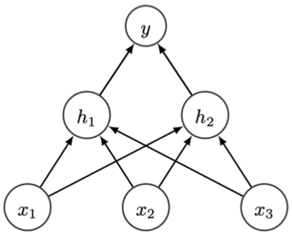
Mathematical Foundations of Machine Learning Assignment

David Felipe Alvear Goyes

1. Backpropagation.
The following graph shows the structure of a simple neural network with a single hidden layer. The input layer consists of three features $x = (x_1, x_2, x_3)$. The hidden layer includes two units $h = (h_1, h_2)$. The output layer includes one unit $\hat{y}$. We ignore bias terms for simplicity.



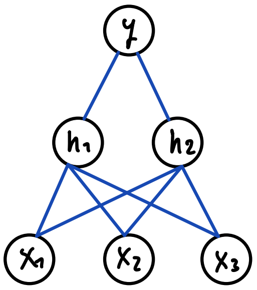
We use ReLU units $\varphi(z) = \max(0, z)$ as activation functions for the hidden and the output layer.
Let $J = \ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ be the loss function. Here y is the target value for the output unit.
Denote by W and V weight matrices connecting input and hidden layer, and hidden layer and output respectively. They are initialized as follows:

$$W = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} \text{ and } V = \begin{bmatrix} 0 & 1 \end{bmatrix}$$

- write out symbolically (in terms of φ, W, and V) the function $x \to \hat{y}$ represented by this network.
- consider a data sample with features $x = (1, 2, 1)$ and target $y = 1$. Compute the output value $\hat{y}$.
- compute the gradient of the loss function with respect to the weights. Specifically,
  - write down symbolically the gradient with respect to V, $\frac{\partial J}{\partial V}$
  - write down symbolically the gradient with respect to W, $\frac{\partial J}{\partial W}$
  - compute the values of this gradient numerically for the parameters above

Solution

→ write symbolically the function $x \to \hat{y}$



$h = \phi(w.x)$ → h hidden layer

$\hat{y} = \phi(v.h)$

$\hat{y}$ → Predicted output

ϕ → ReLU activation function

$\phi(z) = max(0, z)$

$\hat{y} = max(0, v\, max(0, w\, x))$

$\hat{y} = \phi(v.\phi(w.x))$

→ Considere $x=(1,2,1)$ $y=1$, compute $\hat{y}$.

In matrix form

$$\hat{y} = \max\left[0, \left[\underbrace{\begin{bmatrix}0 & 1\end{bmatrix}}_{V}\underbrace{\left[\max\left[0, \underbrace{\begin{bmatrix}1&0&1\\1&-1&0\end{bmatrix}}_{W}\underbrace{\begin{bmatrix}1\\2\\1\end{bmatrix}}_{X}\right]\right]}_{\text{hidden layer}}\right]\right]$$

$$\hat{y} = \max\left[0, \left[\underbrace{\begin{bmatrix}0 & 1\end{bmatrix}}_{V}\left[\max\left[0, \begin{bmatrix}2\\-1\end{bmatrix}\right]\right]\right]\right] = \max\left[0, \left[\underbrace{\begin{bmatrix}0 & 1\end{bmatrix}}_{V}\begin{bmatrix}2\\0\end{bmatrix}\right]\right]$$

$$\hat{y} = \max\left[0, \begin{bmatrix}0\\0\end{bmatrix}\right] = 0 \quad \rightarrow \quad \boxed{\hat{y} = 0}$$

→ loss function: $J(y,\hat{y}) = \frac{1}{2}(y-\hat{y})^2 = \frac{1}{2}$ $\boxed{\frac{dJ}{d\hat{y}} = (\hat{y}-y) = -1}$

→ write $\frac{dJ}{dV} = \frac{dJ}{d\hat{y}}\frac{d\hat{y}}{dV} = (\hat{y}-y)\,\phi'(v.h)\,h$

→ write $\frac{dJ}{dw} = \frac{dJ}{d\hat{y}}\frac{d\hat{y}}{dh}\frac{dh}{dw} = (\hat{y}-y)\,\phi'(v.h)\,V\,\phi(w\,x)\,X$

$$h = \phi\left[\begin{bmatrix}1&0&1\\1&-1&0\end{bmatrix}\begin{bmatrix}1\\2\\1\end{bmatrix}\right] = \phi\begin{bmatrix}2\\-1\end{bmatrix} = \begin{bmatrix}2\\0\end{bmatrix}$$

$$\hat{y} = \phi\left[\begin{bmatrix}0&1\end{bmatrix}\begin{bmatrix}2\\0\end{bmatrix}\right] = 0$$

→ Numerically: $\frac{dJ}{dV} = (\hat{y}-y)\,\phi'(v.h)\,h = (-1)\,(0)\begin{bmatrix}2\\0\end{bmatrix} = \begin{bmatrix}-2\\0\end{bmatrix}$

$$\frac{dJ}{dw} = (\hat{y}-y)\,\phi'(v.h)\,V\,\phi'(w\,x)\,X = 0$$

5. Sentiment analysis with trainable embeddings

Instead of using fixed (pre-trained) word embeddings for the sentiment classification task as we did in the previous exercise, we will now make the word embeddings trainable as well. We will use an embedding of dimension $D = 50$.

- the classification network for this problem consists, as above, of an embedding layer, an averaging operation (with no trainable variables), a dense hidden layer of size 32, and a final dense output layer. Write down the mathematical description of the operations involved in evaluating the function defined by this network.

$E = \frac{1}{N} \sum_{i=1}^{N} E_i$ → Embedding layer + averaging operation

$H = \phi(W_1 . E + b_1)$ → Hidden layer $\quad W_1^{32 \times 50}$

$\hat{y} = W_2 . H + b_2$ → output $\quad W_2^{1 \times 32}$