Artificial Neural Networks

Lecture 5: Multi-Layer Networks

Dr Juan Neirotti

j.p.neirotti@aston.ac.uk

2023

Multi-Layer Networks

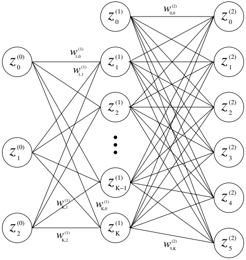
- 

Figure: Feed-forward network with 3 layers, input, hidden and output.

- In figure 1 we have a feed-forward networks with an input layer with 2+1 units, a hidden layer with K+1 units and an output layer with 6 units.
- The input units $z_k^{(0)}$ can be identified with the inputs x_k, which are the entries of the input vector $\boldsymbol{x} \in \mathscr{X}$.

Multi-Layer Networks

- The general form of the activation functions is:

$$z_k^{(\ell+1)} = \sigma\left(\left\langle \boldsymbol{w}_k^{(\ell)} \middle| \boldsymbol{z}^{(\ell)} \right\rangle + w_0^{(\ell)}\right),$$

where:

$$\lim_{a \to -\infty} \sigma(a) = 0 \qquad (\text{or } -1),$$
$$\lim_{a \to \infty} \sigma(a) = 1.$$

Functions like these are called sigmoidal.

Regression Problem: Cybenko's Theorem

- Theorem: Let $\sigma(x)$ be a bounded sigmoidal function, and $f : \mathbb{R} \to \mathbb{R}$ continuous, satisfying $\lim_{x \to -\infty} f(x) = A$ and $\lim_{x \to \infty} f(x) = B$, where A and B are constants, then for any $\varepsilon > 0$, there exists N, c_i, y_i, and θ_i, such that:

$$\left| f(x) - \sum_{i=1}^{N} c_i \sigma \left(y_i x + \theta_i \right) \right| < \varepsilon$$

holds for all $x \in \mathbb{R}$.

Proof

- By continuity of f we have that for all $\varepsilon > 0$ there must exists $M \in \mathbb{N}$ such that $|f(x) - A| < \frac{\varepsilon}{4}$ if $x < -M$; $|f(x) - B| < \frac{\varepsilon}{4}$ if $x > M$; and $|f(x') - f(x'')| < \frac{\varepsilon}{4}$ if $|x'| \leq M$, $|x''| \leq M$, and $|x' - x''| \leq \frac{1}{M}$.
- Let us divide the interval $[-M, M]$ into $N = 2M^2$ equal segments of length $\frac{1}{M}$, and let define: $x_i = -M + \frac{i}{M}$.
- Let us also define $t_i = \frac{x_i + x_{i+1}}{2}$, which is the center of the interval $[x_i, x_{i+1}]$.
- Now let us construct:

$$g(x) = f(-M) + \sum_{i=1}^{N} [f(x_i) - f(x_{i-1})] \, \sigma(K(x - t_{i-1})).$$

- By the properties of the sigmoidal function there exists $W > 0$ such that $1 - \sigma(u) < \frac{1}{M^2}$ if $u > W$ and $\sigma(u) < \frac{1}{M^2}$ if $u < -W$. Let us choose $K = 2MW$.
- If $x < -M$, then

$$|f(x) - f(-M)| < |f(x) - A| + |A - f(-M)| < \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}.$$

- Also:

$$K(x - t_{i-1}) = -K|x + M| - W(2i - 1) < -W$$

$$\begin{aligned} |g(x) - f(-M)| &= \left| \sum_{i=1}^{N} [f(x_i) - f(x_{i-1})] \, \sigma\left(K(x - t_{i-1})\right) \right| \\ &< \sum_{i=1}^{N} |f(x_i) - f(x_{i-1})| \frac{1}{M^2} \\ &< 2M^2 \frac{\varepsilon}{4} \frac{1}{M^2} = \frac{\varepsilon}{2}. \end{aligned}$$

.

- In consequence $|g(x) - f(x)| < |g(x) - f(-M)| + |f(-M) - f(x)| < \varepsilon$ for all $x < -M$.
- If $x > M$, then

$$|f(x) - f(M)| < |f(x) - B| + |B - f(M)| < \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}.$$

- Also:

$$K(x - t_{i-1}) \geq K(x - M) + \frac{K}{2M} > W.$$

- Then:

$$
\begin{aligned}
g(x) &= f(-M) + \sum_{i=1}^{N} [f(x_i) - f(x_{i-1})] \{\sigma(K(x - t_{i-1})) - 1 + 1\} \\
&= f(-M) + \sum_{i=1}^{N} [f(x_i) - f(x_{i-1})] \{\sigma(K(x - t_{i-1})) - 1\} \\
&+ \sum_{i=1}^{N} [f(x_i) - f(x_{i-1})] \\
&= f(-M) + \sum_{i=1}^{N} [f(x_i) - f(x_{i-1})] \{\sigma(K(x - t_{i-1})) - 1\} \\
&+ f(x_N) - f(x_0) \\
&= f(M) + \sum_{i=1}^{N} [f(x_i) - f(x_{i-1})] \{\sigma(K(x - t_{i-1})) - 1\} \\
|g(x) - f(M)| &< \sum_{i=1}^{N} |f(x_i) - f(x_{i-1})| \{1 - \sigma(K(x - t_{i-1}))\} \\
&< 2M^2 \frac{\varepsilon}{4} \frac{1}{M^2} = \frac{\varepsilon}{2}.
\end{aligned}
$$

.

- In consequence $|g(x) - f(x)| < |g(x) - f(M)| + |f(M) - f(x)| < \varepsilon$ for all $x > M$.
- If $-M < x < M$ there exists a $0 \le k \le 2M^2$ such that $x \in [x_{k-1}, x_k]$ and then $|x - t_{i-1}| \le \frac{1}{2M}$ if $i = k$ and $|x - t_{i-1}| > \frac{1}{2M}$ if $i \neq k$..
- Furthermore, if $i < k$ then $x > t_{i-1}$ and:

  $$K(x - t_{i-1}) > \frac{K}{2M} = W,$$

  and

  $$1 - \sigma\left(K(x - t_{i-1})\right) < \frac{1}{M^2}.$$

- If $i > k$ then $x < t_{i-1}$ and:

  $$K(x - t_{i-1}) < -\frac{K}{2M} = -W,$$

  and

  $$\sigma\left(K(x - t_{i-1})\right) < \frac{1}{M^2}.$$

.

- Consequently we have:

$$
\begin{aligned}
g(x) = {} & f(-M) + \sum_{i=1}^{k-1} [f(x_i) - f(x_{i-1})] \, \sigma(K(x - t_{i-1})) + \\
& + [f(x_k) - f(x_{k-1})] \, \sigma(K(x - t_{k-1})) + \\
& + \sum_{i=k+1}^{N} [f(x_i) - f(x_{i-1})] \, \sigma(K(x - t_{i-1})) \\
= {} & f(-M) + \sum_{i=1}^{k-1} [f(x_i) - f(x_{i-1})] \{\sigma(K(x - t_{i-1})) - 1\} + \\
& + f(x_{k-1}) - f(x_0) + \\
& + [f(x_k) - f(x_{k-1})] \, \sigma(K(x - t_{k-1})) + \\
& + \sum_{i=k+1}^{N} [f(x_i) - f(x_{i-1})] \, \sigma(K(x - t_{i-1}))
\end{aligned}
$$

- Therefore

$$
\begin{aligned}
A &= \left|g(x) - f(x_{k-1}) - [f(x_k) - f(x_{k-1})]\,\sigma\left(K(x - t_{k-1})\right)\right| \\
&< \sum_{i=1}^{k-1} |f(x_i) - f(x_{i-1})|\,\sigma\left(K(x - t_{i-1})\right) + \\
&+ \sum_{i=k+1}^{N} |f(x_i) - f(x_{i-1})| \left\{1 - \sigma\left(K(x - t_{i-1})\right)\right\} \\
&< \sum_{i=1}^{k-1} \frac{\varepsilon}{4}\frac{1}{M^2} + \sum_{i=k+1}^{N} \frac{\varepsilon}{4}\frac{1}{M^2} = \frac{2M^2 - 1}{2M^2}\frac{\varepsilon}{2} < \frac{\varepsilon}{2}.
\end{aligned}
$$

- Finally:

$$
\begin{aligned}
|g(x) - f(x)| &< |f(x_{k-1}) - f(x)| + \frac{\varepsilon}{4}\sigma\left(K(x - t_{k-1})\right) + \frac{\varepsilon}{2} \\
&< \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{2} = \varepsilon. \clubsuit
\end{aligned}
$$

Classification Problem: Grand Mother Cells

- Consider $\boldsymbol{x} \in \mathscr{X} = \{\pm 1\}^d \subset \mathbb{R}^d$.
- Let us verify that any dichotomy $\{\mathscr{X}^+, \mathscr{X}^-\}$ of $\mathscr{X}$ can be realized by a neural network with one hidden layer.
- Let us suppose that $|\mathscr{X}^+| \le |\mathscr{X}^-|$. Let us consider the vector $\boldsymbol{x}_k^+ \in \mathscr{X}^+$, for any $1 \le k \le |\mathscr{X}^+|$.

Observe that

$$\langle \boldsymbol{x} \,|\boldsymbol{x}_k^+ \rangle \begin{cases} < d-1 & \mathscr{X} \ni \boldsymbol{x} \neq \boldsymbol{x}_k^+ \\ = d & \mathscr{X} \ni \boldsymbol{x} = \boldsymbol{x}_k^+ . \end{cases}$$

- Therefore

$$\operatorname{sgn}\left(\langle \boldsymbol{x} \,|\boldsymbol{x}_k^+ \rangle - d + \frac{1}{2}\right) = \begin{cases} 1 & \boldsymbol{x} = \boldsymbol{x}_k^+ \\ -1 & \boldsymbol{x} \neq \boldsymbol{x}_k^+ . \end{cases}$$

- The network with $|\mathscr{X}^+| = K \leq 2^{d-1}$ units in tis hidden layer, each one of them implementing the activation function $z_k^{(1)} = \operatorname{sgn}\left(\langle \boldsymbol{x} \,|\boldsymbol{x}_k^+ \rangle - d + \frac{1}{2}\right)$, i.e. $\boldsymbol{w}_k^{(1)} = \boldsymbol{x}_k^+$ and $w_{0,k} = -d + \frac{1}{2}$, and an output layer that implements

  $z^{(2)} = \operatorname{sgn}\left(\sum_{k=1}^{K} z_k^{(1)} + K - \frac{1}{2}\right)$,i.e. $\left\langle \boldsymbol{w}^{(2)} \right| = \overbrace{(1, \ldots, 1)}^{K}$, and $w_0^{(2)} = K - \frac{1}{2}$, produces the correct classification of the dichotomy $\{\mathscr{X}^+, \mathscr{X}^-\}$.
- The problem with this approach is that the total number of units needed may be 2^{d-1}, which could be very large indeed.

General Feed-Forward Networks (with a hidden layer)

- We consider the case of a network with one hidden layer and one linear output.
- Linear outputs are usually used for regression problems (binary outputs are used for classification)
- Let us consider the data set $\mathcal{D} = \{(\boldsymbol{y}_\ell, \boldsymbol{t}_\ell)\}_{\ell=1}^{L}$ with $\boldsymbol{y}_\ell \in \mathscr{Y}$ and $\boldsymbol{t}_\ell \in \mathbb{R}^o$ for all $1 \leq \ell \leq L$, and let us suppose there is a continuous function $\boldsymbol{h} : \mathscr{Y} \to \mathbb{R}^o$ such that $\boldsymbol{h}(\boldsymbol{y}_\ell) = t_\ell$ for all $1 \leq \ell \leq L$.
- Observe that $\langle \boldsymbol{h}| = (h_1, \ldots, h_o)$ where $h_n : \mathscr{Y} \to \mathbb{R}$ for all $1 \leq n \leq o$.

- By Cybenko's approach we should be able to construct a network of sigmoidal units such that the Kth approximation:

$$z_n^{(2)}(\boldsymbol{y}) = w_{0,n}^{(2)} + \sum_{k=1}^{K} w_{k,n}^{(2)} \sigma \left(w_{k,0}^{(1)} + \sum_{i=1}^{d} w_{k,i}^{(1)} y_i \right),$$

is as close as $h_n(\boldsymbol{y})$ as required.
- Inputs $\boldsymbol{y} \in \mathscr{Y} \subset \mathbb{R}^d$ are augmented to $\boldsymbol{x} \in \{1\} \times \mathscr{Y} \subset \mathbb{R}^{d+1}$.
- Let us identify the variables $\boldsymbol{z}^{(0)} = \boldsymbol{x}$. Observe that $z_0^{(0)} = 1$ always.
- The kth *activation* of the hidden layer is given by

$$a_k^{(1)} = \left\langle \boldsymbol{x} \,\middle|\, \boldsymbol{w}_k^{(1)} \right\rangle = \left\langle \boldsymbol{z}^{(0)} \,\middle|\, \boldsymbol{w}_k^{(1)} \right\rangle$$

for a suitable $\boldsymbol{w}_k^{(1)} \in \Omega^{(1)} \subset \mathbb{R}^{d+1}$.

- The kth *activation function* $(1 \leq k \leq K)$ is implemented by a sigmoidal function σ :

$$z_k^{(1)} = \sigma(a_k^{(1)}),$$

where

$$\lim_{x \to \infty} \sigma(x) = 1$$
$$\lim_{x \to -\infty} \sigma(x) = -1.$$

- We define $z_0^{(1)} = 1$.
- The linear output is $z_n^{(2)} = a_n^{(2)} = \left\langle \boldsymbol{z}^{(1)} \middle| \boldsymbol{w}_n^{(2)} \right\rangle$ for a suitable $\boldsymbol{w}_n^{(2)} \in \Omega^{(2)} \subset \mathbb{R}^{K+1}$.

Error Back Propagation (EBP)

- An adequate loss function for the problem is defined as:

$$\mathcal{L}\left(\{\boldsymbol{w}_n^{(2)}\}_{n=1}^{o}, \{\boldsymbol{w}_k^{(1)}\}_{k=1}^{K}\right) = \frac{1}{2}\sum_{\ell=1}^{L}\left\langle \boldsymbol{t}_\ell - \boldsymbol{z}^{(2)}(\boldsymbol{x}_\ell) \middle| \boldsymbol{t}_\ell - \boldsymbol{z}^{(2)}(\boldsymbol{x}_\ell)\right\rangle = \sum_{\ell=1}^{L}\mathcal{L}_\ell.$$

- Let us compute the derivatives of the Loss function with respect to the vectors $\boldsymbol{w}^{(2)}$ and $\boldsymbol{w}_k^{(1)}$:

$$\left|\nabla_{\boldsymbol{w}_n^{(2)}}\mathcal{L}_\ell\right\rangle = \frac{\partial \mathcal{L}_\ell}{\partial a_n^{(2)}}\nabla_{\boldsymbol{w}_n^{(2)}} a_n^{(2)}$$

$$\delta_n^{(2)} := \frac{\partial \mathcal{L}_\ell}{\partial a_n^{(2)}} = z_n^{(2)}(\boldsymbol{x}_\ell) - t_{n,\ell}$$

$$\left|\nabla_{\boldsymbol{w}_n^{(2)}} a_n^{(2)}\right\rangle = \left|\nabla_{\boldsymbol{w}_n^{(2)}}\left\langle \boldsymbol{z}^{(1)} \middle| \boldsymbol{w}_n^{(2)}\right\rangle\right\rangle = \left|\boldsymbol{z}^{(1)}\right\rangle.$$

- Thus

$$\left|\nabla_{\boldsymbol{w}_n^{(2)}}\mathcal{L}_\ell\right\rangle = \delta_n^{(2)}\left|\boldsymbol{z}^{(1)}\right\rangle.$$

.

- The derivatives with respect the weights linking hidden units to inputs are:

$$\left|\nabla_{\boldsymbol{w}_k^{(1)}}\mathcal{L}_\ell\right\rangle = \frac{\partial \mathcal{L}_\ell}{\partial a_k^{(1)}}\left|\nabla_{\boldsymbol{w}_k^{(1)}} a_k^{(1)}\right\rangle$$

$$\left|\nabla_{\boldsymbol{w}_k^{(1)}} a_k^{(1)}\right\rangle = \left|\nabla_{\boldsymbol{w}_k^{(1)}}\left\langle \boldsymbol{x}\,\middle|\,\boldsymbol{w}_k^{(1)}\right\rangle\right\rangle = |\boldsymbol{x}\rangle$$

$$\frac{\partial a_n^{(2)}}{\partial a_k^{(1)}} = \frac{\partial}{\partial a_k^{(1)}}\left\langle \boldsymbol{z}^{(1)}\,\middle|\,\boldsymbol{w}_n^{(2)}\right\rangle = w_{n,k}^{(2)}\sigma'\left(a_k^{(1)}\right)$$

$$\delta_k^{(1)} := \frac{\partial \mathcal{L}_\ell}{\partial a_k^{(1)}} = \sum_{n=1}^{o}\frac{\partial \mathcal{L}_\ell}{\partial a_n^{(2)}}\frac{\partial a_n^{(2)}}{\partial a_k^{(1)}} = \sigma'\left(a_k^{(1)}\right)\sum_{n=1}^{o}\delta_n^{(2)} w_{n,k}^{(2)},$$

- Thus

$$\left|\nabla_{\boldsymbol{w}_k^{(1)}}\mathcal{L}_\ell\right\rangle = \sigma'\left(a_k^{(1)}\right)\sum_{n=1}^{o}\delta_n^{(2)} w_{n,k}^{(2)}\,|\boldsymbol{x}\rangle\,.$$

.

Observe that:

$$\delta_j^{(\mathrm{out})} = \frac{\partial \mathcal{L}_\ell}{\partial z_j^{(\mathrm{out})}} \frac{\mathrm{d} z_j^{(\mathrm{out})}}{\mathrm{d} a_j^{(\mathrm{out})}}$$

$$\delta_j^{(m)} = \frac{\mathrm{d} z_j^{(m)}}{\mathrm{d} a_j^{(m)}} \sum_{i=1}^{K^{(m+1)}} \delta_i^{(m+1)} w_{i,j}^{(m+1)}$$

$$\left| \nabla_{\boldsymbol{w}_k^{(m)}} a_k^{(m)} \right\rangle = \left| z^{(m-1)} \right\rangle$$

$$\left| \nabla_{\boldsymbol{w}_k^{(m)}} \mathcal{L}_\ell \right\rangle = \delta_k^{(m)} \left| z^{(m-1)} \right\rangle .$$

- Information is propagated from the input nodes towards the output (all activations $z^{(m)}$ are computed). Forward step.
- Deltas ($\delta_k^{(m)}$) are computed from output to input to complete the derivatives. Backward step.
- Observe that the derivatives of the sigmoidal functions can be expressed in terms of the sigmoidals themselves:

$$\frac{\mathrm{d}\tanh(x)}{\mathrm{d}x} = 1 - \tanh^2(x),$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{2}{1+\mathrm{e}^{-x}} - 1\right) = \frac{2\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2} = \frac{1}{2}\left[1 - \left(\frac{2}{1+\mathrm{e}^{-x}} - 1\right)^2\right].$$

Gradient descent

- The EBP algorithm is completed by applying the gradients to the weight-update algorithm:

$$\left|\boldsymbol{w}_k^{(K)}\right\rangle_{\ell+1} = \left|\boldsymbol{w}_k^{(K)}\right\rangle_{\ell} - \eta_\ell \left|\nabla_{\boldsymbol{w}_k^{(K)}} \mathcal{L}_\ell\right\rangle$$