Notation for ML

Joel Laity

April 11, 2018

0.1 Indexing conventions

- $\square^{[\ell]}$ - ℓ-th layer.
- $\square^{(i)}$ - i-th example.
- $\square_j$ - j-th node in layer.
- Capital letter - Matrix where the i-th column (or row) is the i-th example of lowercase variable, e.g.

$$X = \begin{pmatrix} | & | & & | \\ x^{(0)} & x^{(1)} & \dots & x^{(m-1)} \\ | & | & & | \end{pmatrix}$$

0.2 Shape of network

The hyperparameters that determine the shape of the network are as follows.

- $L \in \mathbb{N}$ or `L` - the number of layers excluding input and output layer.
- $n^{[\ell]} \in \mathbb{N}$ or `layer_dims[l]` - the number of nodes in the ℓ-th layer. The input layer is $\ell = 0$ and $\ell = L$ is output layer, so $n^{[0]} = n_x$ and $n^{[L]} = n_y$.
- $n \in \mathbb{N}^{L+1}$ or `layer_dims` - the vector n is a row vector

$$n = \begin{pmatrix} n^{[0]}, & n^{[1]}, & \dots, & n^{[L]} \end{pmatrix} \in \mathbb{R}^{L+1}.$$

0.3 Parameters of the network

The goal of the learning algorithm is to learn the parameters for each node in the vertex. The j-th node in the ℓ-th layer has parameters $w_j^{[\ell]}$ and $b_j^{[\ell]}$ as shown on the last page of this document.

0.3.1 The weighting vector

- $w_j^{[\ell]} \in \mathbb{R}^{n[\ell-1]}$ or `Wl[j]` - the weighting vector in the j-th node of the ℓ-level. Note that since $w_j^{[\ell]}$ is a parameter of the network, it does not depend on any example. It's normally presented as a column vector

$$w_j^{[\ell]} = \overbrace{\begin{pmatrix} * & * & \cdots & * \end{pmatrix}}^{n^{[\ell-1]} \text{ times}}$$

- $W^{[\ell]}$ - the $n^{[\ell-1]} \times n^{[\ell]}$ matrix with $w_j^{[\ell]}$ as the j-th row vector

$$W^{[\ell]} = \begin{pmatrix} — & w_1^{[\ell]} & — \\ — & w_2^{[\ell]} & — \\ & \vdots & \\ — & w_{n^{[\ell]}}^{[\ell]} & — \end{pmatrix}.$$

0.3.2 The bias vector

- $b_j^{[\ell]} \in \mathbb{R}$ or `bl[j]` - the bias in the j-th node in the ℓ-th level.
- $b^{[\ell]} \in \mathbb{R}^{n^{[\ell]}}$ or `bl` - a column vector giving the bias at each node in the ℓ-th level

$$b^{[\ell]} = \begin{pmatrix} b_1^{[\ell]} \\ b_2^{[\ell]} \\ \vdots \\ b_{n^\ell}^{[\ell]} \end{pmatrix}$$

0.4 Input and output data

- m or `m` - number of training examples in dataset (number of pictures).
- n_x or `n_x` - size of each data point in data set (number of pixels per picture).
- $n_y = C$ or `n_y` - size of output (number of classifications of the pictures e.g. cat, dog, koala, other).

0.4.1 The example vector X

- $x^{(i)} \in \mathbb{R}^{n_x}$ or `X[i]` - the i-th example as a column vector (pixels as a column vector).

$$x^{(i)} = \begin{pmatrix} x_0^{(i)} \\ x_1^{(i)} \\ \vdots \\ x_{n_x}^{(i)} \end{pmatrix}$$

- X or `X` - The $n_x \times m = n^{[0]} \times m$ matrix containing all examples as column vectors

$$X = \begin{pmatrix} | & | & & | \\ x^{(0)} & x^{(1)} & \dots & x^{(m-1)} \\ | & | & & | \end{pmatrix} = \begin{pmatrix} x_0^{(0)} & x_0^{(1)} & \dots & x_0^{(m-1)} \\ x_1^{(0)} & x_1^{(1)} & \dots & x_1^{(m-1)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_x-1}^{(0)} & x_{n_x-1}^{(1)} & \dots & x_{n_x-1}^{(m-1)} \end{pmatrix}.$$

0.4.2 The classification vector y

- $y^{(i)} \in \mathbb{R}^{n_y}$ - the correct output label for example $x^{(i)}$ (so $y^{(i)} \in \{0,1\}^{n_y}$ and the Hamming weight of $y^{(i)}$ is 1 (the nonzero entry corresponds to the correct classification cat, dog, koala, other etc.).

- y - the $n_y \times m = C \times m$ matrix containng all the correct outputs as column vectors
$$y = \begin{pmatrix} | & | & & | \\ y^{(0)} & y^{(1)} & \dots & y^{(m-1)} \\ | & | & & | \end{pmatrix} \in \{0,1\}^{n_y \times m}.$$

- $\hat{y}^{(i)} = a^{[L](i)} \in \mathbb{R}^{n_y}$ - The vector of probabilities where $\hat{y}_j^{(i)}$ is the probability that $x^{(i)}$ belongs to the j-th category.

- $\hat{y} \in \mathbb{R}^{n_y \times m}$ - the $n_y \times m = C \times m$ matrix containng all the probability vectors as columns
$$\hat{y} = \begin{pmatrix} | & | & & | \\ \hat{y}^{(0)} & \hat{y}^{(1)} & \dots & \hat{y}^{(m-1)} \\ | & | & & | \end{pmatrix} \in \mathbb{R}^{n_y \times m}.$$

0.5 Forward propagation

0.5.1 The activation functions

- $g^{[\ell]} : \mathbb{R}^{n^{[\ell]}} \to \mathbb{R}^{n^{[\ell]}}$ - the activation function in the ℓ-th layer. Usually $g^{[\ell]}$ is just a real-valued function of one real variable, such as the sigmoid function, applied component-wise.

0.5.2 The neuron values A

- $a_j^{[\ell](i)} \in \mathbb{R}$ or `A[i][j]` - the output value of the j-th neuron in the ℓ-th layer when calculating the i-th example.

- $a^{[\ell](i)} \in \mathbb{R}^{n^{[\ell]}}$ or `A[i]`- the output of the ℓ-layer on the i-th example as a column vector
$$a^{[\ell](i)} = \begin{pmatrix} a_1^{[\ell](i)} \\ a_2^{[\ell](i)} \\ \vdots \\ a_{n^{[\ell]}}^{[\ell](i)} \end{pmatrix} \in \mathbb{R}^{n^{[\ell]}}$$

- $a^{[\ell]} \in \mathbb{R}^{m \times n^{[\ell]}}$ or `A` - the output of the ℓ-th layer over all examples.

-

0.6 The preactivation value Z

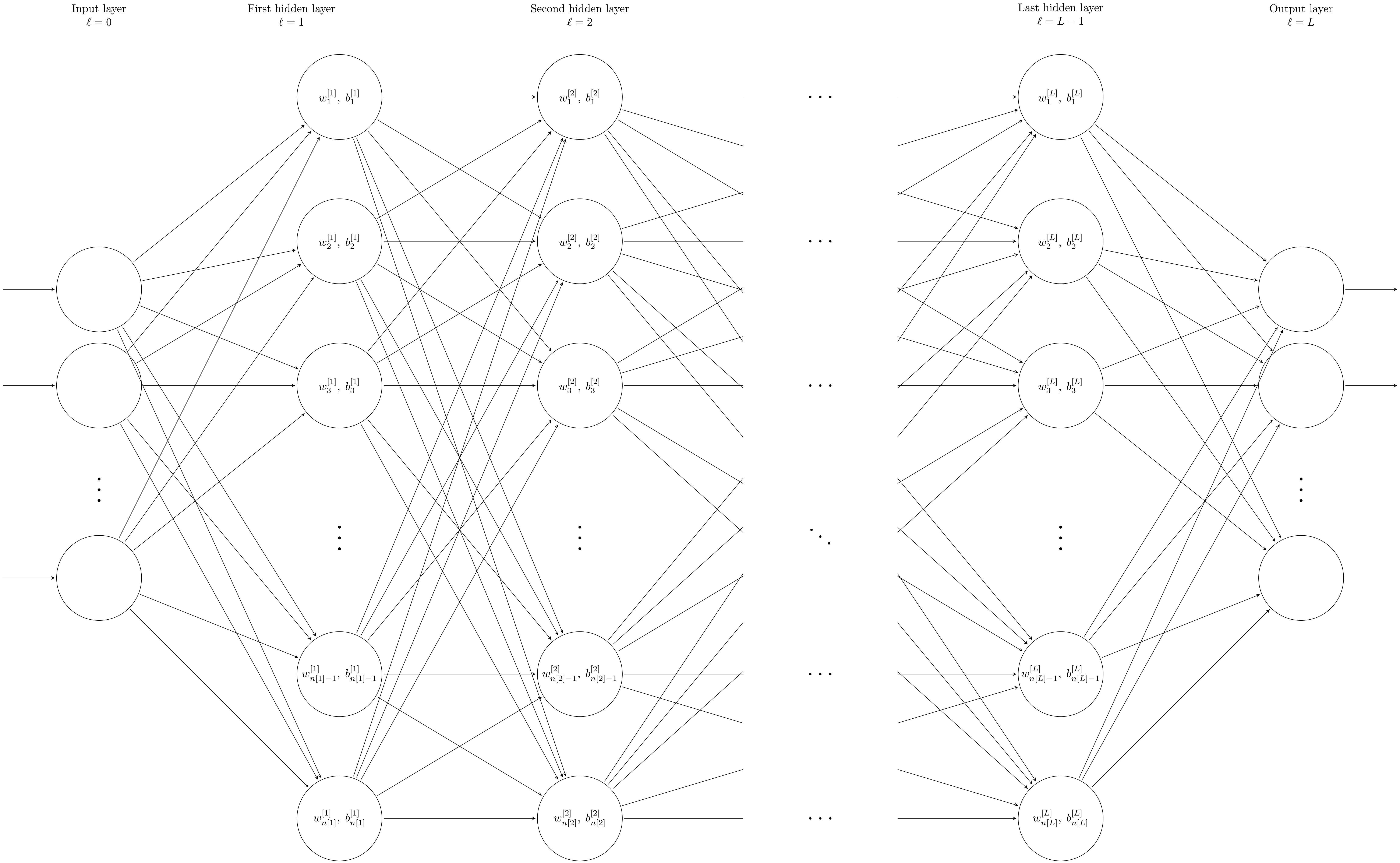
- $z_j^{[\ell](i)} \in \mathbb{R}$ or `Z[i][j]` - the preactivation value of the j-th node in the ℓ-th layer when calculating the i-th example

$$z_j^{[\ell](i)} = w_1^{[\ell]} a^{[\ell-1](i)} + b_j^{[\ell]} \in \mathbb{R}.$$

- $z^{[\ell](i)} \in \mathbb{R}^{[n^{[\ell]}}$ or `Z[i]` - the preactivation values of the ℓ-th layer on the i-th example as a column vector

$$z^{[\ell](i)} = \begin{pmatrix} z_1^{[\ell](i)} \\ z_2^{[\ell](i)} \\ \vdots \\ z_{n^{[\ell]}}^{[\ell](i)} \end{pmatrix} \in \mathbb{R}^{n^{[\ell]}}$$

- $z^{[\ell]} \in \mathbb{R}^{m \times n^{[\ell]}}$ or `Z` - the output of the ℓ-th layer over all examples.



Input layer
$\ell = 0$
First hidden layer
$\ell = 1$
Second hidden layer
$\ell = 2$
Last hidden layer
$\ell = L-1$
Output layer
$\ell = L$
$w_1^{[1]},\ b_1^{[1]}$
$w_2^{[1]},\ b_2^{[1]}$
$w_3^{[1]},\ b_3^{[1]}$
$w_{n[1]-1}^{[1]},\ b_{n[1]-1}^{[1]}$
$w_{n[1]}^{[1]},\ b_{n[1]}^{[1]}$
$w_1^{[2]},\ b_1^{[2]}$
$w_2^{[2]},\ b_2^{[2]}$
$w_3^{[2]},\ b_3^{[2]}$
$w_{n[2]-1}^{[2]},\ b_{n[2]-1}^{[2]}$
$w_{n[2]}^{[2]},\ b_{n[2]}^{[2]}$
$w_1^{[L]},\ b_1^{[L]}$
$w_2^{[L]},\ b_2^{[L]}$
$w_3^{[L]},\ b_3^{[L]}$
$w_{n[L]-1}^{[L]},\ b_{n[L]-1}^{[L]}$
$w_{n[L]}^{[L]},\ b_{n[L]}^{[L]}$