Solutions to assignment1 of CS224n

陈承勃

May 22, 2018

1 Softmax

(a) Omitted. (b) See **q1_softmax.py**

2 Neural Network Basics

(a) $\sigma'(x) = \sigma(x)\sigma(1-x)$

(b) Assume k is the correct class, then

$CE(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = -y_k \log \widehat{y_k} = -\log \widehat{y_k} = -\log \frac{\exp(\boldsymbol{\theta}_k)}{\sum_i \exp(\boldsymbol{\theta}_i)} = -\boldsymbol{\theta}_k + \log \sum_i \exp(\boldsymbol{\theta}_i)$.

$\therefore \frac{\partial CE(\boldsymbol{y}, \widehat{\boldsymbol{y}})}{\partial \boldsymbol{\theta}_k} = -1 + \frac{\exp(\boldsymbol{\theta}_k)}{\sum_i \exp(\boldsymbol{\theta}_i)} = \widehat{y_k} - 1$,

$\frac{\partial CE(\boldsymbol{y}, \widehat{\boldsymbol{y}})}{\partial \boldsymbol{\theta}_j} = \frac{\exp(\boldsymbol{\theta}_j)}{\sum_i \exp(\boldsymbol{\theta}_i)} = \widehat{y_j},\ j \neq k$.

$\therefore \frac{\partial CE(\boldsymbol{y}, \widehat{\boldsymbol{y}})}{\partial \boldsymbol{\theta}} = \widehat{\boldsymbol{y}} - \boldsymbol{y}$

(c) The forward propagation steps:

$\boldsymbol{Z_1} = \boldsymbol{x}\boldsymbol{W_1} + \boldsymbol{b_1}, \quad \boldsymbol{h} = sigmoid(\boldsymbol{Z_1})$

$\boldsymbol{Z_2} = \boldsymbol{h}\boldsymbol{W_2} + \boldsymbol{b_2}, \quad \widehat{\boldsymbol{y}} = sigmoid(\boldsymbol{Z_2})$

$\boldsymbol{J} = CE(\boldsymbol{y}, \widehat{\boldsymbol{y}})$

The backward propagation:

$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{Z_2}} = \widehat{\boldsymbol{y}} - \boldsymbol{y} \triangleq \boldsymbol{\delta_1}, \quad \frac{\partial \boldsymbol{J}}{\partial \boldsymbol{h}} = \boldsymbol{\delta_1}\boldsymbol{W_2^{\mathrm{T}}} \triangleq \boldsymbol{\delta_2}$

$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{Z_1}} = \boldsymbol{\delta_2} * \sigma'(\boldsymbol{Z_1}) \triangleq \boldsymbol{\delta_3}$, $*$ denotes element-wise product.

$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{x}} = \boldsymbol{\delta_3}\boldsymbol{W_1^{\mathrm{T}}}$

(d) $(1 + D_x) \times H + (1 + H) \times D_y$

(e) See **q2_sigmoid.py**

(f) See **q2_gradcheck.py**

(g) See **q2_neural.py**

3 word2vec

(a) $J_{softmax\text{-}CE}(o, \boldsymbol{v}_c, \boldsymbol{U}) = CE(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = -\sum_i yi \log(\widehat{y_i}) = -\log \widehat{y}_o = -\log \frac{\exp(\boldsymbol{u}_o^{\mathrm{T}}\boldsymbol{v}_c)}{\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}}\boldsymbol{V}_c)}$

$= -\boldsymbol{u}_o^{\mathrm{T}}\boldsymbol{v}_c + \log \sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}}\boldsymbol{v}_c)$

$\therefore \frac{\partial J}{\partial \boldsymbol{v}_c} = -\boldsymbol{u_o} + \frac{\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{v}_c)\boldsymbol{u_w}}{\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{v}_c)} = -\boldsymbol{u}_o + \sum_{w=1}^{V} \frac{\exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{v}_c)}{\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{v}_c)} \boldsymbol{u}_w = -\boldsymbol{u}_o + \sum_{w=1}^{V} \widehat{y_w} \boldsymbol{u}_w$

$\therefore \frac{\partial J}{\partial \boldsymbol{v}_c} = \sum_{w=1}^{V} \widehat{y_w} \boldsymbol{u}_w - \sum_{w=1}^{V} y_w \boldsymbol{u}_o = \boldsymbol{U}(\widehat{\boldsymbol{y}} - \boldsymbol{y})$

(b) $\frac{\partial J}{\partial \boldsymbol{u}_o} = -\boldsymbol{v}_c + \frac{\exp(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c)\boldsymbol{v}_c}{\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{V}_c)} = (\widehat{\boldsymbol{y}}_o - 1)\boldsymbol{v}_c$

$\frac{\partial J}{\partial \boldsymbol{u}_k} = \frac{\exp(\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_c)\boldsymbol{v}_c}{\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{V}_c)} = \widehat{\boldsymbol{y}}_k \boldsymbol{v}_c, \ for \ k \neq o.$

$\therefore \frac{\partial J}{\partial \boldsymbol{U}} = \boldsymbol{v}_c(\widehat{\boldsymbol{y}} - \boldsymbol{y})^{\mathrm{T}}, \ or \ \frac{\partial J}{\partial \boldsymbol{u}_k} = \begin{cases} (\widehat{\boldsymbol{y}}_o - 1)\boldsymbol{v}_c & k = o \\ \widehat{\boldsymbol{y}}_k \boldsymbol{v}_c & k \neq o \end{cases}$

(c) $\frac{\partial J}{\partial \boldsymbol{v}_c} = -\frac{1}{\sigma(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c)} \sigma(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c)(1 - \sigma(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c))\boldsymbol{u}_o + \sum_{k=1}^{K} \frac{1}{\sigma(-\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_c)} \sigma(-\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_c)(1 - \sigma(-\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_c))\boldsymbol{u}_k$

$= (\sigma(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c) - 1)\boldsymbol{u}_o - \sum_{k=1}^{K} (\sigma(-\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_c) - 1)\boldsymbol{u}_k$

$\frac{\partial J}{\partial \boldsymbol{u}_o} = (\sigma(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c) - 1)\boldsymbol{v}_c$

$\frac{\partial J}{\partial \boldsymbol{u}_k} = -(\sigma(-\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_c) - 1)\boldsymbol{v}_c$

(d) Let U be the collection of all output vectors for all words in the vocabulary.

For Skip-Gram model,

$\frac{\partial J_{skip\text{-}gram}}{\partial \boldsymbol{v}_c} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial \boldsymbol{F}(w_{t+j}, \boldsymbol{v}_c)}{\partial \boldsymbol{v}_c}, \quad \frac{\partial J_{skip\text{-}gram}}{\partial \boldsymbol{U}} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial \boldsymbol{F}(w_{t+j}, \boldsymbol{v}_c)}{\partial \boldsymbol{U}}$

For CBOW model,

$\frac{\partial J_{CBOW}}{\partial \boldsymbol{v}_{w_{t+j}}} = \frac{\partial \boldsymbol{F}(w_t, \widehat{\boldsymbol{v}})}{\partial \widehat{\boldsymbol{v}}}, \ for \ all \ j \in \{-m, ..., -1, 1, ...m\}$

$\frac{\partial J_{CBOW}}{\partial \boldsymbol{v}_{w_{t+j}}} = 0, \ for \ all \ j \notin \{-m, ..., -1, 1, ...m\}$

$\frac{\partial J_{CBOW}}{\partial \boldsymbol{U}} = \frac{\partial \boldsymbol{F}(w_t, \widehat{\boldsymbol{v}})}{\partial \boldsymbol{U}}$