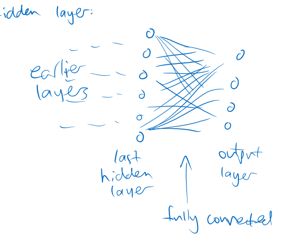
1

Consider a neural network solving a classification problem. There are 4 classes, and 6 nodes in the final hidden layer:



In the notation of ljvmiranda's article, the output values before the activation function is applied are f_1, f_2, f_3, f_4. After the softmax activation is applied, the output values are probabilities denoted p_1, p_2, p_3, p_4 and calculated according to

$$p_i = \frac{e^{f_i}}{\sum_{j=1}^{4} e^{f_j}} = \frac{e^{f_i}}{V(f)}$$

where $f =$ vector (f_1, f_2, f_3, f_4)

and $V(f) = \sum_{j=1}^{4} e^{f_j}$

2

We would like to train this neural network using a single training example, which happens to be from class 3.

We will define a loss function L for the training. We expect that

$\frac{\partial L}{\partial f_3}$ will be negative, since increasing f_3 leads to higher probability for class 3, which is a better prediction, and should have lower loss.

$\frac{\partial L}{\partial f_1}, \frac{\partial L}{\partial f_2}, \frac{\partial L}{\partial f_4}$ will be positive since higher probabilities for those classes leads to a worse prediction which should have higher loss.

(✡)

Let's calculate the derivative and see if our expectation (✡) is correct.

3

First we need to define the loss function L. We will use the negative log likelihood:

$$L = -\log \mathcal{L}$$, where $\mathcal{L}$ is the likelihood of the observed data.

The likelihood of our single observation (class 3) is just p_3. So $\mathcal{L} = p_3$.

If this is confusing, suppose we had observed more samples instead e.g. maybe there were 3 observations that produced values

$f^{(1)} = (f_1^{(1)}, f_2^{(1)}, f_3^{(1)}, f_4^{(1)})$ true class 4 encoded as $t^{(1)} = (0,0,0,1) = (t_1^{(1)}, t_2^{(1)}, t_3^{(1)}, t_4^{(1)})$

$f^{(2)} = (f_1^{(2)}, f_2^{(2)}, f_3^{(2)}, f_4^{(2)})$ true class 3 encoded as $t^{(2)} = (0,0,1,0) = (t_1^{(2)}, t_2^{(2)}, t_3^{(2)}, t_4^{(2)})$

$f^{(3)} = (f_1^{(3)}, f_2^{(3)}, f_3^{(3)}, f_4^{(3)})$ true class 3 encoded as $t^{(1)} = (0,0,1,0) = (t_1^{(3)}, t_2^{(3)}, t_3^{(3)}, t_4^{(3)})$

Here the likelihood would be

$$\mathcal{L} = p_4^{(1)} \times p_3^{(2)} \times p_3^{(3)} = \mathcal{L}^{(1)}\mathcal{L}^{(2)}\mathcal{L}^{(3)}$$

Note that $\mathcal{L}^{(1)}$ can be written rather more elaborately as

$$\mathcal{L}^{(1)} = {p_1^{(1)}}^{t_1^{(1)}} {p_2^{(1)}}^{t_2^{(1)}} {p_3^{(1)}}^{t_3^{(1)}} {p_4^{(1)}}^{t_4^{(1)}}$$ and same for $\mathcal{L}^{(2)}, \mathcal{L}^{(3)}$

So $\mathcal{L}$ can be written in a more general but equivalent way as

$$\mathcal{L} = \prod_{i=1}^{\text{num observations}} \prod_{j=1}^{\text{num classes}} \left(p_j^{(i)}\right)^{t_j^{(i)}}$$

(It's just a multinomial pdf without the normalization factor)

4

Anyway, for single observation of class 3 we get

$$L = -\log \mathcal{L} = -\log p_3$$

More precisely,

$$L(f) = L(f_1, f_2, f_3, f_4) = -\log p_3(f) = -\log p_3(f_1, f_2, f_3, f_4)$$

We need to find the derivatives $\frac{\partial L}{\partial f_1}, \frac{\partial L}{\partial f_2}, \frac{\partial L}{\partial f_3}, \frac{\partial L}{\partial f_4}$. We do each one separately.

① ② ③ ④

① $$\frac{\partial L}{\partial f_1} = \sum_{i=1}^{4} \frac{\partial L}{\partial p_i} \frac{\partial p_i}{\partial f_1}$$

(multi-dimensional chain rule. Note that ljvmiranda article skips this step, writing RHS as $\frac{\partial L}{\partial p_1}\frac{\partial p_1}{\partial f_1}$, which is incorrect in cases ①, ② and ④.)

this is zero except when $i=3$, since $L = -\log p_3$

$$= \frac{\partial L}{\partial p_3} \frac{\partial p_3}{\partial f_1}$$

$$\frac{\partial L}{\partial p_3} = \frac{\partial}{\partial p_3}(-\log p_3) = \frac{-1}{p_3}$$

$$\frac{\partial p_3}{\partial f_1} = \frac{\partial}{\partial f_1} \frac{e^{f_3}}{V(f)} = \frac{V\left(\overbrace{\frac{\partial}{\partial f_1} e^{f_3}}^{\text{zero}}\right) - e^{f_3}\left(\overbrace{\frac{\partial}{\partial f_1} V}^{e^{f_1}}\right)}{V^2}$$

[derivative of quotient rule]

$$= \frac{-e^{f_3} e^{f_1}}{V^2} = \frac{-e^{f_3}}{V} \cdot \frac{e^{f_1}}{V}$$

$$= -p_3 p_1$$

$$= \frac{-1}{p_3} \cdot (-p_3 p_1)$$

$$= p_1$$

Note this is positive, in agreement with expectation (✡) above.

5

② $$\frac{\partial L}{\partial f_2} = \ldots \text{ by same reasoning as ①}$$
$$= p_2$$

③ $$\frac{\partial L}{\partial f_3} = \sum_{i=1}^{4} \frac{\partial L}{\partial p_i} \frac{\partial p_i}{\partial f_3}$$

zero except when $i=3$

$$= \frac{\partial L}{\partial p_3} \frac{\partial p_3}{\partial f_3}$$

here we get to the expression in ljvmiranda's article, and our results are the same as in that article, for this case (case ③).

$$= \frac{\partial}{\partial p_3}(-\log p_3) = \frac{-1}{p_3}$$

$$= \frac{\partial}{\partial f_3} \frac{e^{f_3}}{V(f)}$$

$$= \frac{V\left(\frac{\partial}{\partial f_3} e^{f_3}\right) - e^{f_3}\left(\frac{\partial}{\partial f_3} V\right)}{V^2}$$

(quotient rule for derivatives)

$$= \frac{V e^{f_3} - e^{f_3} \cdot e^{f_3}}{V^2}$$

$$= \frac{e^{f_3}}{V} - \frac{e^{f_3}}{V} \cdot \frac{e^{f_3}}{V}$$

$$= p_3 - p_3^2$$

$$= p_3(1 - p_3)$$

$$= \frac{-1}{p_3} \cdot (p_3(1-p_3))$$

$$= p_3 - 1$$

Since p_3 is a probability, we know $p_3 < 1$ so $p_3 - 1 < 0$ so the derivative is negative in agreement with hypothesis (✡)

④ .. same reasoning as ① and ②, obtain

$$\frac{\partial L}{\partial f_4} = p_4$$

6

To summarize,

$$\begin{pmatrix} \frac{\partial L}{\partial f_1} \\ \frac{\partial L}{\partial f_2} \\ \frac{\partial L}{\partial f_3} \\ \frac{\partial L}{\partial f_4} \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 - 1 \\ p_4 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} - \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{pmatrix} = p - t$$

where t is the one-hot encoded class value.

Finally, note that all of the above applied to computing the gradient for the loss function related to a single train example. However, this generalizes very easily to multiple training examples. When we observe multiple independent training examples, the likelihood function is the product of the individual likelihoods. So the log likelihood is the sum of the individual log likelihoods. And the derivative is the sum of the individual derivatives.

Thus,

if we observed m samples with output vectors $p^{(1)}, p^{(2)}, \ldots p^{(m)}$ and ground truth one-hot encoded vectors $t^{(1)}, t^{(2)}, \ldots t^{(m)}$, the softmax derivative is

$$\frac{1}{m} \sum_{i=1}^{m} \left(p^{(i)} - t^{(i)}\right)$$

The $\frac{1}{m}$ factor is there because we usually define the cost function J as average cost:

$$J = \frac{1}{m} \sum_{i=1}^{m} L^{(i)}$$