Assignment 1 for CS224d

Lifu Huang

Feburary 2016

1 Softmax

Proof:

$$\begin{aligned}(softmax(x))_i &= \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}} \\ &= \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}} \cdot \frac{e^c}{e^c} \\ &= \frac{e^{x_i+c}}{\sum_{k=1}^{K} e^{x_k+c}} \\ &= (softmax(x+c))_i\end{aligned}$$

Therefore,

$$softmax(x) = softmax(x+c)$$

2 Neural Network Basics

(a)

Answer:

$$\begin{aligned}\frac{\partial}{\partial x}\sigma(x) &= \frac{\partial}{\partial x}\frac{1}{1+e^{-x}} \\ &= \frac{e^{-x}}{(1+e^{-x})^2} \\ &= \frac{1}{1+e^{-x}} \cdot (1 - \frac{1}{1+e^{-x}}) \\ &= \sigma(x) \cdot (1-\sigma(x))\end{aligned}$$

(b)

Answer:

Assume y is an one-hot vector with k^{th} element being one, other elements being zeros,

$$\begin{aligned}
CE(y, \hat{y}) &= -\sum_i y_i \log(\hat{y}_i) \\
&= -\log(\hat{y}_k) \\
&= -\log(softmax(\theta)_k) \\
\frac{\partial}{\partial \theta_i} CE(y, \hat{y}) &= \frac{\partial}{\partial \theta_i} \left(-\log(softmax(\theta)_k)\right) \\
&= \frac{\partial}{\partial \theta_i} \left(-\theta_k + \log \sum_{j=1}^{s_3} e^{\theta_j}\right) \\
&= \frac{e^{\theta_i}}{\log \sum_{j=1}^{s_3} e^{\theta_j}} - y_i \\
&= \hat{y}_i - y_i
\end{aligned}$$

Therefore,

$$\frac{\partial}{\partial \theta} CE(y, \hat{y}) = \hat{y} - y$$

(c)

Assume $z = W_1 x + b_1$

$$
\begin{aligned}
\delta_i^{(2)} &= \frac{\partial}{\partial z_i} CE(y, \hat{y}) \\
&= \sum_{j=1}^{s_3} \left(\frac{\partial}{\partial \theta_j} CE(y, \hat{y}) \cdot \frac{\partial \theta_j}{\partial z_i} \right) \\
&= \sum_{j=1}^{s_3} \left((\hat{y}_j - y_j) \cdot \frac{\partial}{\partial z_i} (W^{(2)} \sigma(z) + b^{(2)})_j \right) \\
&= \sum_{j=1}^{s_3} \left((\hat{y}_j - y_j) \cdot W_{ji}^{(2)} \cdot \sigma'(z) \right) \\
&= \sigma'(z) \sum_{j=1}^{s_3} \left((\hat{y}_j - y_j) \cdot (W^{(2)})_{ji} \right) \\
\delta^{(2)} &= \frac{\partial}{\partial z} CE(y, \hat{y}) \\
&= \sigma'(z) \cdot (W^{(2)})^T \cdot (\hat{y} - y) \\
&= \sigma'(z) \cdot (W^{(2)})^T \cdot \delta^{(3)} \\
\delta_i^{(1)} &= \frac{\partial}{\partial x_i} CE(y, \hat{y}) \\
&= \sum_{j=1}^{s_2} \left(\frac{\partial}{\partial z_j} CE(y, \hat{y}) \cdot \frac{\partial z_j}{\partial x_i} \right) \\
&= \sum_{j=1}^{s_2} \left(\delta_j^{(2)} \cdot \frac{\partial}{\partial x_i} \cdot (W^{(1)} \cdot x + b)_j \right) \\
&= \sum_{j=1}^{s_2} \left(\delta_j^{(2)} \cdot W_{ji}^{(1)} \right) \\
\delta^{(1)} &= (W^{(1)})^T \cdot \delta^{(2)}
\end{aligned}
$$

In conclusion,

$$
\frac{\partial}{\partial x} = (W^{(1)})^T \cdot (W^{(2)})^T \cdot \delta^{(3)} \cdot \sigma(z) \cdot (1 - \sigma(z))
$$

(d)

$$
\begin{aligned}
|W^{(1)}| &= D_x \cdot H \\
|W^{(2)}| &= H \cdot D_y \\
|b^{(2)}| &= H \\
|b^{(3)}| &= D_y
\end{aligned}
$$

Therefore, there are $D_x \cdot H + H \cdot D_y + H + D_y$ arguments in total.

3 word2vec

Please see Derivation of Word2Vec Models for detailed derivation of word2vec model.

4 Sentiment Analysis

(a)

Answer: Regularization is introduced to penalize the magnitude of parameters, preventing them from fitting training data by unlimited increasing, which might lead to the problem of overfitting.

(b)