Assignment 1

Arvid Frydenlund

August 30, 2015

1 Q1

softmax($\mathbf{x}$) = softmax($\mathbf{x} + c$) (Note, remember softmax is defined as a function taking in a vector)

Considering element i

$$\frac{e^{x_i}}{\sum_j e^{x_j}} = \frac{e^{x_i+c}}{\sum_j e^{x_j+c}} \tag{1}$$

since,

$$\frac{e^{x_i}e^c}{\sum_j e^{x_j}e^c} = \frac{e^{x_i}\cancel{e^c}}{\sum_j e^{x_j}\cancel{e^c}} \tag{2}$$

QED

This is useful for the trick where $c = -\max_{\mathrm{i}}^{\mathrm{n}} \mathbf{x}_{\mathrm{i}}$, since the largest number in the exponential possible is now 0. That is it prevents overflow.

2 Q2 a)

The sigmoid function is

$$y = \frac{1}{1+e^{-x}} = \sigma(x) = (1+e^{-x})^{-1} = (z)^{-1} \tag{3}$$

Where $z = 1 + e^{-x} = 1 + e^r$, and $r = -x$

Then by simple chain rule

$$\frac{dy}{dx} = \frac{dy}{dz} * \frac{dz}{dr} * \frac{dr}{dx} \tag{4}$$

$$\frac{dy}{dx} = -(1+e^{-x})^{-2} * e^{-x} * (-1) = e^{-x} * \left(\frac{1}{1+e^{-x}}\right)^2 = e^{-x}\sigma(x)\sigma(x) \tag{5}$$

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

$$\frac{1 - \frac{1}{1+e^{-x}}}{\frac{1}{1+e^{-x}}} = (1+e^{-x}) - \frac{1+e^{-x}}{1+e^{-x}} = (1+e^{-x}) - 1 = e^{-x} \tag{6}$$

Then

$$e^{-x}\sigma(x)\sigma(x) = \frac{1-\sigma(x)}{\cancel{\sigma(x)}}\cancel{\sigma(x)}\sigma(x) = (1-\sigma(x))(\sigma(x)) \tag{7}$$

QED

3 Q2 b)

Asume K classes, $\hat{y} = \text{softmax}(\theta) = [\hat{y}_1, ..., \hat{y}_K]$ =[softmax(θ_1), ...,softmax(θ_K)] $= g(\theta_i) = \frac{e^{\theta_i}}{\sum_c e^{\theta_c}}$

$$J = CE(y, \hat{y}) = -\sum_i^K y_i \log(\hat{y}_i) \tag{8}$$

Note that the sum is over all the classes and that y is 1-hot so that the sum only picks out one value

$$\nabla_\theta J = -\sum_i^K \frac{\partial}{\partial\theta} y_i \log(\hat{y}_i) = -\sum_i^K y_i \frac{\partial}{\partial\theta}\log(g(\theta_i)) = -\sum_i^K y_i \frac{1}{g(\theta_i)}\frac{\partial}{\partial\theta} g(\theta_i) \tag{9}$$

Considering $\frac{\partial}{\partial\theta} g(\theta_i)$, for an element j of θ, θ_j there are two cases.
Case 1) $j = i$ then note

$$\frac{\partial}{\partial\theta_j} g(\theta_i) = \frac{\partial}{\partial\theta_j} g(\theta_j) = \frac{\partial}{\partial\theta_i} g(\theta_i) \tag{10}$$

$$\frac{dg(\theta_i)}{d\theta_j} = \frac{d}{d\theta_j}\frac{e^{\theta_i}}{\sum_c e^{\theta_c}} = \frac{(\frac{d}{d\theta_j}e^{\theta_i})(\sum_c e^{\theta_c}) - (\frac{d}{d\theta_j}\sum_c e^{\theta_c})(e^{\theta_i})}{(\sum_c e^{\theta_c})^2} \tag{11}$$

Where $(\frac{d}{d\theta_j}\sum_c e^{\theta_c})(e^{\theta_i}) = \frac{d}{d\theta_j} e^{\theta_1} + ... + e^{\theta_j} + ... + e^{\theta_K} = e^{\theta_j} = e^{\theta_i}$
So

$$= \frac{(e^{\theta_i})(\sum_c e^{\theta_c}) - (e^{\theta_i})(e^{\theta_i})}{(\sum_c e^{\theta_c})(\sum_c e^{\theta_c})} = \frac{(e^{\theta_i})}{(\sum_c e^{\theta_c})}\frac{(\sum_c e^{\theta_c} - e^{\theta_i})}{(\sum_c e^{\theta_c})} = g(\theta_i)\left(\frac{\sum_c e^{\theta_c}}{\sum_c e^{\theta_c}} - \frac{e^{\theta_i}}{\sum_c e^{\theta_c}}\right) \tag{12}$$

$$= g(\theta_i)(1 - g(\theta_i)) = g(\theta_i)(1 - g(\theta_j)) \tag{13}$$

Now in the case there $j \neq i$

$$\frac{dg(\theta_i)}{d\theta_j} = \frac{d}{d\theta_j}\frac{e^{\theta_i}}{\sum_c e^{\theta_c}} = \frac{d}{d\theta_j}e^{\theta_i}\left(\sum_c e^{\theta_c}\right)^{-1} = -e^{\theta_i}\left(\sum_c e^{\theta_c}\right)^{-2}e^{\theta_j} \tag{14}$$

$$= \frac{-(e^{\theta_i})}{(\sum_c e^{\theta_c})}\frac{(e^{\theta_j})}{(\sum_c e^{\theta_c})} = -g(\theta_i)g(\theta_j) \tag{15}$$

Let $t_{ij} = 1$ if $i = j$ and $= 0$, if $i \neq j$, then we can combine both cases as

$$\frac{dg(\theta_i)}{d\theta_j} = g(\theta_i)(t_{ij} - g(\theta_i)) \tag{16}$$

Which can be confirmed with $i = j$, $g(\theta_i)(1 - g(\theta_i)) = g(\theta_i)(1 - g(\theta_j))$ and with $j \neq i$, $g(\theta_i)(0 - g(\theta_j)) = g(\theta_i)(-g(\theta_j)) = -g(\theta_i)(g(\theta_j))$

Now putting it all together we get

$$\nabla_\theta J = -\sum_i^K y_i \frac{1}{\cancel{g(\theta_i)}}\cancel{g(\theta_i)}(t_{ij} - g(\theta_i)) \tag{17}$$

Let $t = y$, and since y is 1-hot, this gets vectorized as

$$= -y - g(\theta) = -y - \hat{y} = \hat{y} - y \tag{18}$$