Detailed solutions to CS224n assignments

October 11, 2017

Assignment 1

2 Neural Network Basics

(a)

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

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

(b)

$CE(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_i y_i \ln \hat{y}_i$ where:

$\hat{y}_i = f(\theta_i; \Theta)$

$\Theta = [\theta_1, ..., \theta_j, ...\theta_c]$

c is the number of classes

$\mathbf{y}$ is non zero only in one entry, in which case it is 1. Let's call that entry t (like "true"):

$CE(\mathbf{y}, \hat{\mathbf{y}}) = -\ln \hat{y}_t = -\ln f(\theta_t, \Theta)$

$$\frac{\partial CE(\mathbf{y}, \hat{\mathbf{y}})}{\partial \theta_k} = -\frac{\partial \ln f(\theta_t, \Theta)}{\partial \theta_k} = -\frac{\frac{\partial f(\theta_t, \Theta)}{\partial \theta_k}}{f(\theta_t, \Theta)}$$

Softmax derivative:

First, let's have a look a the derivative of the denominator:

$$\frac{\partial \sum_j e^{x_j}}{\partial x_k} = \sum_j \frac{\partial e^{x_j}}{\partial x_k}$$

which is e^{x_j} when $k = j$ and 0 otherwise, so

$$\frac{\partial \sum_j e^{x_j}}{\partial x_k} = e^{x_k}$$

Now, for the derivative of the whole

$\frac{\partial \text{sm}_i}{\partial x_k} = \dfrac{\dfrac{\partial e^{x_i}}{\partial x_k}\sum_j e^{x_i} - e^{x_i}\dfrac{\partial \sum_j e^{x_j}}{\partial x_k}}{(\sum_j e^{x_j})^2}$

Case 1: $k = i$

$\frac{\partial \text{sm}_i}{\partial x_k} = e^{x_k}\frac{\sum_j e^{x_j} - e^{x_k}}{(\sum_j e^{x_j})^2}$

Case 2: $k \neq i$

$\frac{\partial \text{sm}_i}{\partial x_k} = -e^{x_i}\dfrac{e^{x_k}}{(\sum_j e^{x_j})^2}$

So in the case where $f = \text{softmax}$

Case 1: $k = t$

$\dfrac{\partial CE(\mathbf{y}, \hat{\mathbf{y}})}{\partial \theta_k} = -\frac{e^{\theta_k}\frac{\sum_j e^{\theta_j} - e^{\theta_k}}{(\sum_j e^{\theta_j})^2}}{\frac{e^{\theta_k}}{\sum_j e^{\theta_j}}} = -\frac{\sum_j e^{\theta_j} - e^{\theta_k}}{\sum_j e^{\theta_j}} = \text{softmax}(\theta_k; \Theta) - 1 = \hat{y}_k - 1$

Case 2: $k \neq t$

$\dfrac{\partial CE(\mathbf{y}, \hat{\mathbf{y}})}{\partial \theta_k} = -y_k\frac{-e^{\theta_t}\frac{e^{\theta_k}}{(\sum_j e^{\theta_j})^2}}{\frac{e^{\theta_t}}{\sum_j e^{\theta_j}}} = \frac{e^{\theta_k}}{\sum_j e^{\theta_j}} = \text{softmax}(\theta_k; \Theta) = \hat{y}_k$

Since $\mathbf{y}$ is 0 everywhere except at $k = t$, we can rewrite this as

$\dfrac{\partial CE(\mathbf{y}, \hat{\mathbf{y}})}{\partial \theta_k} = \hat{\mathbf{y}} - \mathbf{y}$