


Deep Learning

Chuks Okoli

Last Updated: September 11, 2024

1	Welcome to Deep Learning	2
1.1	Introduction to Deep learning	2
1.1.1	Supervised Learning with Neural Networks	3
1.1.2	Why is Deep Learning taking off?	4
1.2	Neural Network Basics	5
1.2.1	Logistic Regression as a Neural Network	5
1.2.2	Gradient Descent	9
1.3	Shallow Neural Network	12
1.3.1	Neural Network Representation	12
1.3.2	Neural Network Structure	12
1.3.3	Activation Functions	15
1.4	On Cross-Referencing	19
1.5	On Math	19

WELCOME TO DEEP LEARNING



Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI (Artificial Intelligence) will transform in the next several years.

— Andrew Ng, *DeepLearning.AI*

HELLO THERE, and welcome to Deep Learning. This work is a culmination of hours of effort to create my reference for deep learning. All the explanations are in my own words but majority of the contents are based on DeepLearning.AI's specialization in [Deep Learning](#).

1.1 Introduction to Deep learning

The term, Deep Learning, refers to training Neural Networks, sometimes very large Neural Networks. In order to predict the price of a house based on its size for example, we can apply linear regression as a method for fitting a function to predict house prices. An alternative approach would be to use a simple neural network to model the relationship between house size and price.

The simple neural network consists of a single neurons, which takes an input (e.g., house size) and outputs a prediction (e.g., house price). The neuron computes a linear function of the input and applies a rectified linear unit (ReLU) activation function to ensure non-negativity. Each neuron in the network computes a function of the input features and contributes to the overall prediction. When extended to multiple features, each feature is represented by a separate neuron, and the network learns to predict the house price based on these features.

Neural networks are useful in supervised learning scenarios, where the goal is to map input features to corresponding output labels. Given enough training data, neural networks can learn complex mappings from inputs to outputs.

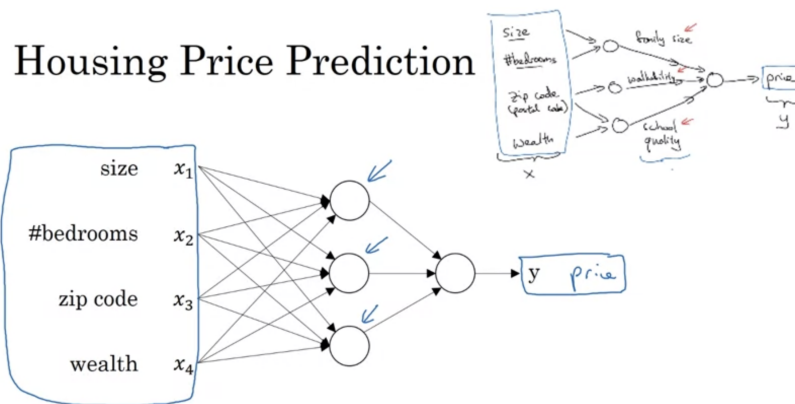


Figure 1.1: A Simple Neural Network for House Price Prediction

1.1.1 Supervised Learning with Neural Networks

Neural networks have gained a lot of attention lately for their ability to solve complex problems effectively. In supervised learning, you input data and aim to predict an output. Examples include predicting house prices or online ad clicks. Neural networks have been successful in various applications, like online advertising, computer vision, speech recognition, and machine translation. Different types of neural networks are used based on the nature of the data, such as convolutional neural networks for images and recurrent neural networks for sequential data. Structured data, like database entries, and unstructured data, like images or text, are both now interpretable by neural networks, thanks to recent advancements. While neural networks are often associated with recognizing images or text, they also excel in processing structured data, leading to improved advertising and recommendation systems. The techniques covered in this course apply to both structured and unstructured data, reflecting the versatility of neural networks in various applications.

Supervised Learning

Input(x) ↙	Output (y) ↙	Application
Home features	Price	Real Estate
Ad, user info ↙	Click on ad? (0/1)	Online Advertising
Image	Object (1,...,1000)	Photo tagging
Audio	Text transcript	Speech recognition
English	Chinese	Machine translation
Image, Radar info ↗	Position of other cars	Autonomous driving

Figure 1.2: Examples of Supervised learning

Neural Network examples

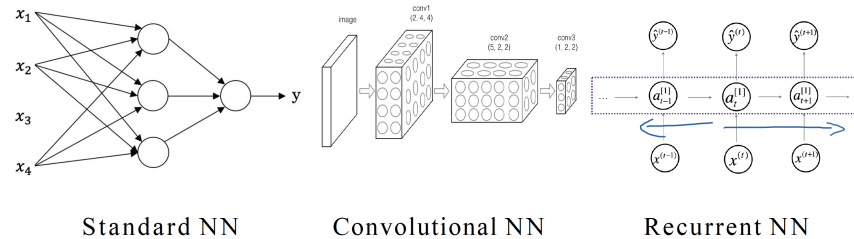


Figure 1.3: Neural network examples

1.1.2 Why is Deep Learning taking off?

The rise of deep learning has been fueled by several key factors. One major driver is the abundance of data available for training machine learning models. With the digitization of society, activities performed on digital devices generate vast amounts of data, enabling neural networks to learn from large datasets. Additionally, advancements in hardware, such as GPUs and specialized processors, have facilitated the training of large neural networks by providing faster computation speeds. Algorithmic innovations, like the adoption of the ReLU activation function, have also played a crucial role in accelerating learning processes. By reducing the time required to train models and enabling faster experimentation, these innovations have enhanced productivity and fostered rapid progress in deep learning research. Moving forward, the continued growth of digital data, advancements in hardware technology, and ongoing algorithmic research are expected to further drive improvements in deep learning capabilities. As a result, deep learning is poised to continue evolving and delivering advancements in various applications for years to come.

FUNFACT: What's drives Deep Learning

Deep Learning took off in the last few years and not before mainly because of great computing power and huge amount of data. These two are the key components for the successes of deep learning. The performance of a neural network improves with more training data.

Scale drives deep learning progress

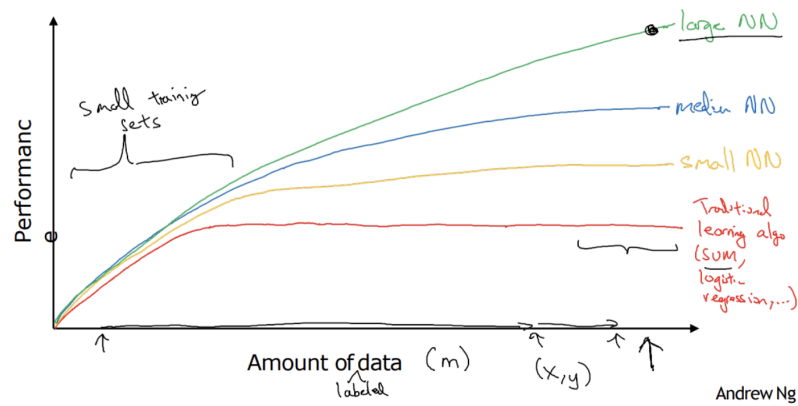


Figure 1.4: Scale drives neural networks

1.2 Neural Network Basics

1.2.1 Logistic Regression as a Neural Network

Logistic regression is an algorithm for binary classification problem. In a binary classification problem, the goal is to train a classifier for which the input is an image represented by a feature vector, x , and predicts whether the corresponding label y is 1 or 0. In this case, whether this is a cat image (1) or a non-cat image (0).



Figure 1.5: Binary classification - Cat vs Non-Cat

An image is stored in the computer in three separate matrices corresponding to the Red, Green, and Blue color channels of the image. The three matrices have the same size as the image, for example, the resolution of the cat image is 64 pixels x 64 pixels, the three matrices (RGB) are 64 x 64 each. The value in a cell represents the pixel intensity which will be used to create a feature vector of n dimension. In pattern recognition and machine learning, a feature vector represents an image. Then the classifier's job is to determine whether it contains a picture of a cat or not. To create a feature vector, x , the pixel intensity values will be “unrolled” or “reshaped” for each color. The dimension of the input feature vector x is $n = 64 * 64 * 3 = 12288$. Hence, we use $n_x = 12288$ to represent the dimensions of the feature vectors.

In binary classification, our goal is to learn a classifier that can input an image represented by this feature vector x and predict whether the corresponding label y is 1 or 0, that is, whether this is a cat image or a non-cat image.

$$x = \begin{bmatrix} 255 \\ 231 \\ 42 \\ \vdots \\ 255 \\ 134 \\ 202 \\ \vdots \\ 255 \\ 134 \\ 93 \\ \vdots \end{bmatrix} \begin{array}{l} \text{red} \\ \text{green} \\ \text{blue} \end{array}$$

Figure 1.6: Reshaped feature vector

Logistic Regression

In Logistic regression, the goal is to minimize the error between the prediction and the training data. Given an image represented by a feature vector x , the algorithm will evaluate the probability of a cat being in that image.

$$\text{Given } x, \hat{y} = P(y = 1|x), \text{ where } 0 \leq \hat{y} \leq 1 \quad (1.1)$$

The parameters used in Logistic regression are:

- The input features vector: $x \in \mathbb{R}^{n_x}$, where n_x is the number of features
- The training label: $y \in \{0, 1\}$
- The weights: $w \in \mathbb{R}^{n_x}$, where n_x is the number of features
- The threshold: $b \in \mathbb{R}$
- The output: $\hat{y} = \sigma(w^T * x + b)$
- Sigmoid function: $s = \sigma(w^T * x + b) = \sigma(z) = \frac{1}{1+e^{-z}}$

$w^T x + b$ is a linear function ($ax + b$), but since we are looking for a probability constraint between $[0, 1]$, the sigmoid function is used. The function is bounded between $[0, 1]$ as shown in the graph above. Some observations from the graph:

1. If z is a large positive number, then $\sigma(z) = 1$
2. If z is small or large negative number, then $\sigma(z) = 0$
3. If $z = 0$, then $\sigma(z) = 0.5$

The difference between the cost function and the loss function for logistic regression is that the loss function computes the error for a single training example while the cost function is the average of the loss functions of the entire training set.

Logistic regression makes use of the sigmoid function which outputs a probability between 0 and 1. The sigmoid function with some weight parameter θ or w and some input $x^{(i)}$ is defined as follows.

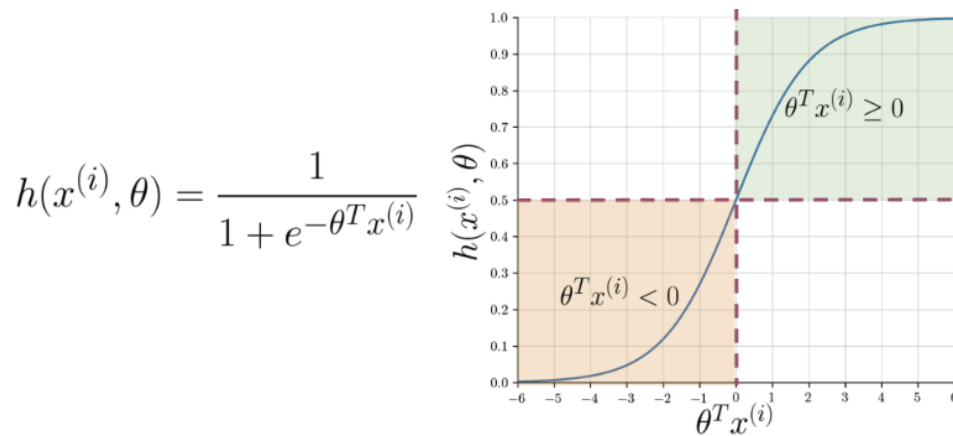


Figure 1.7: Logistic Regression Overview

Note that as $\theta^T x^{(i)}$ i.e. $w^T x$ gets closer and closer to $-\infty$ the denominator of the sigmoid function gets larger and larger and as a result, the sigmoid gets closer to 0. On the other hand, as $\theta^T x^{(i)}$ i.e. $w^T x$ gets closer and closer to ∞ the denominator of the sigmoid function gets closer to 1 and as a result the sigmoid also gets closer to 1.

When we implement logistic regression, our job is to try to learn parameters w and b so that \hat{y} becomes a good estimate of the chance of y being equal to one.

Logistic Regression Cost function

The *loss function* (\mathcal{L}) is a function we need to define to measure how good our output \hat{y} is when the true label is y . The loss function measures how well we are doing on a single training example.

Loss function $\Rightarrow \mathcal{L}(\hat{y}, y) = -y * \log \hat{y} + (1 - y) * \log(1 - \hat{y})$

The *cost function* (\mathcal{J}), which measures how we are doing on the entire training set. So the cost function, which is applied to your parameters w and b , is going to be the average, really one of the m of the sum of the loss function apply to each of the

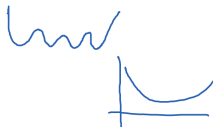
training examples.

Cost function $\Rightarrow \mathcal{J}(w, b) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\hat{y}, y) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} * \log \hat{y}^{(i)} + (1 - y^{(i)}) * \log(1 - \hat{y}^{(i)})]$

Logistic Regression cost function

$\rightarrow \hat{y}^{(i)} = \sigma(w^T x^{(i)} + b)$, where $\sigma(z^{(i)}) = \frac{1}{1+e^{-z^{(i)}}}$ $z^{(i)} = w^T x^{(i)} + b$

Given $\{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$, want $\hat{y}^{(i)} \approx y^{(i)}$. $x^{(i)}$
 $y^{(i)}$
 $z^{(i)}$ i -th example.

Loss (error) function: $\mathcal{L}(\hat{y}, y) = \frac{1}{2} (\hat{y} - y)^2$ 

Cost function: $\mathcal{L}(\hat{y}, y) = - (y \log \hat{y} + (1-y) \log(1-\hat{y})) \leftarrow$

If $y=1$: $\mathcal{L}(\hat{y}, y) = -\log \hat{y} \leftarrow$ Want $\log \hat{y}$ large, want \hat{y} large.

If $y=0$: $\mathcal{L}(\hat{y}, y) = -\log(1-\hat{y}) \leftarrow$ Want $\log(1-\hat{y})$ large ... Want \hat{y} small

Cost function: $\mathcal{J}(w, b) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) \log(1-\hat{y}^{(i)})]$

Figure 1.8: Logistic Regression Cost function

Recap: $\hat{y} = \sigma(w^T x + b)$, $\sigma(z) = \frac{1}{1+e^{-z}} \leftarrow$

$\mathcal{J}(w, b) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})]$

Want to find w, b that minimize $\mathcal{J}(w, b)$

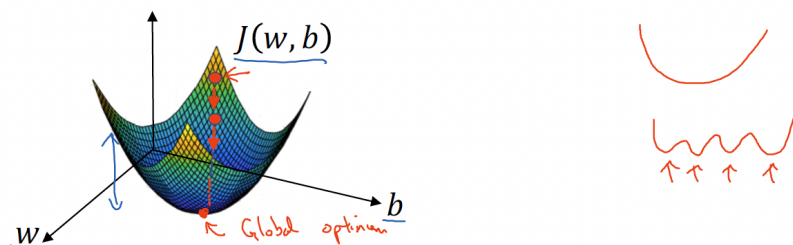


Figure 1.9: Gradient Descent

In logistic regression, you use the cost function $\mathcal{J}(w, b)$ to measure how well your parameters perform on the entire training set. The goal is to minimize $\mathcal{J}(w, b)$ using gradient descent, which iteratively updates the parameters by moving in the direction of steepest descent. Because the cost function is convex, gradient descent will converge to the global minimum, regardless of initialization.

Gradient Descent

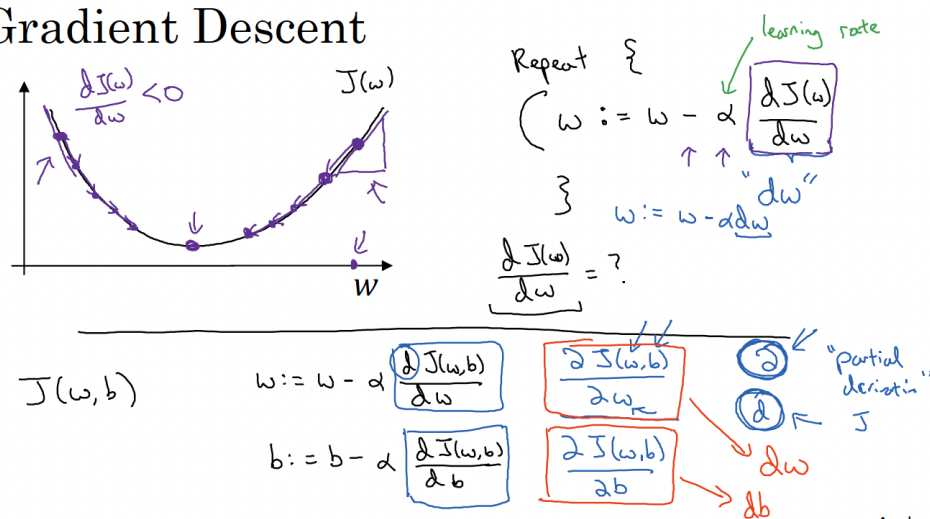


Figure 1.10: Gradient descent Optimization

1.2.2 Gradient Descent

Imagine you're on top of a big hill where the goal is to get to the lowest point. Here's how you would do it:

- Look around you to see which way the ground slopes down the most.
- Take a step in that direction.
- Now that you're in a new spot, look around again to see which way is downhill.
- Take another step in that direction.
- Keep doing this over and over: look for the downhill direction, then take a step.
- Eventually, you'll reach a point where there's no more downhill to go. You're at the bottom!

This is basically what gradient descent does. The "hill" is like a mathematical function we're trying to minimize. "Looking around" is like calculating the gradient (which tells us which direction is downhill). "Taking a step" is like updating our parameters (our position on the hill). We keep doing this until we can't go any lower (we've reached the minimum of the function).

The trick is to not take steps that are too big (or you might overshoot the bottom) or too small (or it will take forever to get there). In the algorithm, we control this with something called the "learning rate". That's gradient descent! It's a way of finding the lowest point by always moving downhill, little by little.

Gradient descent is an optimization technique used to minimize a function, often applied in machine learning for model training. The algorithm starts with an initial guess for the parameters and iteratively updates them by moving in the direction of the negative gradient (the direction of steepest descent) of the function, until it reaches a local minimum. The step size, or learning rate, controls how big each move is.

The gradient descent update rule is:

$$\theta_{\text{new}} = \theta_{\text{old}} - \alpha \nabla J(\theta)$$

where:

- θ are the parameters we want to optimize.
- α is the learning rate (step size).
- $\nabla J(\theta)$ is the gradient of the cost function $J(\theta)$ with respect to θ .

This rule ensures that we move in the direction of the steepest descent, reducing the value of $J(\theta)$ with each iteration.

Why It Works

The gradient of a function points in the direction of the steepest ascent. By moving in the opposite direction of the gradient, the function value decreases, eventually reaching a local or global minimum. As the gradient approaches zero, the parameter values converge toward the minimum.

Pseudocode

Below is the pseudocode for gradient descent in simple terms:

```
1 initialize  $\theta$  (e.g., random values)
2 choose learning rate  $\alpha$ 
3 repeat until convergence:
4     compute gradient  $\nabla J(\theta)$ 
5     update  $\theta : \theta = \theta - \alpha * \nabla J(\theta)$ 
```

The process repeats until the gradient is close to zero, indicating the function has been minimized.

Gradient descent is an optimization algorithm used to minimize the cost function of a model by iteratively adjusting the model's parameters. The goal is to find the values of the parameters (like weights in a machine learning model) that minimize the cost function, which measures how well the model fits the data.

ALGORITHM 1.1: Gradient Descent

Input: Initial parameter θ^0 , gradient of loss function $\nabla J(\theta)$, learning rate α , tolerance ϵ , maximum iterations max_iter

Output: Optimized parameter θ

$iter \leftarrow 0$

while $iter < max_iter$ **do**

 Compute $\nabla J(\theta)$

if $\|\nabla J(\theta)\| < \epsilon$ **then**

break

end if

$\theta \leftarrow \theta - \alpha \nabla J(\theta)$

$iter \leftarrow iter + 1$

end while

return θ

How Gradient Descent Works:

1. **Initialize Parameters:** Start with an initial guess for the parameters (e.g., weights). This can be a set of random values or zeros.
2. **Calculate the Gradient:** Compute the gradient (i.e., the partial derivative) of the cost function with respect to each parameter. The gradient indicates the direction of the steepest increase in the cost function.
3. **Update the Parameters:** Adjust the parameters in the opposite direction of the gradient (i.e., the direction that reduces the cost function). The size of the step taken is controlled by a learning rate, a hyperparameter that determines how big the steps are.

$$\text{new parameter} = \text{current parameter} - \text{learning rate} \times \text{gradient}$$

4. **Repeat:** Continue recalculating the gradient and updating the parameters until the algorithm converges, meaning the changes in the cost function become very small, indicating that a minimum has been reached.

FUNFACT: Computation Graph

The computation graph organizes a computation from left-to-right computation. Through a left-to-right pass, we can compute the value of \mathcal{J} . In order to compute derivatives there'll be a right-to-left pass, kind of going in

the opposite direction called backwards propagation. That would be most natural for computing the derivatives.

Forward Propagation: Computes the output y from the input X by passing through the network layers.

Backward propagation adjusts the network's parameters (weights and biases) by propagating the error backward from the output to the input, using the gradients to minimize the loss.

1.3 Shallow Neural Network

1.3.1 Neural Network Representation

A neural network is a machine learning model inspired by the human brain. It consists of layers of interconnected nodes (neurons) where each connection has a weight, representing its importance. Neural networks process input data by passing it through these layers, applying activation functions to transform the data, and adjusting the weights based on the output error during training. The goal is to minimize the error and improve the model's predictions. Neural networks are widely used in tasks like image recognition, language processing, and complex data modeling.

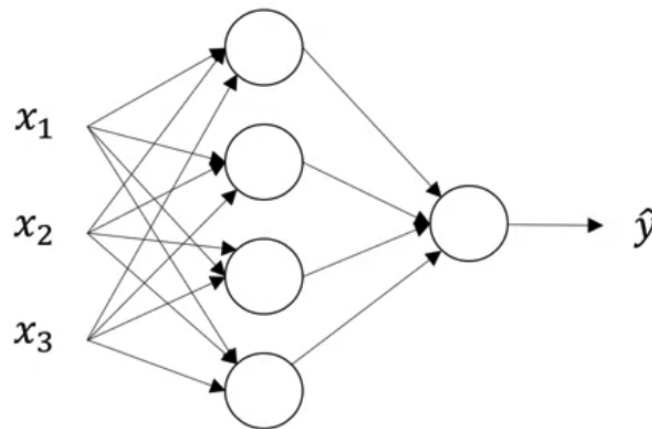


Figure 1.11: Shallow Neural Network

1.3.2 Neural Network Structure

A simple neural network has similar structure as a linear classifier:

- A neuron takes inputs from other neurons (-> input into linear classifier)
- The inputs are summed in a weighted manner (-> weighted sum)

DEEP LEARNING

- ▢ Learning is through a modification of the weights (gradient descent in the case of NN)
- If it receives enough inputs, it “fires” (if it exceeds the threshold or weighted sum plus bias is high enough)
- The output of a neuron can be modulated by a non linear function (e.g sigmoid).

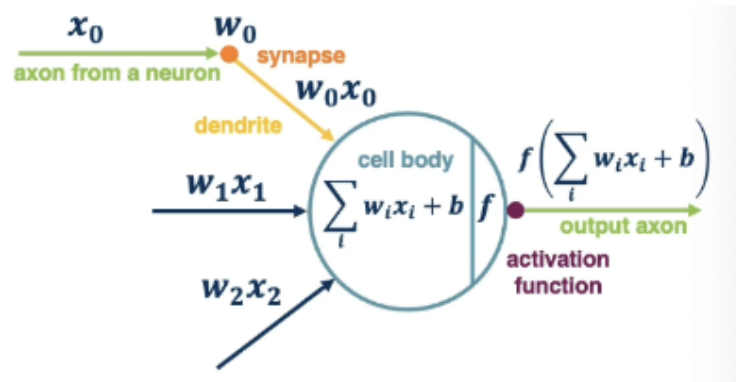


Figure 1.12: Structure of a simple neural network

A neural network consists of three primary layers: input, hidden, and output layers as shown in **Fig. 1.13**.

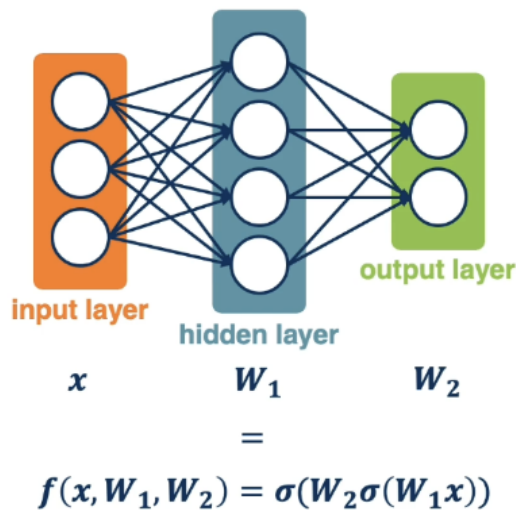


Figure 1.13: Layers in a neural network

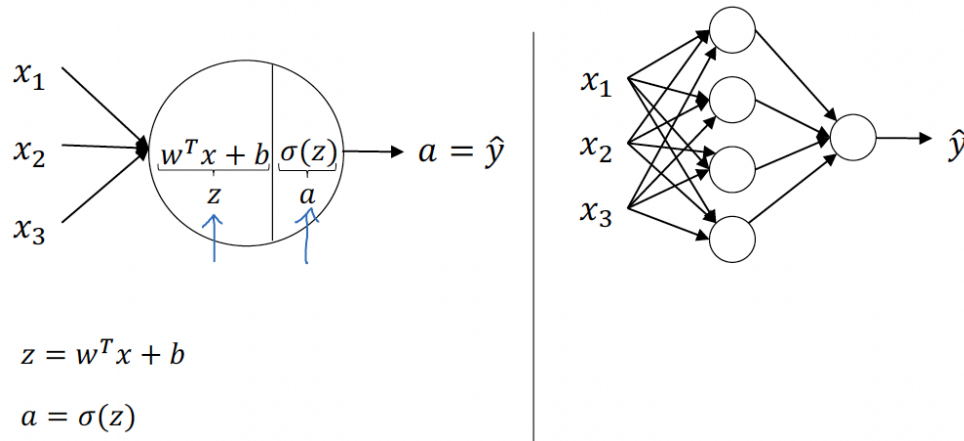


Figure 1.14: Neural Network Representation

Input Layer

The **input layer** is responsible for taking in the features of the data. For example, features x_1, x_2, x_3 are passed into the network in **Fig. 1.14**, and these values are referred to as the *activations* of the input layer, denoted $A^{[0]}$.

Hidden Layer

The **hidden layer** processes the input features. It is called *hidden* because, during training, the true values for these nodes are not seen in the training data. The activations of this layer are denoted $A^{[1]}$, where each node $A_i^{[1]}$ represents an activation value computed using a weight matrix $W^{[1]}$ and bias vector $b^{[1]}$. For example, if there are four hidden units, $A^{[1]}$ will be a 4-dimensional vector.

Output Layer

The **output layer** produces the final prediction \hat{y} , based on the activations from the hidden layer. The output is represented as $A^{[2]}$, a single scalar value in this example.

Notation

We use the following notations to represent the activations in different layers:

- $A^{[0]}$: Activations of the input layer.
- $A^{[1]}$: Activations of the hidden layer.
- $A^{[2]}$: Output, representing \hat{y} .

In the neural network in **Fig. 1.14**, each layer has associated weight matrices and biases:

- $W^{[1]}$ is a 4×3 matrix (4 hidden units, 3 input features).

- $b^{[1]}$ is a 4×1 vector (for 4 hidden units).
- $W^{[2]}$ is a 1×4 matrix (1 output unit, 4 hidden units).
- $b^{[2]}$ is a 1×1 scalar (for the output unit).

While this neural network has three layers (input, hidden, output), it is commonly referred to as a **two-layer network** because the input layer is not counted. Therefore, the hidden layer is called *layer 1* and the output layer *layer 2*.

Training

The parameters W and b are optimized during training to minimize the error between the predicted output \hat{y} and the actual output y . In this process, the neural network learns the best weights and biases for making accurate predictions.

1.3.3 Activation Functions

Activation functions are mathematical equations that determine the output of a neural network node. They introduce non-linearity into the model, enabling the network to learn complex patterns. Below are some common activation functions used in neural networks:

1. Sigmoid Function

The **sigmoid** activation function is defined as:

$$a = \sigma(x) = \frac{1}{1 + e^{-z}}$$

It compresses input values to a range between 0 and 1, making it useful for models that need probabilities as output or binary classification. However, it suffers from the *vanishing gradient problem*, where gradients become very small, slowing down learning.

2. Tanh Function

The **tanh** activation function is defined as:

$$a = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

It outputs values between -1 and 1, centered around zero, making learning more efficient in some cases. Like sigmoid, it also suffers from vanishing gradients but it is superior to the sigmoid function for hidden layers.

3. ReLU (Rectified Linear Unit)

The **ReLU** activation function is defined as:

$$a = \text{ReLU}(z) = \max(0, z)$$

ReLU is simple and efficient, especially in deep networks, because it allows faster convergence. The downside is that it can cause “dead neurons” (neurons that output 0 for all inputs), which may stop learning in certain neurons.

4. Leaky ReLU

The **Leaky ReLU** activation function is defined as:

$$a = \max(0.01z, z)$$

The Leaky ReLU addresses the ReLU’s zero-gradient issue by allowing small negative values. Generally works better than ReLU but is used less frequently.

5. Softmax Function

The **softmax** function is commonly used in the output layer for classification tasks. It converts raw outputs (logits) into probabilities:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

This is useful for multi-class classification, ensuring the sum of output probabilities equals 1.

Choosing an Activation Function

ReLU is popular for hidden layers in deep neural networks due to its simplicity and efficiency. **Sigmoid** and **tanh** are useful in smaller networks or specific scenarios but less common in modern deep networks. **Softmax** is used in the output layer for multi-class classification tasks.

- **Output Layer:** Use **Sigmoid** for binary classification.
- **Hidden layers:** **ReLU** is the default choice, though **tanh** can also be used effectively.
- **Learning Efficiency:** **ReLU** and **Leaky ReLU** often result in faster learning compared to sigmoid or tanh, as their gradients do not saturate easily.

Each activation function has its specific use cases depending on the task and network architecture.

Pros and cons of activation functions

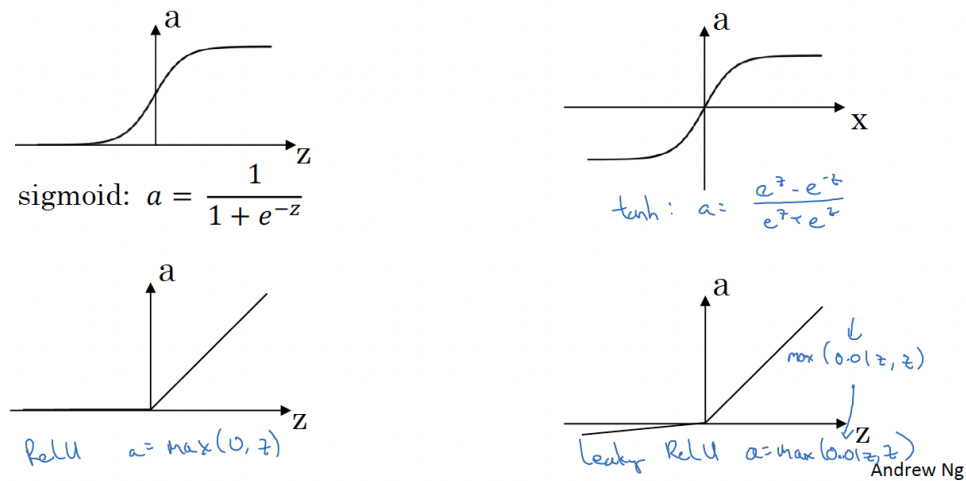


Figure 1.15: Activation Functions in Neural Networks

- Like this one,

which is wrapped in gray. I use it for notes...

- Or this one,

which is wrapped in red. I use it for fun facts or other asides...

- Or this one,

which is wrapped in blue and used for mathy stuff.

- Or this last one,

which is wrapped in green. With a title, it's used for enumerated examples (see `\extitle` and `\excounter`). Observe:

EXAMPLE 1.1: Test

This is an example. What's the answer to $2 + 2$?

ANSWER: Obviously 4, lol.

EXAMPLE 1.2: Test Again

This one will increment the counter automatically, resetting for each chapter.

- For red and blue boxes, there are custom commands for titles, too:

ONE TITLE

Like this

TWO TITLES: A Subtitle

Or this

These styles also automatically apply to theorems and claims.

Theorem 1.1 (Pythagorean Theorem). *For any right triangle with legs a, b and hypotenuse c :*

$$a^2 + b^2 = c^2 \tag{1.2}$$

Proof. This is left as an exercise to the reader. ■

Claim 1.1. *This is the greatest note template in the world.*

There are different ways to quote things, too, depending on how you want to emphasize:

This is a simple, indented quote with small letters and italics usually suitable for in-text quotations when you just want a block.

Alternatively, you can use the `\inspiration` command from the chapter heading, which leverages the `thickleftborder` frame internally, but adds a little more padding and styling (there's also just `leftborder` for a thinner variant):

■ Hello there!

1.4 On Cross-Referencing

You can reference most things—see [Theorem 1.1](#) or [\(1.2\)](#) or the [Welcome to Deep Learning](#) chapter—directly and easily as long as you give them labels. These are “built-ins.” However, you can also create a **custom term** that will be included in the index, then include references to it that link back to the original definition. Try clicking: [custom term](#). Building the index is on you, though. You can also reference by using a different term for the text: [like this](#). Sometimes it doesn’t fit the **grammatical structure** of the sentence so you can define the term one way and visualize it another way (this creates a **grammar** entry in the index). There’s also **math terms** and a way to reference them: [math terms](#) (clickable), but they do **not** show up in the index.

This is the standard way to include margin notes. There are also commands to link to source papers directly (see `\lesson`).

1.5 On Math

Most of the math stuff is just macros for specific things like the convolution operator, \otimes , probabilities, $\Pr[A|B=C]$, or big- O notation, $\mathcal{O}(n^2 \log n)$ but there’s also a convenient way to include explanations on the side of an equation:

$$\begin{array}{ll}
 1 + 1 \stackrel{?}{=} 2 & \text{first we do this} \\
 2 \stackrel{?}{=} 2 & \text{then we do this} \\
 2 = 2 & \blacksquare
 \end{array}$$

These are all in the `CustomCommands.sty` file.