IFT6390

Homework 2: multilayer perceptron (single hidden layer)

Gabriel C-Parent C5912

Q1 a)

What is the dimension of $b^{(1)}$?

$b^{(1)}$ is $d_h \times 1$.

Write down the formula to calculate the vector of activations (i.e. before the non-linearity) of the neurons in the hidden layer, h^a , given an input, x, at first in matrix expression.

$$h^a_{d_h \times 1} = b^{(1)}_{d_h \times 1} + W^{(1)}_{d_h \times d} x_{d \times 1}$$

Element-by-element computation of the entries of h^a.

$$h^a_i = b^{(1)}_i + \sum_{j=1}^{d} w^{(1)}_{i,j} x_j$$

Write down the vector of outputs of the hidden layer neurons, h^s, in terms of the activations, h^a.

$$h^s_{d_h \times 1} = tanh(b^{(1)}_{d_h \times 1} + W^{(1)}_{d_h \times d} x_{d \times 1})$$

Q1 b)

Let $W^{(2)}$ be the weight matrix from the hidden to output layer and $b^{(2)}$ be the vector of biases for the output layer. What are the dimensions of $W^{(2)}$ et $b^{(2)}$?

$b^{(2)}$ is $m \times 1$

$W^{(2)}$ is $m \times d_h$

Write down the formula describing the vector of activations of neurons in the output layer o^a given h^s in matrix form

$$o^a_{m \times 1} = b^{(2)}_{m \times 1} + W^{(2)}_{m \times d_h} h^s_{d_h \times 1}$$

Then in detail element-wise form

$$o_i^a = b_i^{(2)} + \sum_{j=1}^{d_h} w_{i,j}^{(2)} h_j^s$$

Q1 c)

What is contained in the set of all network parameters, θ

- activation function (tanh, sigmoid, linear even)
- number of hidden layer nodes
- $W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}$

What is the number n_θ of parameters in θ ?

- $W^{(1)}$ is $d_h \times d$
- $W^{(2)}$ is $m \times d_h$
- $b^{(1)}$ is $d_h \times 1$
- $b^{(2)}$ is $m \times 1$

n_θ = $d_h * (d + 1 + m) + m$

Q1 d)

Show that the gradients wrt. parameters $W^{(2)}$ and $b^{(2)}$ are given by:

$$\frac{\delta L}{\delta W^{(2)}} = \frac{\delta L}{\delta o^a} (h^s)^T$$

and

$$\frac{\delta L}{\delta b^{(2)}} = \frac{\delta L}{\delta o^a}$$

** (i) the dimensions**

$\frac{\delta L}{\delta W^{(2)}}$ is $m \times d_h$

$\frac{\delta L}{\delta o^a}$ is $m \times 1$

$(h^s)^T$ is $1 \times d_h$

$\frac{\delta L}{\delta W^{(2)}}$ is $m \times 1$

(ii) the weights

$$o^s = softmax(o^a) = softmax(W^{(2)} h^s + b^{(2)})$$

$$f(g(x))' = f'(g(x)) * g'(x)$$

$$\frac{\delta L}{\delta W^{(2)}} = \frac{\delta L}{\delta o^a} * \frac{\delta(W^{(2)} h^s + b^{(2)})}{\delta W^{(2)}} = \frac{\delta L}{\delta o^a} (h^s)^T$$

(iii) the biases

same as for the weights

$$\frac{\delta L}{\delta b^{(2)}} = \frac{\delta L}{\delta o^a} * \frac{\delta(W^{(2)} h^s + b^{(2)})}{\delta b^{(2)}} = \frac{\delta L}{\delta o^a}$$

Q1 e)

Using the chain rule

$$\frac{\delta L}{\delta h_j^s} = \sum_{k=1}^{M} \frac{\delta L}{\delta o_k^a} \frac{\delta o_k^a}{\delta h_j^s}$$

show that the partial derivatives of the cost L with respect to the outputs of the neurons in the hidden layer are given by:

$$\frac{\delta L}{\delta h_j^s} = (W^{(2)})^T \frac{\delta L}{\delta o^a}$$

We start from:

$$\frac{\delta L}{\delta h_j^s} = \sum_{k=1}^{M} \frac{\delta L}{\delta o_k^a} \frac{\delta o_k^a}{\delta h_j^s}$$

$$o_k^a = W_k^{(2)} h_j^s + b_k^{(2)}$$

$$\frac{\delta o_k^a}{\delta h_j^s} = W_k^{(2)}$$

We substitute the derivative term for its value

$$\frac{\delta L}{\delta h_j^s} = \sum_{k=1}^{M} \frac{\delta L}{\delta o_k^a} W_k^{(2)}$$

Which is equivalent in the matrix form to

$$\frac{\delta L}{\delta h^s} = (W^{(2)})^T \frac{\delta L}{\delta o^a}$$

$\frac{\delta L}{\delta o^a}$ is $m \times 1$ and $(W^{(2)})^T$ is $d_h \times m$

Q1 f)

First show that the derivative of $tahn(z) = 1 - tanh^2(z)$

You can see the derivation here: http://math.stackexchange.com/questions/741050/hyperbolic-functions-derivative-of-tanh-x .

It's not really worth the copying.

So the derivative we are looking for is equal to:

$$\frac{\delta L}{\delta h_j^a} = \frac{\delta L}{\delta h_j^s} \frac{\delta h_j^s}{\delta h_j^a}$$

which is therefore equal to

$$\frac{\delta L}{\delta h_j^a} = \frac{\delta L}{\delta h_j^s} \left(1 - tanh^2(h_j^a)\right)$$

$\frac{\delta L}{\delta h_j^s}$ is $d_h \times 1$

Q1 g)

$$\frac{\delta L}{\delta W^{(1)}} = \frac{\delta L}{\delta h^a} \frac{\delta h^a}{\delta W^{(1)}}$$

$$\frac{\delta h^a}{\delta W^{(1)}} = \frac{\delta L}{\delta h^a} x^T$$

and

$$\frac{\delta h^a}{\delta b^{(1)}} = \frac{\delta L}{\delta h^a}$$

since $h^a = W^{(1)} x + b^{(1)}$

Q1 h)

$L_{mod} = \alpha W^{(1)} + \alpha W^{(2)} + \beta(W^{(1)})^2 + \beta(W^{(2)})^2 + L(x,t)$

$\frac{\delta L}{\delta W^{(2)}} = \frac{\delta L}{\delta o^a} (h^s)^T + \alpha + 2\beta W^{(1)}$

$\frac{\delta L}{\delta W^{(2)}} = \frac{\delta L}{\delta o^a} (h^s)^T + \alpha + 2\beta W^{(2)}$

Q1 i)

$$\frac{\delta h^s}{\delta h^a} = 1 \text{ if } h^a > \ 0,\ 0 \text{ otherwise}$$

therefore, only the derivatives depending on this term are affected: that is $\frac{\delta L}{\delta W^{(1)}}$ and $\frac{\delta L}{\delta b^{(1)}}$