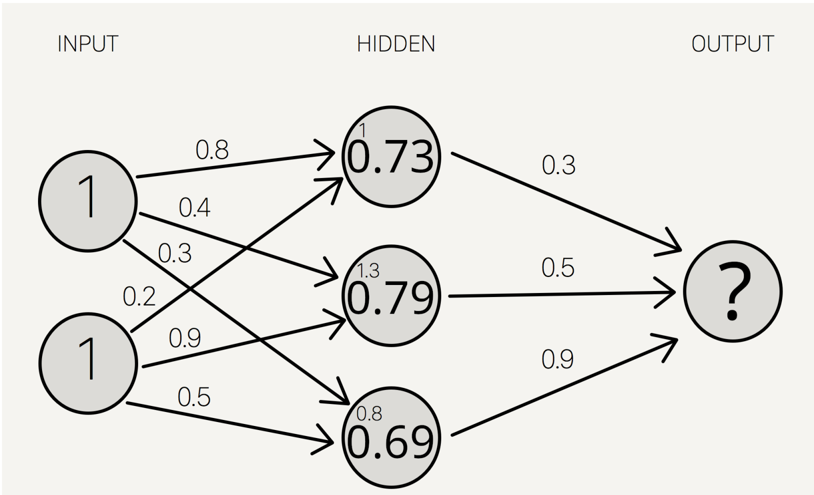
Dense Neural network



Instructions

- Complete the forward computation, given that the activation function is a sigmoid: $\frac{1}{1+e^{-x}}$.
- Calculate the mean square error ($E\left[(\hat{\theta}-\theta)^2\right]$) of the model, given that the expected output is 1.
- Determine the contribution of each node to the error.
- Update the weights using backpropagation, with the derivative of the sigmoid function given by sig'(x)=sig(x)(1-sig(x)) and a learning rate $\lambda = 1$.
- Perform the forward pass with the updated weights and draw conclusions from the results.
- Identify if we applied gradient descent or stochastic gradient descent.
- Explain the difference between gradient descent with momentum and without momentum.

Solution

Notation:
$\times$: Matrix multiplication
$\circ$: Hadamard product

- $$W1 \times L1 = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.9 \\ 0.3 & 0.5 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1.3 \\ 0.8 \end{bmatrix} = l2$$
$$L2 = \begin{bmatrix} sig(1) \\ sig(1.3) \\ sig(0.8) \end{bmatrix} = \begin{bmatrix} 0.73 \\ 0.79 \\ 0.69 \end{bmatrix}$$
$$W2 \times L2 = \begin{bmatrix} 0.3 & 0.5 & 0.9 \end{bmatrix} \times \begin{bmatrix} 0.73 \\ 0.79 \\ 0.69 \end{bmatrix} = \begin{bmatrix} 1.24 \end{bmatrix} = l3$$
$$L3 = sig(l3) = \begin{bmatrix} sig(1.24) \end{bmatrix} = \begin{bmatrix} 0.78 \end{bmatrix}$$
- ErrL3=$\frac{\begin{bmatrix}(0.78-1)^2\end{bmatrix}}{1} = 0.22$
- contribNorm(W2^T)= $\begin{bmatrix} \frac{0.3}{0.3+0.5+0.9} \\ \frac{0.5}{0.3+0.5+0.9} \\ \frac{0.9}{0.3+0.5+0.9} \end{bmatrix} = \begin{bmatrix} 0.18 \\ 0.29 \\ 0.53 \end{bmatrix}$

  ErrL2= contribNorm(W2^T) $\times$ ErrL3 $= \begin{bmatrix} 0.04 \\ 0.06 \\ 0.11 \end{bmatrix}$

  contribNorm(W1^T)= $\begin{bmatrix} \frac{0.8}{0.8+0.2} & \frac{0.4}{0.4+0.9} & \frac{0.3}{0.3+0.5} \\ \frac{0.2}{0.8+0.2} & \frac{0.9}{0.4+0.9} & \frac{0.5}{0.3+0.5} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.31 & 0.375 \\ 0.2 & 0.61 & 0.625 \end{bmatrix}$

  ErrL1= contribNorm(W1^T) $\times$ ErrL2 $= \begin{bmatrix} 0.06 \\ 0.11 \end{bmatrix}$
- sig'(l3)=L3$\circ$(1-L3)=$\begin{bmatrix} 0.78 \end{bmatrix} (1 - \begin{bmatrix} 0.78 \end{bmatrix})$= $\begin{bmatrix} 0.17 \end{bmatrix}$

  NewW2^T=W2^T+λL2$\times$(ErrL3$\circ$sig'(l3))^T

  $= \begin{bmatrix} 0.3 \\ 0.5 \\ 0.9 \end{bmatrix} + \lambda \begin{bmatrix} 0.73 \\ 0.79 \\ 0.69 \end{bmatrix} (\begin{bmatrix} 0.22 \end{bmatrix} \circ \begin{bmatrix} 0.17 \end{bmatrix})^T$

  $= \begin{bmatrix} 0.3 \\ 0.5 \\ 0.9 \end{bmatrix} + \lambda \begin{bmatrix} 0.73 * 0.04 \\ 0.79 * 0.04 \\ 0.69 * 0.04 \end{bmatrix}$

  $= \begin{bmatrix} 0.3 \\ 0.5 \\ 0.9 \end{bmatrix} + \lambda \begin{bmatrix} 0.03 \\ 0.03 \\ 0.03 \end{bmatrix}$

  $= \begin{bmatrix} 0.33 \\ 0.53 \\ 0.93 \end{bmatrix}$

sig'(l2)=L2∘(1-L2)= $\begin{bmatrix} 0.20 \\ 0.17 \\ 0.23 \end{bmatrix}$

NewW1^T=W1^T+λL1×(ErrL2∘sig'(l2))^T

$=\begin{bmatrix} 0.8 & 0.4 & 0.3 \\ 0.2 & 0.9 & 0.5 \end{bmatrix} + \lambda \begin{bmatrix} 1 \\ 1 \end{bmatrix} (\begin{bmatrix} 0.04 \\ 0.06 \\ 0.11 \end{bmatrix} \circ \begin{bmatrix} 0.20 \\ 0.17 \\ 0.23 \end{bmatrix})^T$

$= \begin{bmatrix} 0.8 & 0.4 & 0.3 \\ 0.2 & 0.9 & 0.5 \end{bmatrix} + \lambda \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0.04 * 0.20 \\ 0.06 * 0.17 \\ 0.11 * 0.23 \end{bmatrix}^T$

$= \begin{bmatrix} 0.8 & 0.4 & 0.3 \\ 0.2 & 0.9 & 0.5 \end{bmatrix} + \lambda \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0.01 & 0.01 & 0.03 \end{bmatrix}$

$= \begin{bmatrix} 0.8 & 0.4 & 0.3 \\ 0.2 & 0.9 & 0.5 \end{bmatrix} + \lambda \begin{bmatrix} 0.01 & 0.01 & 0.03 \\ 0.01 & 0.01 & 0.03 \end{bmatrix} = \begin{bmatrix} 0.81 & 0.41 & 0.33 \\ 0.21 & 0.91 & 0.53 \end{bmatrix}$

- $$NewW1 * L1 = \begin{bmatrix} 0.81 & 0.21 \\ 0.41 & 0.91 \\ 0.33 & 0.53 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.02 \\ 1.32 \\ 0.86 \end{bmatrix} = l2$$

  L2= $\begin{bmatrix} sig(1.02) \\ sig(1.32) \\ sig(0.86) \end{bmatrix} = \begin{bmatrix} 0.73 \\ 0.79 \\ 0.70 \end{bmatrix}$

  NewW2*L2= $\begin{bmatrix} 0.33 & 0.53 & 0.93 \end{bmatrix} \times \begin{bmatrix} 0.73 \\ 0.79 \\ 0.70 \end{bmatrix} = \begin{bmatrix} 1.31 \end{bmatrix} = l3$

  L3=sig(l3)=0.79

  We have successfully improved the prediction accuracy by reducing the error from $(0.78 - 1)^2 = 0.05$ to $(0.79 - 1)^2 = 0.04$.

- Gradient descent was used because the batch size is 1. It is not stochastic because it does not rely on batch sampling.

- Gradient Descent with Momentum includes an extra term called momentum, which is a fraction of the previous weight update, this fraction isdetermined by the momentum factor β. This smooths out the updates, accelerates convergence, and reduces oscillations by enabling the algorithm to maintain its direction and build up speed along the most promising path.