CS224d Assignment 1

Denny Britz

October 3, 2015

1 Softmax

$$
\begin{aligned}
\mathrm{softmax}(\mathbf{x}+c) &= \frac{e^{x_i+c}}{\sum_{k=1}^{K} e^{x_k+c}} \quad \text{for } j=1,...,K && (1)\\
&= \frac{e^{x_i}e^{c}}{\sum_{k=1}^{K} e^{x_k}e^{c}} && (2)\\
&= \frac{e^{x_i}e^{c}}{e^{c}\sum_{k=1}^{K} e^{x_k}} && (3)\\
&= \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}} \quad \text{for } j=1,...,K && (4)\\
&= \mathrm{softmax}(\mathbf{x}) && (5)
\end{aligned}
$$

2 Neural Network Basics

2.1 (a)

The sigmoid function is defined as $f(x) = \frac{1}{1+e^{-x}}$. Let $g(x) = 1 + e^{-x}$.

$$\begin{aligned}
\frac{df}{dx} &= \frac{df}{dg}\frac{dg}{dx} && (6)\\
&= -\frac{1}{(1+e^{-x})^2}(-e^{-x}) && (7)\\
&= \left(\frac{1}{1+e^{-x}}\right)^2 e^{-x} && (8)\\
&= e^{-x} f^2(x) && (9)\\
&= \left(\frac{1}{f(x)} - 1\right) f^2(x) && (10)\\
&= f(x)(1 - f(x)) && (11)
\end{aligned}$$

2.2 (b)

We know that only the kth dimension of $\mathbf{y}$ is 1 and so the cross entropy loss simplifies to $CE(\mathbf{y}, \hat{\mathbf{y}}) = -\log(\hat{\mathbf{y}})$ where $\hat{\mathbf{y}}_k = \frac{e^{\theta_k}}{\sum_i e^{\theta_i}}$ (the softmax).

$$\begin{aligned}
\frac{\partial CE}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}}\left(-\log \frac{e^{\theta_k}}{\sum_i e^{\theta_i}})\right) && (12)\\
&= \frac{\partial}{\partial \boldsymbol{\theta}}(-\boldsymbol{\theta}_k) + \frac{\partial}{\partial \boldsymbol{\theta}}\left(\log \sum_i e^{\theta_i}\right) && (13)\\
&= \frac{\partial}{\partial \boldsymbol{\theta}}(-\boldsymbol{\theta}_k) + \frac{\sum_i \frac{\partial e^{\theta_i}}{\partial \boldsymbol{\theta}}}{\sum_i e^{\theta_i}} && (14)
\end{aligned}$$

Let's consider the derivative for a single $\boldsymbol{\theta}_i$, $\frac{\partial CE}{\partial \boldsymbol{\theta}_i}$ The first term falls away for $i \neq k$, which gives us:

$$\begin{aligned}
\frac{\partial CE}{\partial \boldsymbol{\theta}_i} &= \hat{\mathbf{y}}_i \quad \text{if } i \neq k && (15)\\
\frac{\partial CE}{\partial \boldsymbol{\theta}_i} &= \hat{\mathbf{y}}_i - 1 \quad \text{if } i = k && (16)
\end{aligned}$$

Or, more consisely in vector notation, $\frac{\partial CE}{\partial \boldsymbol{\theta}} = \hat{\mathbf{y}} - \mathbf{y}$.

2.3 (c)

To simply the notation, let

$$\begin{aligned} z_2 &= xW_1 + b_1 && (17) \\ z_3 &= hW_2 + b_2 && (18) \\ & && (19) \end{aligned}$$

Using the backpropagation algorithm derived in class we know that $\frac{\partial J}{\partial z_2} = \delta_2 = (W_2\delta_3 \circ \sigma'(z_2)) = (W_2(\hat{y} - y) \circ \sigma'(z_2))$.

Then:

$$\begin{aligned} \frac{\partial J}{\partial x_i} &= \sum_j \frac{\partial J}{\partial z_{2j}} \frac{\partial z_{2j}}{\partial x_i} && (20) \\ &= \sum_j \delta_{2j} \frac{\partial z_{2j}}{\partial x_i} && (21) \\ &= \sum_j \delta_{2j} \sum_k \frac{\partial}{\partial x_i} x_k W_{1kj} + b_j && (22) \\ &= \sum_j \delta_{2j} \frac{\partial}{\partial x_i} \sum_k x_k W_{1kj} + b_j && (23) \\ &= \sum_j \delta_{2j} W_{1ij} && (24) \\ & && (25) \end{aligned}$$

Vectorizing the above we see that $\frac{\partial J}{\partial x} = W_1\delta_2$.

2.4 (d)

W_1 must be of dimension $D_x \times H$. b_1 must be of dimension H. W_2 must be of dimension $H \times D_y$. b_2 must be of dimension D_y. The total number of parameters is the sum of these: $HD_x + H + HD_y + D_y$. If we try to learn the vectors for the input data too we need to add another D_x parameters.

3 word2vec

3.1 (a)

Applying the cross-entropy cost we get:

$$J(\hat{\mathbf{r}}, \mathbf{w}) = -\log \frac{e^{\mathbf{w}_i^T \hat{\mathbf{r}}}}{\sum_{j=1}^{|V|} e^{\mathbf{w}_j^T \hat{\mathbf{r}}}} \tag{26}$$

$$= -\log e^{\mathbf{w}_i^T \hat{\mathbf{r}}} + \log \sum_{j=1}^{|V|} e^{\mathbf{w}_j^T \hat{\mathbf{r}}} \tag{27}$$

$$= -\mathbf{w}_i^T \hat{\mathbf{r}} + \log \sum_{j=1}^{|V|} e^{\mathbf{w}_j^T \hat{\mathbf{r}}} \tag{28}$$

$$\tag{29}$$

Let $z_j = \mathbf{w}_j^T \hat{\mathbf{r}}$ and $\mathbb{1}[j = i]$ the indicator function evlauating to 1 if $j = i$ and 0 otherwise. Then:

$$\frac{\partial J}{\partial z_k} = \frac{e^{z_k}}{\sum_{j=1}^{|V|} e^{z_j}} - \mathbb{1}[k = i] \tag{30}$$

$$\frac{\partial J}{\partial \hat{\mathbf{r}}} = \sum_{k=1}^{|V|} \frac{\partial J}{\partial z_k} \frac{\partial z_k}{\partial \hat{\mathbf{r}}} \tag{31}$$

$$= \sum_{k=1}^{|V|} \mathbf{w}_j \left(\frac{e^{z_k}}{\sum_{j=1}^{|V|} e^{z_j}} - \mathbb{1}[k = i] \right) \tag{32}$$

3.2 (b)

$$\frac{\partial J}{\partial \mathbf{w}_k} = \frac{\partial J}{\partial z_k} \frac{\partial z_k}{\partial \mathbf{w}_k} \tag{33}$$

$$= \hat{\mathbf{r}} \left(\frac{e^{z_k}}{\sum_{j=1}^{|V|} e^{z_j}} - \mathbb{1}[k = i] \right) \tag{34}$$

3.3 (c)

Let $z_j = \mathbf{w}_j^T \hat{\mathbf{r}}$, and let $\mathbb{1}[j = i]$ be the indicator function evlauating to 1 if $j = i$ and 0 otherwise. Let's first look at the case of $i \notin K$:

$$\begin{aligned}
\frac{\partial J}{\partial z_i} &= -\frac{\partial}{\partial z_i} log(\sigma(z_i)) && (35)\\
&= -\frac{\sigma'(z_i)}{\sigma(z_i)} && (36)\\
&= -\frac{\sigma(z_i)(1-\sigma(z_i))}{\sigma(z_i)} && (37)\\
&= \sigma(z_i) - 1 && (38)
\end{aligned}$$

And for $i \in K$:

$$\begin{aligned}
\frac{\partial J}{\partial z_i} &= -\frac{\partial}{\partial z_i} log(\sigma(-z_i)) && (39)\\
&= -\frac{-\sigma'(-z_i)}{\sigma(-z_i)} && (40)\\
&= \frac{\sigma(-z_i)(1-\sigma(-z_i))}{\sigma(-z_i)} && (41)\\
&= 1 - \sigma(-z_i) && (42)\\
&= \sigma(z_i) && (43)
\end{aligned}$$

More generally, $\frac{\partial J}{\partial z_j} = (\sigma(z_j) - \mathbb{1}[j = i])$. We note that this is the prediction error. Then, using the chain rule:

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{w}_j} &= \frac{\partial J}{\partial z_j}\frac{\partial z_j}{\partial \mathbf{w}_j} = (\sigma(\mathbf{w}_j^T \hat{\mathbf{r}}) - \mathbb{1}[j = i])\hat{\mathbf{r}} && (44)\\
\frac{\partial J}{\partial \hat{\mathbf{r}}} &= \frac{\partial J}{\partial z_j}\frac{\partial z_j}{\partial \hat{\mathbf{r}}} = (\sigma(\mathbf{w}_j^T \hat{\mathbf{r}}) - \mathbb{1}[j = i])\mathbf{w}_j && (45)
\end{aligned}$$

The negative sampling loss is much cheaper to evaluate because we don't need to sum over the whole vocabulary, just $|K|$ samples.

3.4 (d)

In the skip-gram model we simply sum the gradients calculated for each context.

$$\frac{\partial J}{\partial \hat{\mathbf{r}}} = \sum_{-c \leq j \leq c, j \neq 0} \frac{\partial F(\mathbf{v}'_{w_{i+j}} | \mathbf{v}_{w_i})}{\partial \hat{\mathbf{r}}} \tag{46}$$

$$\frac{\partial J}{\partial \mathbf{w}_j} = \sum_{-c \leq j \leq c, j \neq 0} \frac{\partial F(\mathbf{v}'_{w_{i+j}} | \mathbf{v}_{w_i})}{\partial \mathbf{w}_j} \tag{47}$$

4 Sentiment Analysis

4.1 (a)

Regularization helps us prevent overfitting on our training data by keeping the parameters small.

4.2 (b)

Figure 1 shows that regularization improves the accuracy on the development set, but too much regularization introduces a bias that results in worse performance.

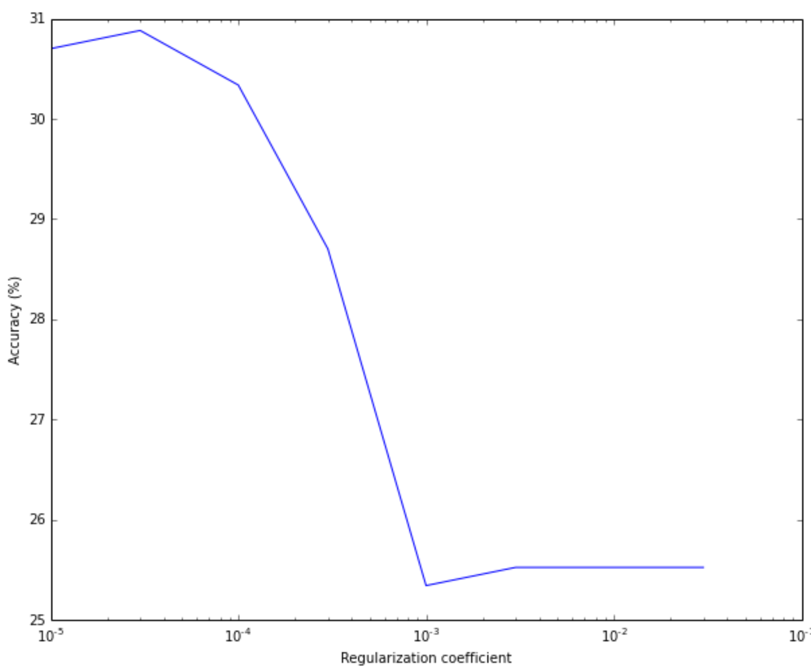

Figure 1: Regularization strength vs. dev accuracy