Machine Learning (CE 40717)

Fall 2024

Ali Sharifi-Zarchi

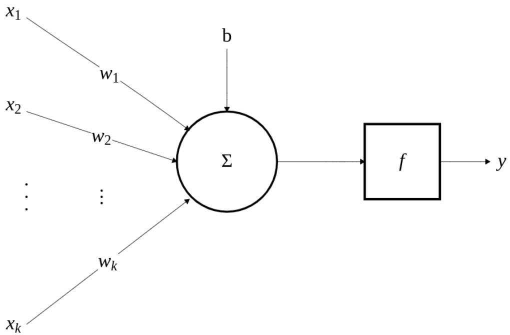
CE Department
Sharif University of Technology

October 9, 2024

Perceptron Reminder

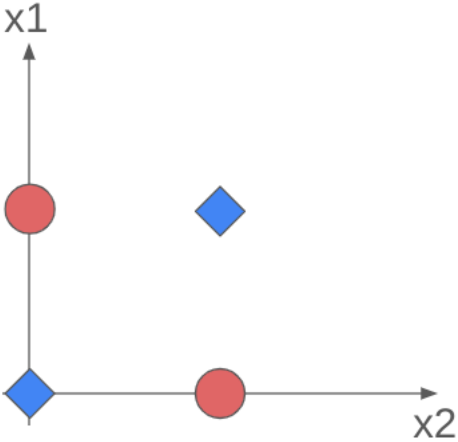
The building block of each neural network is Perceptron:

- $\{x_1, x_2, \ldots, x_k\}$: input features
- $\{w_1, w_2, \ldots, w_k\}$: feature weights
- b : bias term
- $f(\cdot)$: activation function
- y : output of the neuron



Perceptron Capacity

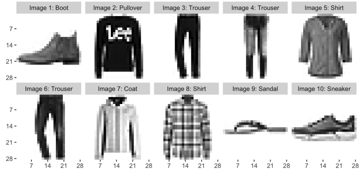
- A perceptron solves linearly separable problems
- The XOR gate is not linearly separable
- We need a multi-layer perceptron for such problems



XOR gate

Why Neural Networks?

- We can find explicit formulas for some problems (no machine learning)
  - $\Delta x = \frac{1}{2} a \cdot t^2 + v_0 \cdot t$
- We can model some problems assuming simple relationships (classical machine learning)
  - House price as a linear function of its features
  - $y = a_1 \cdot x_1 + a_2 \cdot x_2 + \ldots + a_p \cdot x_p$
- How about classifying these images?

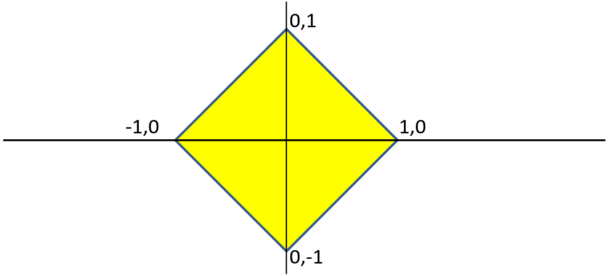


Why Neural Networks? Cont.

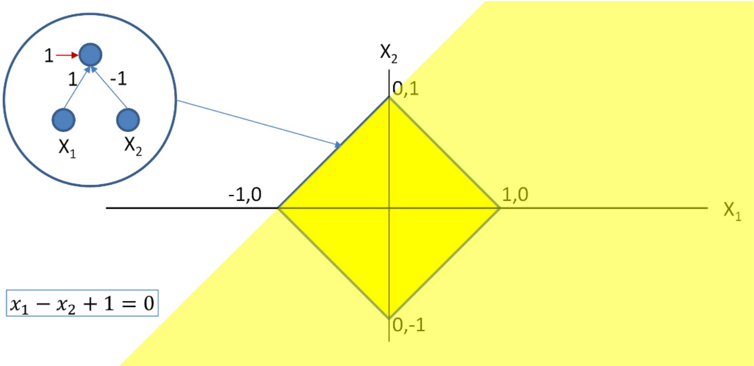
- No explicit formula exists to recognize a sneaker
- We recognize any sneaker intuitively
- Our brains use a complex function for this recognition
- **Deep neural networks** can learn this complex function

Example: MLP for Complex Patterns

- What network to learn this area?
- Example is adapted from [1].

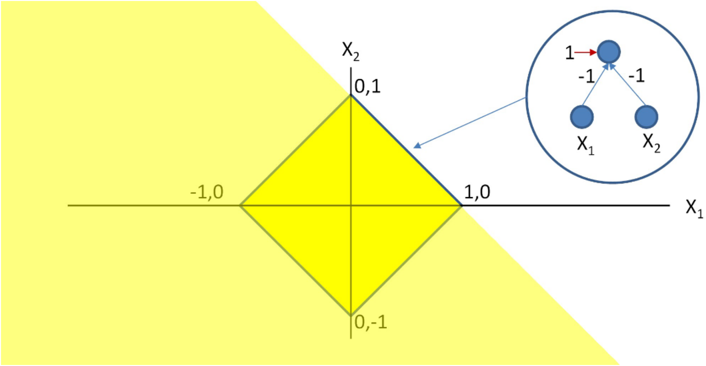


Example: MLP for Complex Patterns Cont.

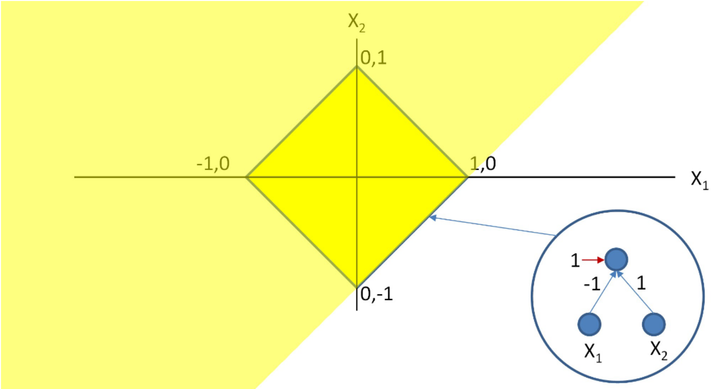


$x_1 - x_2 + 1 = 0$

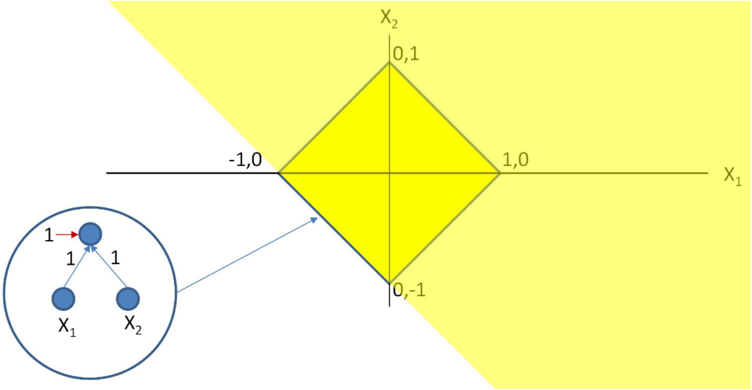
Example: MLP for Complex Patterns Cont.



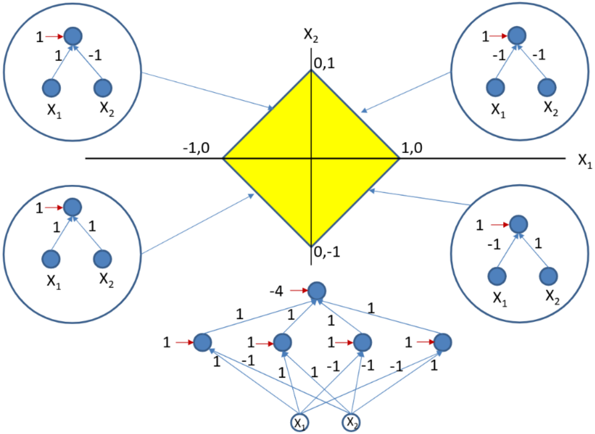
Example: MLP for Complex Patterns Cont.



Example: MLP for Complex Patterns Cont.

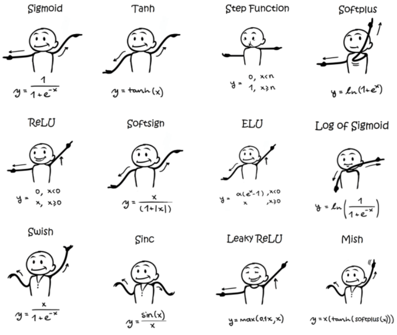


Example: MLP for Complex Patterns Cont.



MLP Capacity

- Increasing width and depth allows us to approximate complex decision boundaries
- An activation function makes a neuron's output non-linear, allowing the network to learn complex data
- Not limited to Boolean or step functions
- With appropriate activation functions, neural networks can approximate any real-valued function (More details later)



Adapted from Sefiks

Single Hidden Layer Neural Network

- Hidden layer pre-activation:

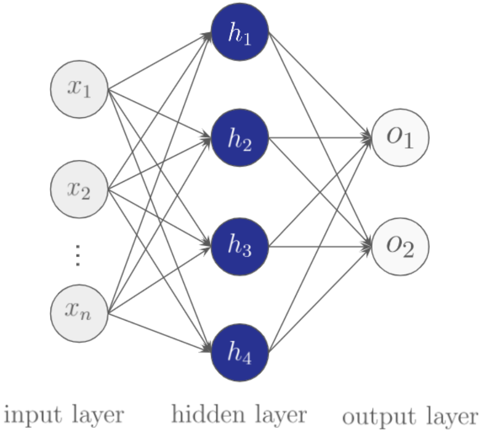
$$a(x)_i = b_i^{(1)} + \sum_j W_{ij}^{(1)} \cdot x_j$$

- Activated hidden layer:

$$h(x) = f(a(x))$$

- Output layer:

$$o(x) = o\left(b^{(2)} + W^{(2)} h^{(1)} x\right)$$



Multi-Hidden Layer Neural Network

- Let $h_i^0 = x_i$ for $i \in \{1, 2, \ldots, n\}$
- For $\ell \in \{0, 1, \ldots, L\}$:

$$a_j^{(\ell+1)} = b_j^{(\ell)} + \sum_i W_{ij}^{(\ell)} \cdot h_i^{(\ell)}$$

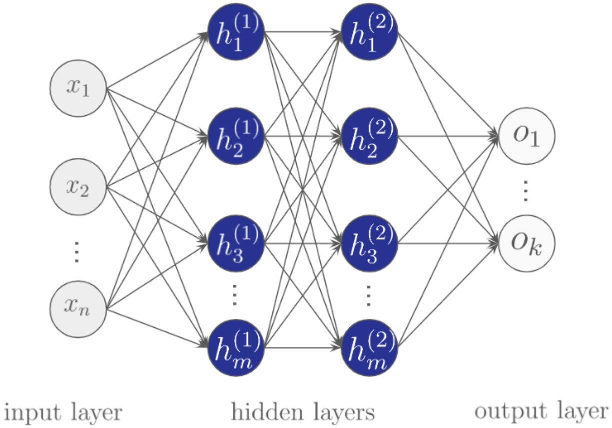
$$h_j^{(\ell+1)} = f(a_j^{(\ell+1)})$$

- Learnable parameters:

$$b_j^{(\ell)}, W_{ij}^{(\ell)}$$

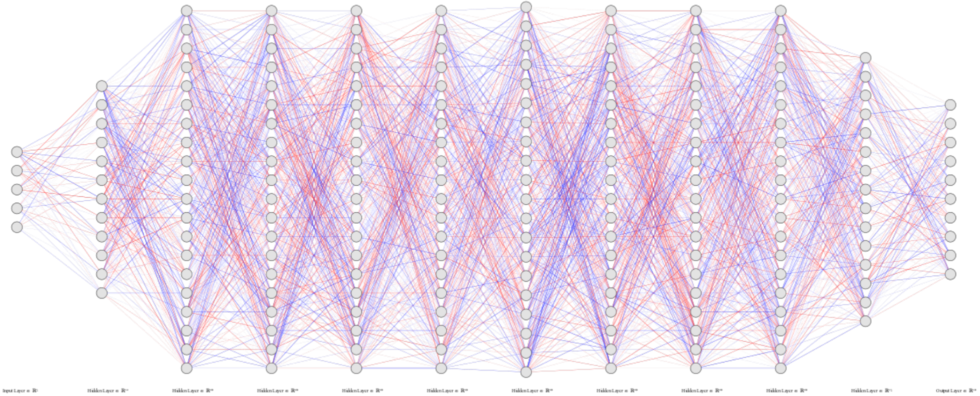
- Number of learnable parameters:

$$(n+1)m_1 + (m_1+1)m_2 + \ldots + (m_L+1)k$$



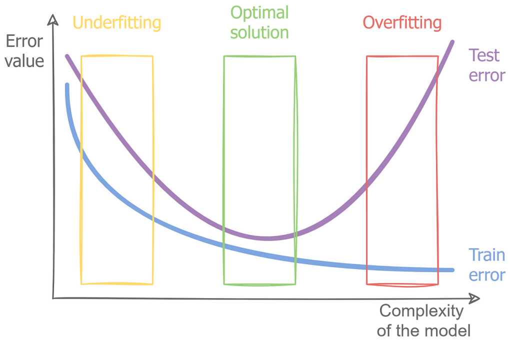
Deep Neural Network Architecture

- More than a few hidden layers: Deep Neural Network (DNN)
- Designing a neural network architecture is more of an art than a science.



Network Width and Depth

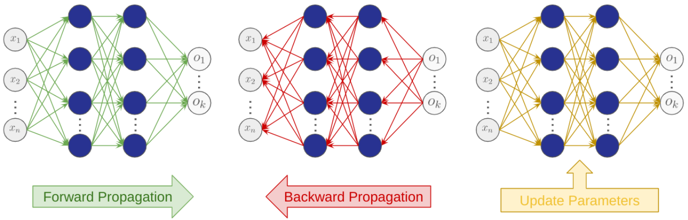
- **Width:** More neurons, more complexity
- **Depth:** More layers, more abstraction
- **Balance:**
  - Too narrow/shallow: risk of underfitting
  - Too wide/deep: risk of overfitting



Adapted from Towards Data Science

Training Phases

- **Initialize weights and biases**: These values control how the network initially processes information (More details later)
- **Forward pass**: Pass the input through the network to get an output
- **Calculate the error**: Compare the network's output to the correct answer to measure the difference (called the 'loss' - More details later)
- **Backpropagation**: Use the loss value to adjust the weights and biases to improve the network's accuracy

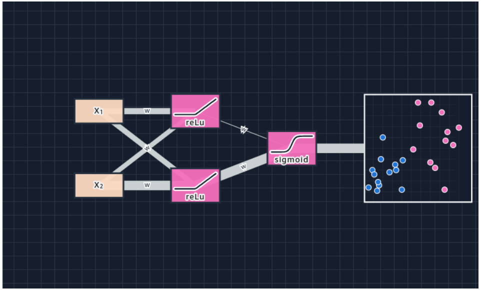


Forward Propagation

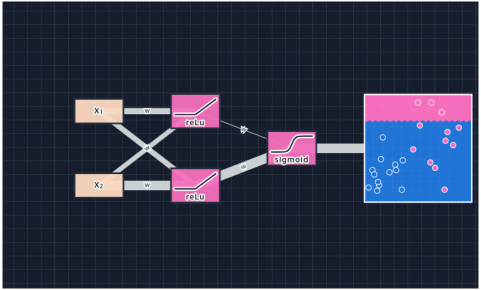
- This is the pass where we send input data through the network to make a prediction (likely inaccurate at first).
- The prediction is made by calculating weighted sums and applying an activation function at each layer

$$o = a^{(L)} = f^{(L)}\left(b^{(L)} + W^{(L)} \cdot f^{(L-1)}\left(\ldots f^{(1)}(b^{(1)} + W^{(1)x}) \ldots\right)\right)$$

Forward Propagation Cont.



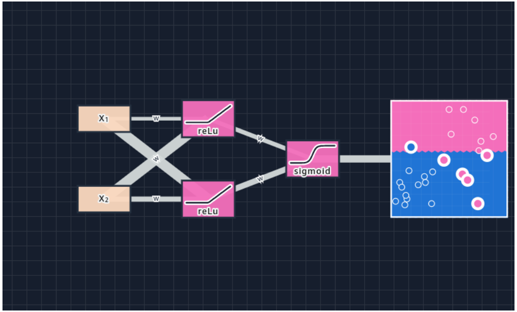
Before producing predictions. Adapted from mlu-explain.



After producing predictions. Adapted from mlu-explain.

Forward Propagation Cont.

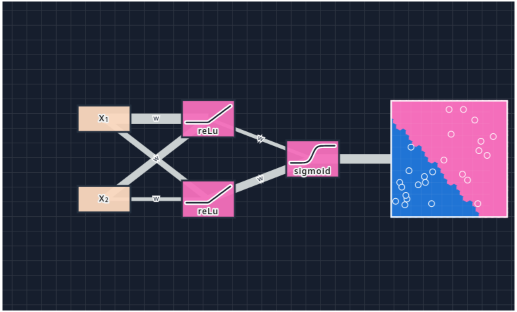
- The goal is to adjust the network's parameters to improve the predictions
- The loss is calculated after the forward pass, telling us how far off our predictions are from the true values



Loss values for predictions. Adapted from mlu-explain.

BackPropagation and Parameter Update

- The network uses the **loss** to adjust its **weights and biases** through a process called **backpropagation**
- Backpropagation calculates how much weights should change to reduce the error
- This will be explained in more detail in the next lecture



Predictions get better as the weights get updated. Adapted from mlu-explain.