Mikhail Sushkov
msushkov

Homework 1

Problem 1

Given a D-dimensional vector **x** ($\mathbf{x}_k$ is the k-th component of **x**, and 1 <= j <= D):

$$softmax(\boldsymbol{x})_j = \frac{e^{x_j}}{\sum_{k=1}^{D} e^{x_k}}.$$

Then,

$$softmax(\boldsymbol{x}+c)_j = \frac{e^{x_j+c}}{\sum_{k=1}^{D} e^{x_k+c}} = \frac{e^c e^{x_j}}{\sum_{k=1}^{D} e^c e^{x_k}} = \frac{e^c e^{x_j}}{e^c \sum_{k=1}^{D} e^{x_k}} = \frac{e^{x_j}}{\sum_{k=1}^{D} e^{x_k}} = softmax(\boldsymbol{x})_j$$

Since we have shown that for all components j $softmax(x+c)_j = softmax(x)_j$, then it follows that softmax(x+c) = softmax(x).

Mikhail Sushkov
msushkov

Problem 2

a) $\sigma(x) = \frac{1}{1+e^{-x}} = (1+e^{-x})^{-1}$

$$\frac{\partial}{\partial x}\sigma(x) = -(1+e^{-x})^{-2} * (e^{-x} * -1) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} * \frac{e^{-x}}{1+e^{-x}} = \frac{1}{1+e^{-x}} * \left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right)$$
$$= \sigma(x) * \big(1 - \sigma(x)\big)$$

$$\boldsymbol{\sigma'^{(x)} = \sigma(x) * \big(1 - \sigma(x)\big)}$$

b) $\frac{\partial}{\partial \boldsymbol{\theta}} CE = \frac{\partial}{\partial \boldsymbol{\theta}}\left[-\sum_i y_i * \log \frac{e^{\boldsymbol{\theta}_i}}{\sum_{k=1}^{D} e^{\boldsymbol{\theta}_k}}\right] = \frac{\partial}{\partial \boldsymbol{\theta}}\left[-\sum_i y_i * \left(\boldsymbol{\theta}_i - \log \sum_{k=1}^{D} e^{\boldsymbol{\theta}_k}\right)\right]$

When j = i:

$$\frac{\partial}{\partial \boldsymbol{\theta}} CE_j = -\left(y_i * 1 - \frac{1}{\sum_{k=1}^{D} e^{\boldsymbol{\theta}_k}} * e^{\boldsymbol{\theta}_i}\right) = y_i * \left(\frac{e^{\boldsymbol{\theta}_i}}{\sum_{k=1}^{D} e^{\boldsymbol{\theta}_k}} - 1\right) = y_i * (\hat{\boldsymbol{y}}_i - 1)$$

$y_i = 1$ since y is the 1-hot label, so:

$\frac{\partial}{\partial \boldsymbol{\theta}} CE_j = \hat{\boldsymbol{y}}_j - 1$ when j = i.

When j ≠ i:

$$\frac{\partial}{\partial \boldsymbol{\theta}} CE_j = -\left(y_i * 0 - \frac{1}{\sum_{k=1}^{D} e^{\boldsymbol{\theta}_k}} * e^{\boldsymbol{\theta}_j}\right) = \frac{e^{\boldsymbol{\theta}_j}}{\sum_{k=1}^{D} e^{\boldsymbol{\theta}_k}} = \hat{\boldsymbol{y}}_j$$

Thus,

$$\boldsymbol{\frac{\partial}{\partial \boldsymbol{\theta}} CE = \hat{y} - y.}$$

c) Let's define $\theta = hW_2 + b_2$, $\hat{y} = softmax(\theta)$, $h = sigmoid(a)$, and $a = xW_1 + b_1$.
Using the chain rule:

$$\frac{\partial J}{\partial x} = \left(\left(\frac{\partial J}{\partial \theta} * \frac{\partial \theta}{\partial h}\right) \circ \frac{\partial h}{\partial a}\right) * \frac{\partial a}{\partial x}$$

Here ∘ represents element-wise product and * represents dot product. We take the element-wise product with $\frac{\partial h}{\partial a}$ because h is an element-wise function.

From (b) we know that for cross-entropy cost J, $\frac{\partial J}{\partial \theta} = \hat{y} - y$.

$$\frac{\partial \theta}{\partial h} = W_2^T$$

$$\frac{\partial h}{\partial a} = h \circ (1 - h)$$

$$\frac{\partial a}{\partial x} = W_1^T$$

Thus,

$$\boldsymbol{\frac{\partial J}{\partial x} = \left((\hat{y} - y) * W_2^T\right) \circ h \circ (1 - h) * W_1^T.}$$

We know that, since J is a scalar, $\frac{\partial J}{\partial x}$ must have the same dimensions as x. Let's do some dimensional analysis on our answer above to make sure this is the case.

$\hat{y} - y$ has dimensions 1 x D_y

W_2 has dimensions H x D_y

h and 1 - h both have dimensions 1 x H

W_1 has dimensions D_x x H

Thus $\frac{\partial J}{\partial x}$ has dimensions (1 x D_y) * (D_y x H) ∘ (1 x H) * (H x D_x) = (1 x H) * (H x D_x) = 1 x D_x, which is the same dimensions as x.

d) The parameters are $W_1\ (D_x by\ H)$, $W_2\ (H\ by\ D_y)$, $b_1\ (1\ by\ H)$, $b_2\ (1\ by\ D_y)$.

$$\boxed{Total\ \#\ of\ params = HD_x + HD_y + H + D_y}$$

Mikhail Sushkov
msushkov

Problem 3

a) We are assuming cross-entropy cost, so for a given predicted word at index i:

$$J = -\log P(word_i \mid \hat{\boldsymbol{r}}, \boldsymbol{w}) = -\log \frac{\exp(\boldsymbol{w_i^T}\hat{\boldsymbol{r}})}{\sum_{j=1}^{|V|}\exp(\boldsymbol{w_j^T}\hat{\boldsymbol{r}})} = -\boldsymbol{w_i^T}\hat{\boldsymbol{r}} + \log\sum_{j=1}^{|V|}\exp(\boldsymbol{w_j^T}\hat{\boldsymbol{r}})$$

$$\frac{\partial J}{\partial \hat{\boldsymbol{r}}} = -\hat{\boldsymbol{r}} + \frac{1}{\sum_{j=1}^{|V|}\exp(\boldsymbol{w_j^T}\hat{\boldsymbol{r}})} * \sum_{k=1}^{|V|}\exp(\boldsymbol{w_k^T}\hat{\boldsymbol{r}}) * \boldsymbol{w_k} = -\hat{\boldsymbol{r}} + \sum_{k=1}^{|V|}\frac{\exp(\boldsymbol{w_k^T}\hat{\boldsymbol{r}}) * \boldsymbol{w_k}}{\sum_{j=1}^{|V|}\exp(\boldsymbol{w_j^T}\hat{\boldsymbol{r}})} =$$
$$= -\hat{\boldsymbol{r}} + \sum_{k=1}^{|V|} P(word_k \mid \hat{\boldsymbol{r}}, \boldsymbol{w}) * \boldsymbol{w_k}$$

$$\boxed{\frac{\partial J}{\partial \hat{\boldsymbol{r}}} = -\boldsymbol{w_i} + \sum_{k=1}^{|V|} P(word_k \mid \hat{\boldsymbol{r}}, \boldsymbol{w}) * \boldsymbol{w_k}}$$

b) $J = -\log P(word_i \mid \hat{\boldsymbol{r}}, \boldsymbol{w}) = -\log \frac{\exp(\boldsymbol{w_i^T}\hat{\boldsymbol{r}})}{\sum_{j=1}^{|V|}\exp(\boldsymbol{w_j^T}\hat{\boldsymbol{r}})} = -\boldsymbol{w_i^T}\hat{\boldsymbol{r}} + \log\sum_{j=1}^{|V|}\exp(\boldsymbol{w_j^T}\hat{\boldsymbol{r}})$

When j = i:

$$\frac{\partial J}{\partial \boldsymbol{w}_j} = -\hat{\boldsymbol{r}} + \frac{1}{\sum_{k=1}^{|V|}\exp(\boldsymbol{w_k^T}\hat{\boldsymbol{r}})} * \exp(\boldsymbol{w_j^T}\hat{\boldsymbol{r}}) * \hat{\boldsymbol{r}}$$

When j ≠ i:

$$\frac{\partial J}{\partial \boldsymbol{w}_j} = \frac{1}{\sum_{k=1}^{|V|}\exp(\boldsymbol{w_k^T}\hat{\boldsymbol{r}})} * \exp(\boldsymbol{w_j^T}\hat{\boldsymbol{r}}) * \hat{\boldsymbol{r}}$$

We have:

$\frac{\partial J}{\partial \boldsymbol{w}_j} = -\hat{\boldsymbol{r}} + \frac{\exp(\boldsymbol{w_j^T}\hat{\boldsymbol{r}}) * \hat{\boldsymbol{r}}}{\sum_{k=1}^{|V|}\exp(\boldsymbol{w_k^T}\hat{\boldsymbol{r}})}$ when j = i

$\frac{\partial J}{\partial \boldsymbol{w}_j} = \frac{\exp(\boldsymbol{w_j^T}\hat{\boldsymbol{r}}) * \hat{\boldsymbol{r}}}{\sum_{k=1}^{|V|}\exp(\boldsymbol{w_k^T}\hat{\boldsymbol{r}})}$ when j ≠ i

c)

$$J = -\log\sigma(\boldsymbol{w_i^T\hat{r}}) - \sum_{k=1}^{K}\log\sigma(-\boldsymbol{w_k^T\hat{r}})$$

$$\frac{\partial J}{\partial \boldsymbol{\hat{r}}} = -\frac{1}{\sigma(\boldsymbol{w_i^T\hat{r}})} * \sigma(\boldsymbol{w_i^T\hat{r}}) * \left(1 - \sigma(\boldsymbol{w_i^T\hat{r}})\right) * \boldsymbol{w_i} - \sum_{k=1}^{K}\frac{1}{\sigma(-\boldsymbol{w_k^T\hat{r}})} * \sigma(-\boldsymbol{w_k^T\hat{r}}) * \left(1 - \sigma(-\boldsymbol{w_k^T\hat{r}})\right)(-\boldsymbol{w_k})$$

$$\frac{\partial J}{\partial \boldsymbol{\hat{r}}} = \boldsymbol{w_i} * \left(\sigma(\boldsymbol{w_i^T\hat{r}}) - 1\right) + \sum_{k=1}^{K}\boldsymbol{w_k}\left(1 - \sigma(-\boldsymbol{w_k^T\hat{r}})\right)$$

When j = i:

$$\frac{\partial J}{\partial \boldsymbol{w_j}} = -\frac{1}{\sigma(\boldsymbol{w_j^T\hat{r}})} * \sigma(\boldsymbol{w_j^T\hat{r}})\left(1 - \sigma(\boldsymbol{w_j^T\hat{r}})\right)\boldsymbol{\hat{r}}$$

$$\frac{\partial J}{\partial \boldsymbol{w_j}} = \boldsymbol{\hat{r}}\left(\sigma(\boldsymbol{w_j^T\hat{r}}) - 1\right)$$

When j ≠ i:

$$\frac{\partial J}{\partial \boldsymbol{w_j}} = -\frac{1}{\sigma(-\boldsymbol{w_j^T\hat{r}})} * \sigma(-\boldsymbol{w_j^T\hat{r}}) * \left(1 - \sigma(-\boldsymbol{w_j^T\hat{r}})\right)(-\boldsymbol{\hat{r}})$$

$$\frac{\partial J}{\partial \boldsymbol{w_j}} = -\boldsymbol{\hat{r}}\left(\sigma(-\boldsymbol{w_j^T\hat{r}}) - 1\right)$$

We have:

$$\frac{\partial J}{\partial \boldsymbol{\hat{r}}} = \boldsymbol{w_i} * \left(\sigma(\boldsymbol{w_i^T\hat{r}}) - 1\right) + \sum_{k=1}^{K}\boldsymbol{w_k}\left(1 - \sigma(-\boldsymbol{w_k^T\hat{r}})\right)$$

$\frac{\partial J}{\partial w_j} = \boldsymbol{\hat{r}}\left(\sigma(\boldsymbol{w_j^T\hat{r}}) - 1\right)$ when j = i

$\frac{\partial J}{\partial w_j} = -\boldsymbol{\hat{r}}\left(\sigma(-\boldsymbol{w_j^T\hat{r}}) - 1\right)$ when j ≠i, j is in {1, …, K}

$\frac{\partial J}{\partial w_j} = 0$ when j ≠ i and j is not in {1, …, K}

d)

$$J_{skip-gram} = \sum_{-c \le j \le c, j \neq 0} F(v'_{w_{i+j}}, v_{wi})$$

Given the previous parts we know how to compute the gradient of F with respect to any word vector.

$$\frac{\partial J}{\partial v_{w_i}} = \sum_{-c \le j \le c, j \neq 0} \frac{\partial F\left(v'_{w_{i+j}}, v_{w_i}\right)}{\partial v_{w_i}}$$

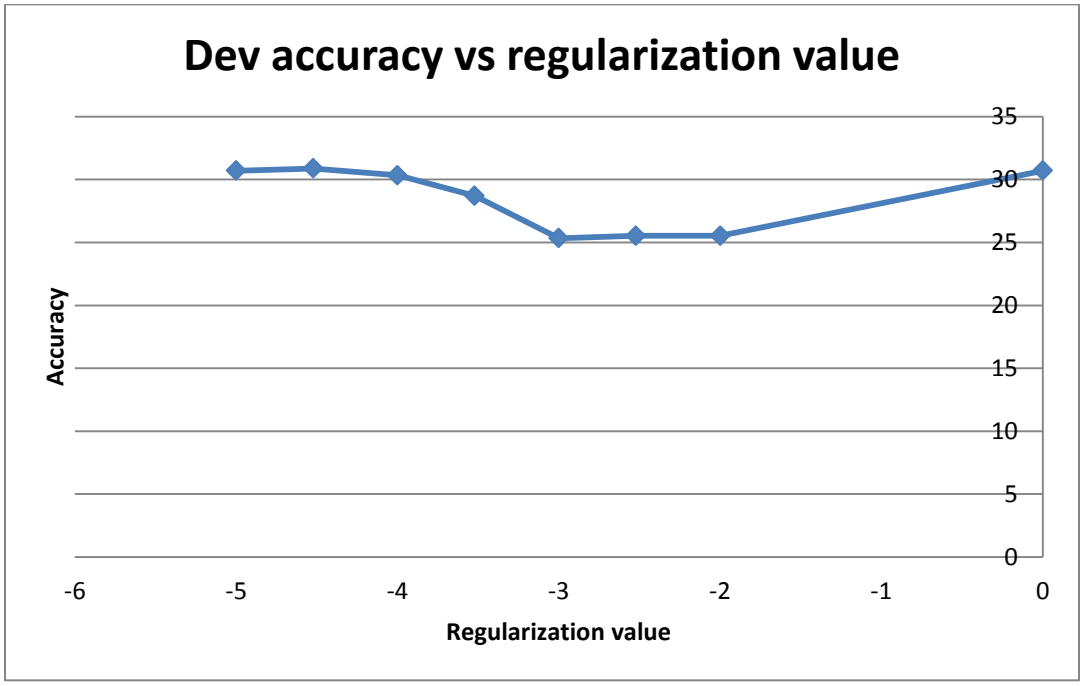
$$\frac{\partial J}{\partial v'_{w_{i+k}}} = \frac{\partial F\left(v'_{w_{i+k}}, v_{w_i}\right)}{\partial v'_{w_{i+k}}} \text{ for } -c \le k \le c, k \neq 0$$

Mikhail Sushkov
msushkov

Problem 4

a) We want to introduce regularization because it helps to prevent overfitting. Regularization penalizes more highly parameters with larger values, so it forces the parameters to be closer to 0. This reduces model complexity and helps to reduce overfitting.

b)



We see that as the regularization value increases, the dev accuracy first slightly increases, then decreases significantly, then increases again. (we note that the rightmost point of the curve corresponds to log 0 which we set to be 0 here). We achieve the best dev accuracy with a regularization value of 0.0003, after which the accuracy drops significantly but then again starts to increase, though to values smaller than the best accuracy.