Notes on backprop slides

Bert Huang

March 2019

1 Second-layer gradients

In my slides, I skip a couple steps in the slide on the gradient of the log likelihood with respect to the hidden layer weights. Let's make these steps more explicit.

The slides have the equation

$$\nabla_{w_{11}} \mathrm{ll} = \sum_{i=1}^{n} \frac{1}{p(y_i|x_i)} \times \nabla_{w_{11}} p(y_i|x_i)$$

Plugging in the definition of $p(y_i|x_i)$, we get

$$\nabla_{w_{11}} \mathrm{ll} = \sum_{i=1}^{n} \frac{1}{\sigma(y_i w_{21}^\top h_i)} \times \nabla_{w_{11}} \sigma(y_i w_{21}^\top h_i)$$

(Here I'm writing h_i to make it explicit that there is a different h vector for each example i, but I didn't do that on the slides out of laziness.)

We can do the same expansion we did on the previous analysis by using chain rule on the expression $\nabla_{w_{11}} \sigma(y_i w_{21}^\top h_i)$, giving us

$$\begin{aligned}
\nabla_{w_{11}} \mathrm{ll} &= \sum_{i=1}^{n} \frac{\sigma(y_i w_{21}^\top h_i)(1 - \sigma(y_i w_{21}^\top h_i))}{\sigma(y_i w_{21}^\top h_i)} \times \nabla_{w_{11}} (y_i w_{21}^\top h_i) \\
&= \sum_{i=1}^{n} (1 - \sigma(y_i w_{21}^\top h_i)) \times \nabla_{w_{11}} (y_i w_{21}^\top h_i) \\
&= \sum_{i=1}^{n} (1 - \sigma(y_i w_{21}^\top h_i)) y_i \times \nabla_{w_{11}} (w_{21}^\top h_i)
\end{aligned} \tag{1}$$

The case analysis from the previous slide (the gradient w.r.t. $w_2 1$ tells us that

$$\begin{aligned}
\nabla_{w_{11}} \mathrm{ll} &= \sum_{i=1}^{n} (1 - \sigma(y_i w_{21}^\top h_i)) y_i \times \nabla_{w_{11}} (w_{21}^\top h_i) \\
&= \sum_{i=1}^{n} (I(y_i = 1) - \sigma(w_{21}^\top h_i)) \times \nabla_{w_{11}} w_{21}^\top h_i
\end{aligned} \tag{2}$$

which gets us to the second line in the slide.

It's useful to see what that expression on the left, $(I(y_i = 1) - \sigma(w_{21}^\top h_i))$, actually means. Recall the original nested function form of this simple neural net,

$$\sigma(w_{21}^\top [\sigma(w_{11}^\top x), \sigma(w_{12}^\top x)]^\top).$$

Since we're interested in the log-likelihood, we are actually differentiating the function

$$\log \sigma(y_i w_{21}^\top [\sigma(w_{11}^\top x), \sigma(w_{12}^\top x)]^\top).$$

The expression $(I(y_i = 1) - \sigma(w_{21}^\top h_i))$ then turns out to be the derivative of this one-dimensional function $\log \sigma(y_i z)$, i.e.,

$$\frac{d \ \log \sigma(y_i z)}{d \ z}.$$

In other words, it's the derivative of the log-likelihood with respect to the input to the final logistic squashing function.