NLP assignment #1

Andrius Buinovskij - 18-940-270

Q1 b) i)

Well since all inputs are 1 and all weights are 1 then

$$\mathbf{s}_1^1 = \sum_{i=1}^{3} \mathbf{x}_i \cdot \mathbf{w}_{i,1}^1 = 3 \tag{1}$$

$$\mathbf{s}_2^1 = \sum_{i=1}^{3} \mathbf{x}_i \cdot \mathbf{w}_{i,2}^1 = 3 \tag{2}$$

Where $\mathbf{s}_j^i$ is the sum input to the j'th neuron in the i'th layer, and $\mathbf{s}^i$ is a vector whose length is equal to the number of neurons in the i'th layer, with the input being the 0'th layer. So $\mathbf{s}^1$ is of length 2 since 1'st layer has 2 neurons.

Let $\mathbf{n}_j^i$ be the output of the j'th neuron in the i'th layer, then of course $\mathbf{n}^i$ is a vector of length equal to the number of neurons in the i'th layer:

$$n_j^i = ReLU\big(\mathbf{s}_j^i\big) \tag{3}$$

So in our case We get

$$\mathbf{n}_1^1 = ReLU\big(\mathbf{s}_1^1\big) = ReLU\big(3\big) = 3 \tag{4}$$

$$\mathbf{n}_2^1 = ReLU\big(\mathbf{s}_2^1\big) = ReLU\big(3\big) = 3 \tag{5}$$

Same steps in the next layer:

$$\mathbf{s}_1^2 = 3 \cdot 1 + 3 \cdot 1 = 6 \tag{6}$$

And now We pass this through a sigmoid for our output instead of a $ReLU$ so We get

$$out = \sigma\big(\mathbf{s}_1^2\big)\big) = \frac{1}{1 + e^{-6}} = 0.99752737684 \tag{7}$$

Q1 b) ii)

$$\frac{d}{d\mathbf{w}^1_{j,k}}\mathbf{x}_i \cdot \mathbf{w}^1_{j,k} = \mathbf{x}_i \tag{8}$$

Since the inputs are all 1 this simplifies to

$$\frac{d}{d\mathbf{w}^1_{j,k}}\mathbf{x}_i \cdot \mathbf{w}^1_{j,k} = \frac{d}{d\mathbf{w}^1_{j,k}}1 \cdot \mathbf{w}^1_{j,k} = 1 \tag{9}$$

Now for the second layer:

$$\frac{d}{d\mathbf{w}^2_{1,1}}f = \frac{d}{d\mathbf{w}^2_{1,1}}\mathbf{n}_{1,1} \cdot \mathbf{w}^2_{1,1} = \frac{d}{d\mathbf{w}^2_{1,1}}ReLU\big(\mathbf{s}_{1,1}\big) \cdot \mathbf{w}^2_{1,1} = ReLU\big(\mathbf{s}_{1,1}\big) \tag{10}$$

Since $ReLU\big(\mathbf{s}_{1,1}\big)$ is just a constant w.r.t. $\mathbf{w}^2_{1,1}$. In this case We simply get

$$\frac{d}{d\mathbf{w}^2_{1,1}}f = ReLU\big(\mathbf{s}_{1,1}\big) = 3 \tag{11}$$

And likewise for the other weight.

Q1 b) iii)

$$\begin{aligned} L_{BCE} &= -(y\log(\hat{y}) + (1-y)\log(1-\hat{y})) && (12)\\ &= -(0 \cdot \log(0.99752737684) + (1-0)\log(1-0.99752737684)) && (13)\\ &= -(\log(0.00247262316)) && (14)\\ &= 2.60684206722 && (15) \end{aligned}$$

redone

Alright let's just roll all of these questions into one and do an iteration of backprop.

The forward pass is trivial.

Let $\mathbf{x}^l_i$ be the output of the i'th neuron in the l'th layer.

Let $\mathbf{w}^l_{i,j}$ be the weight belonging to j'th neuron multiplying the i'th input in the l'th layer. $\mathbf{w}^l_{:,j}$ is then simply the weight vector associated with the j'th neuron in the l'th layer.

Let $\mathbf{s}^l_j$ be $(\mathbf{x}^{l-1})^\top \mathbf{w}^l_{:,j}$, the weighted sum of inputs to the j'th neuron in the l'th layer.

We can then say that $x^l_i = \sigma(\mathbf{s}^l_j)$, where σ is the nonlinearity of choice.

Then define

$$\delta^l_j = \frac{\partial L_{BCE}}{\partial \mathbf{s}^l_j} \tag{16}$$

And finally

$$\delta_j^{l-1} = \sum_{i=1}^{d^l} \frac{\partial L_{BCE}}{\partial \mathbf{s}_i^l} \frac{\partial \mathbf{s}_i^l}{\partial \mathbf{x}_j^{l-1}} \frac{\partial \mathbf{x}_j^{l-1}}{\partial \mathbf{s}_j^{l-1}} \tag{17}$$

$$= \sum_{i=1}^{d^l} \delta_i^l \frac{\partial \mathbf{s}_i^l}{\partial \mathbf{x}_j^{l-1}} \frac{\partial \mathbf{x}_j^{l-1}}{\partial \mathbf{s}_j^{l-1}} \tag{18}$$

Where d^l is the number of neurons in layer l. So now We have a recursive definition which uses dynamic programming. This could be further nuanced by expressing things in matrix notation, but it's good enough for present purposes.

Here is also a picture since just looking at symbols is a nightmare.

So now

$$\frac{\partial L_{BCE}}{\partial \mathbf{s}_1^2} = \frac{\partial L_{BCE}}{\partial \hat{y}} \frac{\partial \hat{y}}{\mathbf{s}_1^2} \tag{19}$$

$$\frac{\partial L_{BCE}}{\partial \hat{y}} = \frac{\partial}{\partial \hat{y}} - \big(y \log(\hat{y}) + (1-y)\log(1-\hat{y})\big) \tag{20}$$

$$= -y\frac{1}{\hat{y}} - (1-y)\frac{\partial}{\partial \hat{y}} \log(1-\hat{y}) \tag{21}$$

$$= -y\frac{1}{\hat{y}} - (1-y)\frac{-1}{1-\hat{y}} \tag{22}$$

$$= -y\frac{1}{\hat{y}} + \frac{1-y}{1-\hat{y}} \tag{23}$$

$$\frac{\partial \hat{y}}{\mathbf{s}_1^2} = \frac{\partial}{\mathbf{s}_1^2}\sigma(\mathbf{s}_1^2) \tag{24}$$

$$= \sigma(\mathbf{s}_1^2) \cdot (1 - \sigma(\mathbf{s}_1^2)) \tag{25}$$

So then We have

$$\delta_1^2 = \frac{\partial L_{BCE}}{\partial \mathbf{s}_1^2} = \frac{\partial L_{BCE}}{\partial \hat{y}} \frac{\partial \hat{y}}{\mathbf{s}_1^2} \tag{26}$$

$$= \Bigg(- y\frac{1}{\hat{y}} + \frac{1-y}{1-\hat{y}}\Bigg)\Bigg(\sigma(\mathbf{s}_1^2) \cdot (1 - \sigma(\mathbf{s}_1^2))\Bigg) \tag{27}$$

Alright. We only have one more of these to figure out:

$$\delta_j^1 = \sum_{i=1}^{d^2} \delta_i^2 \frac{\partial \mathbf{s}_i^2}{\partial \mathbf{x}_j^1} \frac{\partial \mathbf{x}_j^1}{\partial \mathbf{s}_j^1} \tag{28}$$

But since $d^2 = 1$, i.e. there is only one neuron in the output layer, We have

$$\delta_j^1 = \sum_{i=1}^{d^2} \delta_i^2 \frac{\partial \mathbf{s}_i^2}{\partial \mathbf{x}_j^1} \frac{\partial \mathbf{x}_j^1}{\partial \mathbf{s}_j^1} \tag{29}$$

$$= \delta_1^2 \frac{\partial \mathbf{s}_1^2}{\partial \mathbf{x}_j^1} \frac{\partial \mathbf{x}_j^1}{\partial \mathbf{s}_j^1} \tag{30}$$

$$= \delta_1^2 \mathbf{w}_j^2 \sigma(\mathbf{s}_j^1)(1 - \sigma(\mathbf{s}_j^1)) \tag{31}$$

Now for the gradient update We will of course need

$$\frac{\partial L_{BCE}}{\partial \mathbf{w}_{i,j}^l} \tag{32}$$

But this can be easily derived since

$$\frac{\partial L_{BCE}}{\partial \mathbf{w}_{i,j}^l} = \frac{\partial L_{BCE}}{\partial \mathbf{s}_j^l} \frac{\partial \mathbf{s}_j^l}{\partial \mathbf{w}_{i,j}^l} \tag{33}$$

$$= \delta_j^l \mathbf{x}_i^{l-1} \tag{34}$$

Now We do a forward pass and We know that $\mathbf{s}_i^1 = 3$ and $\mathbf{s}_1^2 = 6$. Plugging in values We then get

$$\delta_1^2 = \left(-y\frac{1}{\hat{y}} + \frac{1-y}{1-\hat{y}} \right) \left(\sigma(\mathbf{s}_1^2) \cdot (1 - \sigma(\mathbf{s}_1^2)) \right) \tag{35}$$

$$= \left(\frac{1}{1 - 0.99752737684} \right) \left(\sigma(6) \cdot (1 - \sigma(6)) \right) \tag{36}$$

$$= \left(\frac{1}{1 - 0.99752737684} \right) \left(0.99752737684 \cdot (1 - 0.99752737684) \right) \tag{37}$$

$$= 404.428792942 \cdot 0.00246650929 \tag{38}$$

$$= 0.99752737493 \tag{39}$$

Similarly We have

$$\delta_j^1 = \delta_1^2 \mathbf{w}_j^2 \sigma(\mathbf{s}_j^1)(1 - \sigma(\mathbf{s}_j^1)) \tag{40}$$

$$\delta_1^1 = \delta_1^2 \mathbf{w}_1^2 \sigma(\mathbf{s}_1^1)(1 - \sigma(\mathbf{s}_1^1)) \tag{41}$$

$$= 0.997527374931 \cdot \sigma(3)(1 - \sigma(3)) \tag{42}$$

$$= 0.997527374931 \cdot 0.95257412682 \cdot 0.04742587317 \tag{43}$$

$$= 0.04506495478 \tag{44}$$

δ_2^1 is equal to δ_1^1.
Now that We have the deltas We can use

$$\frac{\partial L_{BCE}}{\partial \mathbf{w}_{i,j}^l} = \delta_j^l \mathbf{x}_i^{l-1} \tag{45}$$

For weights in the first layer this means

$$\begin{align}
\frac{\partial L_{BCE}}{\partial \mathbf{w}_{i,j}^1} &= \delta_j^1 \mathbf{x}_i^0 \tag{46}\\
&= 0.04506495478 \cdot 1 \tag{47}\\
&= 0.04506495478 \tag{48}
\end{align}$$

Where the 0'th layer is the input layer.

$$\begin{align}
\frac{\partial L_{BCE}}{\partial \mathbf{w}_{i,j}^2} &= \delta_j^2 \mathbf{x}_i^1 \tag{49}\\
&= 0.99752737493 \cdot 3 \tag{50}\\
&= 2.99258212479 \tag{51}
\end{align}$$

All that is left is to perform a step for each weight. All weights in the first layer simplify to

$$\begin{align}
w_{i,j}^1 &= 1 - 0.1 \cdot 0.04506495478 \tag{52}\\
&= 0.99549350452 \tag{53}
\end{align}$$

Sidenote
So when actually implementing this, for the next assignemnt, We get

$$\begin{align}
\delta_1^2 = \frac{\partial L_{BCE}}{\partial \mathbf{s}_1^2} &= \frac{\partial L_{BCE}}{\partial \hat{y}} \frac{\partial \hat{y}}{\mathbf{s}_1^2} \tag{54}\\
&= \left(-y\frac{1}{\hat{y}} + \frac{1-y}{1-\hat{y}} \right) \left(\sigma(\mathbf{s}_1^2) \cdot (1 - \sigma(\mathbf{s}_1^2)) \right) \tag{55}\\
&= \left(-y\frac{1}{\hat{y}} + \frac{1-y}{1-\hat{y}} \right) \left(\hat{y} \cdot (1 - \hat{y}) \right) \tag{56}\\
&= -\hat{y} \cdot (1-\hat{y}) \cdot \frac{y}{\hat{y}} + \hat{y} \cdot (1-\hat{y}) \cdot \frac{1-y}{1-\hat{y}} \tag{57}\\
&= -(1-\hat{y}) \cdot y + \hat{y} \cdot (1-y) \tag{58}
\end{align}$$