{b}_1)) \quad (38)$$

$$\text{Apply law of trace and element-wise product :} \quad (39)$$

$$\text{trace}(A^T(B \odot C)) = \text{trace}((A \odot B)^T C) \quad (40)$$

$$= \text{trace}([(\hat{\mathbf{y}} - \mathbf{y})\mathbf{W}_2^T \odot \sigma'(\cdot)]^T (\mathbf{x}d\mathbf{W}_1 + d\mathbf{b}_1)) \quad (41)$$

$$\frac{\partial J}{\partial \mathbf{W}_1} = \{[(\hat{\mathbf{y}} - \mathbf{y})\mathbf{W}_2^T \odot \sigma'(\cdot)]^T \mathbf{x}\}^T \quad (42)$$

$$= \mathbf{x}^T [(\hat{\mathbf{y}} - \mathbf{y})\mathbf{W}_2^T \odot \sigma'(\cdot)] \quad (43)$$

$$\frac{\partial J}{\partial \mathbf{b}_1} = [(\hat{\mathbf{y}} - \mathbf{y})\mathbf{W}_2^T \odot \sigma'(\cdot)] \quad (44)$$

3 word2vec

Notice: $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ are column vectors. Previously, all vectors are row vectors.

3.1 a

$$J_{softmax-CE}(o, \boldsymbol{\nu}_c, \mathbf{U}) = CE(\mathbf{y}, \hat{\mathbf{y}}) \quad (45)$$

$$\hat{\mathbf{y}} = softmax(\{\mathbf{U}_{\{d \times V\}}^T \times \boldsymbol{\nu}_{c\{d \times 1\}}\}^T) \quad (46)$$

$$dCE(\cdot) = (\hat{\mathbf{y}} - \mathbf{y}) d\boldsymbol{\nu}_c^T \mathbf{U} \quad (47)$$

$$= trace[(\hat{\mathbf{y}} - \mathbf{y}) d\boldsymbol{\nu}_c^T \mathbf{U}] \quad (48)$$

$$= trace[\mathbf{U}(\hat{\mathbf{y}} - \mathbf{y}) d\boldsymbol{\nu}_c^T] \quad (49)$$

$$\frac{\partial CE(\cdot)}{\partial \boldsymbol{\nu}_c^T} = (\hat{\mathbf{y}} - \mathbf{y})^T \mathbf{U}^T \quad (50)$$

$$\frac{\partial CE(\cdot)}{\partial \boldsymbol{\nu}_c} = \mathbf{U}(\hat{\mathbf{y}} - \mathbf{y}) \quad (51)$$

3.2 b

$$dCE(\cdot) = (\hat{\mathbf{y}} - \mathbf{y}) \boldsymbol{\nu}_c^T d\mathbf{U} \quad (52)$$

$$= trace[(\hat{\mathbf{y}} - \mathbf{y}) \boldsymbol{\nu}_c^T d\mathbf{U}] \quad (53)$$

$$\frac{\partial CE}{\partial \mathbf{U}} = \boldsymbol{\nu}_c (\hat{\mathbf{y}} - \mathbf{y})^T \quad (54)$$

3.3 c

Assume that for $\boldsymbol{\nu}_c$, \mathbf{k}_{neg} is a row vector of K -hot (ensure $\mathbf{k}_{neg\{o\}} \neq 1$) and \mathbf{k} is one-hot vector that $\mathbf{k}_o = 1$. Rewrite $J_{neg-sample}(\cdot)$ as follow,

$$J_{neg-sample}(\cdot) = -\log(\sigma(\boldsymbol{\nu}_c^T \mathbf{U})) \mathbf{k}^T - \log(\sigma(-\boldsymbol{\nu}_c^T \mathbf{U})) \mathbf{k}_{neg}^T \quad (55)$$

$$= -\log(\sigma(\boldsymbol{\nu}_c^T \mathbf{U} \mathbf{k}^T)) - \log(\sigma(-\boldsymbol{\nu}_c^T \mathbf{U} \mathbf{k}_{neg}^T)) \quad (56)$$

In the last equation, we put all vectors into element-wise operator.

$$dJ_{neg-sample}(\cdot) = \frac{\sigma(\boldsymbol{\nu}_c^T \mathbf{U} \mathbf{k}^T)(1 - \sigma(\boldsymbol{\nu}_c^T \mathbf{U} \mathbf{k}^T))}{\sigma(\boldsymbol{\nu}_c^T \mathbf{U} \mathbf{k}^T)} \quad (57)$$

$$\frac{\partial J_{neg-sample}(\cdot)}{\partial \boldsymbol{\nu}_c} = \quad (58)$$

TODO

$$\frac{\partial J}{\partial \boldsymbol{\nu}_c} = -\boldsymbol{\mu}_o \frac{\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)(1 - \sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c))}{\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)} + \sum_{k=1}^K \boldsymbol{\mu}_k (1 - \sigma(-\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c)) \quad (59)$$

$$= -\boldsymbol{\mu}_o (1 - \sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)) + \sum_{k=1}^K \boldsymbol{\mu}_k (1 - (1 - \sigma(\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c))) \quad (60)$$

$$= -\boldsymbol{\mu}_o (1 - \sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)) + \sum_{k=1}^K \boldsymbol{\mu}_k (\sigma(\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c)) \quad (61)$$

$$dJ(\cdot) = -\frac{d(\sigma(\cdot))}{\sigma(\cdot)} + \sum_{k=1}^K (1 - \sigma(-\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c)) d\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c \quad (62)$$

$$= -\text{trace}((1 - \sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)) d\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c) \quad (63)$$

$$= \text{trace}(\boldsymbol{\nu}_c (\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c) - 1) d\boldsymbol{\mu}_o^T) \quad (64)$$

$$\frac{\partial J}{\partial \boldsymbol{\mu}_o} = \boldsymbol{\nu}_c (\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c) - 1) \quad (65)$$

$$(66)$$

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = \boldsymbol{\nu}_c \sigma(\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c), \text{ for } k = [1, 2, \dots, K] \quad (67)$$

Computation efficiency of negative-sampling loss, speed up ratio is approximately

$$\frac{V}{K+1}$$