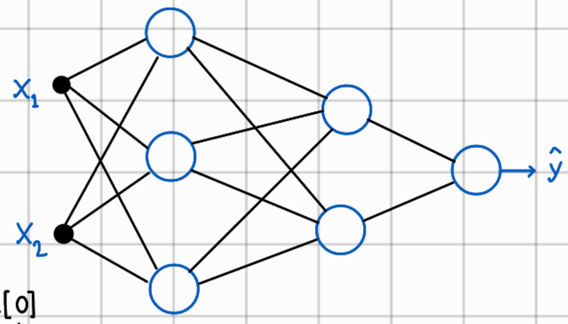
PHY654

Machine Learning in Particle Physics

Thursday, 22 August, 2024

- Initialization of w & b parameters:

- Importance of vectorization



$x = a[0]$ → layer

$$x = a[0] = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Layer 1 (3 units)

$$Z^{[1]} = W^{[1]} a^{[0]} + b^{[1]}$$

$$= \underset{3\times 2}{\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{pmatrix}} \underset{2\times 1}{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}} + \underset{3\times 1}{\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}}$$

$$= \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

where $$\boxed{a^{[1]} = g^{[1]}\left(Z^{[1]}\right) = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}}$$

and $g^{[1]} = \tanh$

Layer 2 :

$$Z^{[2]} = W^{[2]} a^{[1]} + b^{[2]} = \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$$

Now, $a^{[2]} = g^{[2]}\left(Z^{[2]}\right) = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ (say)

Layer 3 : $Z^{[3]} = W^{[3]} a^{[2]} + b^{[3]} = \begin{pmatrix} w_1 & w_2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + b_1 = \begin{pmatrix} Z_1 \end{pmatrix}$

$$a^{[3]} = \hat{y} = \underset{\downarrow \text{ sigmoid}}{g^{[3]}}\left(Z^{[3]}\right)$$

General Rule (l-th layer):

$$Z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]}, \quad a^{[l]} = g^{[l]}\left(Z^{[l]}\right)$$

m training example   Vectorization

$$Z^{[\ell]} = W^{[\ell]} A^{[\ell-1]} + b^{[\ell]}$$

$$A^{[\ell]} = g^{[\ell]}\left(Z^{[\ell]}\right)$$

$$Z^{[\ell]} = \begin{pmatrix} z^{[\ell]1} & z^{[\ell]2} & \cdots & z^{[\ell]m} \end{pmatrix} \quad A^{[0]} = X = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(m)} \end{pmatrix}$$

- ## Regularization

1. Logistic Regression (L_2 regularization)

   Minimize the cost function $J(w, b) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right) + \frac{\lambda}{2m}\|w\|_2^2$

   regularization parameter
   another hyperparameter

   L_2-norm of w

   $$\|w\|_2^2 = w^T w \rightarrow L_2 \text{ regularization}$$
   $$= \sum_{j=1}^{n} w_j^2$$

Next, consider : (for NN (neural network))

$$J\left(w^{(1)}, b^{(1)}; \ldots; w^{(L)}, b^{(L)}\right) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right) + \frac{\lambda}{2m}\sum_{\ell=1}^{L}\|w^{(\ell)}\|_F^2$$

$$\|w^{(\ell)}\|_F^2 = \sum_{i=1}^{n^{[\ell-1]}} \sum_{j=1}^{n^{[\ell]}} \left(w_{ij}^{[\ell]}\right)^2 \longrightarrow \text{Frobenius norm}$$

[now w is a matrix, instead of a row/column vector]

Also called weight decay

2. Drop-out Regularization:

3. Early stopping Regularization