Table of Contents

[References 3](#_Toc179103080)

[7th Sep 4](#_Toc179103081)

[8th Sep 4](#_Toc179103082)

[Neural Network Architecture 4](#_Toc179103083)

[14th Sep 4](#_Toc179103084)

[Azure Account 4](#_Toc179103085)

[Service Models 4](#_Toc179103086)

[Infrastructure as a Service (IaaS) 4](#_Toc179103087)

[Platform as a Service (PaaS) 5](#_Toc179103088)

[Software as a Service (SaaS) 5](#_Toc179103089)

[15th Sep 5](#_Toc179103090)

[Sigmoid function 5](#_Toc179103091)

[Alternatives 6](#_Toc179103092)

[21st Sep 6](#_Toc179103093)

[22nd Sep 6](#_Toc179103094)

[Maximum (Singular) 6](#_Toc179103095)

[Maxima (Plural) 6](#_Toc179103096)

[Global Minima 6](#_Toc179103097)

[Local Minima 6](#_Toc179103098)

[28th September 6](#_Toc179103099)

[29th September 6](#_Toc179103100)

[BGD - Batch Gradient Descent 6](#_Toc179103101)

[SGD - Stochastic Gradient Descent 8](#_Toc179103102)

[BGD vs SGD 9](#_Toc179103103)

[BGD - Batch Gradient Descent 9](#_Toc179103104)

[5th October 10](#_Toc179103105)

[Perceptron or Single Layer Perceptron 10](#_Toc179103106)

# References

GitHub Link

* [shikharkumar13 (Data Science With Shikhar) (github.com)](https://github.com/shikharkumar13)

Drive:

* [Data Science Files - Google Drive](https://drive.google.com/drive/folders/1jiKxJCI3Cnd_pgznxQ-690O8RyfmgiPV)
* <https://drive.google.com/drive/folders/16YcAr_oKbrtNjJkMhTSXVpldvy9m-p15>

# 7th Sep

In this class just given introduction about the Deep Learning.

# 8th Sep

## Neural Network Architecture

Neural network architecture refers to the structure of a neural network, which is a map of the neural layers and processes. It defines how information flows through the network and how the neurons are interconnected. The architecture determines the network's capacity to learn and solve specific problems.

Here's a breakdown of the key components of a neural network architecture:

* **Input Layer:** This layer consists of neurons that receive input data. Each neuron in this layer represents a feature of the input data.
* **Hidden Layers:** These are intermediate layers that process the inputs received from the input layer. A neural network can have one or more hidden layers, and the number of neurons in each hidden layer can vary. The hidden layers apply weights and biases to the inputs and pass the results through an activation function to produce outputs.
* **Output Layer:** The final layer of the network, which produces the output. The number of neurons in this layer typically corresponds to the number of classes in a classification problem or the number of output variables in a regression problem.

# 14th Sep

## Azure Account

Today, the session is to create the Azure account.

Created Azure account by using below:

URL: <https://portal.azure.com>

UserName: [jagadeeshmushm2@gmail.com](mailto:jagadeeshmushm2@gmail.com)

Password: W\*\*c\*\*\*@\*\*8\*

## Service Models

Service models are the different ways in which cloud services are delivered to users, typically defined in terms of the level of control and management responsibility they offer. The three main cloud service models are:

### Infrastructure as a Service (IaaS)

* Provides virtualized computing resources over the internet, such as virtual machines, storage, and networks.
* Users have control over the infrastructure (like storage, servers, and networking) but must manage the operating systems, applications, and runtime environments.
* **Example providers**: Amazon Web Services (AWS), Microsoft Azure, Google Cloud.

### Platform as a Service (PaaS)

* Offers a platform that allows developers to build, run, and manage applications without worrying about the underlying infrastructure.
* The provider manages infrastructure, operating systems, and runtime environments, while the user focuses on deploying and managing their applications.
* **Example providers**: Google App Engine, Heroku, Microsoft Azure App Service.

### Software as a Service (SaaS)

* Delivers software applications over the internet, fully managed by the service provider.
* Users access the software without worrying about the underlying infrastructure, maintenance, or updates.
* **Example providers**: Google Workspace, Microsoft 365, Salesforce.

These models cater to different levels of responsibility and control, allowing businesses to choose the most suitable option for their needs.

# 15th Sep

Deep Learning basics explanation

* Pixels -> Edges -> Patterns -> Digits

Note:

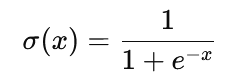
* In this class, focus on Weights and Bias

## Sigmoid function

The **Sigmoid function** is a widely used activation function in deep learning, particularly for binary classification tasks and in the output layer of neural networks.

**Formula:**

The Sigmoid function is mathematically defined as:



Where:

* x is the input (the weighted sum of the neuron's inputs).
* e is Euler's number (approximately 2.718).

**Pros:**

* **Smooth gradient**: The smoothness helps avoid sudden jumps in predictions.
* **Probabilistic interpretation**: Output can be interpreted as probabilities.

**Cons:**

* **Vanishing gradient problem**: For large positive or negative inputs, the gradient approaches zero, leading to slow learning during backpropagation (especially in deep networks).
* **Outputs not centered at zero**: This can slow down convergence, as the output of the Sigmoid function is always positive.

### Alternatives

In modern deep learning, alternatives like the **ReLU** (Rectified Linear Unit) or **Leaky ReLU** are often preferred for hidden layers, as they mitigate some of the limitations of the Sigmoid function, especially the vanishing gradient problem. However, the Sigmoid is still used in the output layer for binary classification tasks.

# 21st Sep

Deep Learning basic explanation.

# 22nd Sep

## Maximum (Singular)

* Refers to **a single largest value** in a given set or function.
* Example: If a function has one highest point, that point is called the maximum.
* In optimization, the maximum is the greatest value a function can achieve at a particular point

## Maxima (Plural)

* Refers to **multiple maximum points** in a given set or function.
* A function may have multiple maxima, especially in non-monotonic functions (functions that increase and decrease at different intervals).

## Global Minima

* The **global minimum** is the lowest value of the function across its entire domain.
* It represents the absolute minimum point, meaning no other point in the function has a smaller value than the global minimum.
* A function can have one or more global minima, but they all must have the same value.

## Local Minima

* A **local minimum** is the lowest value of the function in a specific neighborhood (or interval) but **not necessarily the lowest value** across the entire domain.
* Local minima occur at points where the function changes direction from decreasing to increasing, but there could be lower points elsewhere in the function.
* A function can have multiple local minima.

# 28th September

Hands session on Gradient Descent Demonstration

# 29th September

## BGD - Batch Gradient Descent

**BGD** is one of the most common optimization algorithms used in machine learning to minimize a cost (or loss) function by updating model parameters (like weights) in the direction of the steepest descent (i.e., negative gradient).

**How BGD Works:**

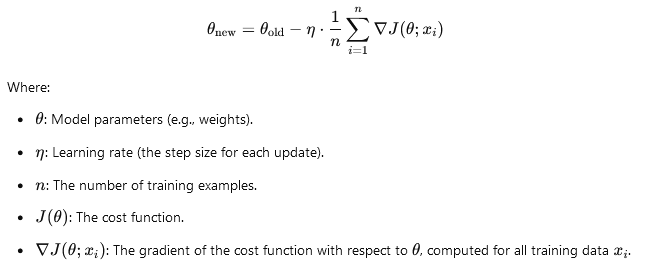
In **Batch Gradient Descent**, the gradient of the cost function is computed using the **entire dataset** at once. The algorithm then updates the model parameters in one step based on this cumulative gradient.

**Steps of Batch Gradient Descent:**

1. **Calculate the gradient** of the cost function for all the training data. This means you look at how much the model's error changes as you slightly adjust the model's parameters (weights and biases).
2. **Update the model parameters** by moving them in the opposite direction of the gradient. This process is repeated iteratively until the algorithm converges to the minimum of the cost function.

**Update Rule:**

The update rule for the model parameters θ\thetaθ in Batch Gradient Descent is:



**Key Features of BGD:**

1. **Uses the entire dataset**: At each iteration, BGD calculates the gradient based on all data points, ensuring that each step is a true representation of the entire dataset.
2. **Stable convergence**: Since it considers all the data at once, the updates are typically more stable, and it ensures a smooth path toward the minimum of the cost function.
3. **Global minima**: Batch Gradient Descent is less prone to being stuck in local minima (for convex problems) compared to Stochastic Gradient Descent, which has more noise due to the single-sample updates.

**Advantages:**

* **Accurate gradient estimation**: Since the gradient is computed using the entire dataset, the updates are precise, leading to more stable convergence.
* **Suitable for convex problems**: For convex functions, BGD is guaranteed to converge to the global minimum.

**Disadvantages:**

* **Slow for large datasets**: Since BGD requires calculating the gradient over the entire dataset before updating the parameters, it can be slow and computationally expensive, especially for large datasets.
* **Memory-intensive**: It needs to load the entire dataset into memory at once, which may not be feasible for very large datasets.
* **Slower updates**: The model parameters are updated less frequently compared to Stochastic Gradient Descent (SGD), which updates parameters for each individual sample.

**Example Use:**

In smaller datasets, where the entire dataset can fit into memory and the computational load is not too heavy, BGD is a practical and effective optimization algorithm. However, for larger datasets, more efficient alternatives like **Stochastic Gradient Descent (SGD)** or **Mini-batch Gradient Descent** are preferred.

## SGD - Stochastic Gradient Descent

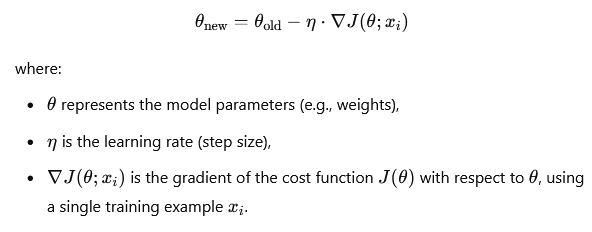
**SGD** is an optimization algorithm widely used in machine learning and deep learning to minimize the cost (loss) function and improve model accuracy.

**How SGD Works:**

* The core idea of SGD is similar to **Gradient Descent**, where the algorithm iteratively updates the model's parameters (like weights) by following the gradient of the loss function in the direction that minimizes it.
* However, instead of using the entire dataset to compute the gradient (as in **Batch Gradient Descent**), SGD uses only a **single sample** (or a small mini-batch) from the dataset to perform the update at each iteration.

**Key Features of SGD:**

1. **Stochastic vs Batch Gradient Descent**:
   * In **Batch Gradient Descent**, the gradient is computed using the entire dataset, which can be slow and computationally expensive, especially for large datasets.
   * In **Stochastic Gradient Descent**, the gradient is computed using only a single data point (or a mini-batch), which makes the updates faster but also introduces some noise into the optimization process.
2. **Update Rule**: The general parameter update rule in SGD is:



1. **Advantages of SGD**:
   * **Faster updates**: Since it updates the parameters more frequently (once per sample), it can converge faster than batch gradient descent.
   * **Better for large datasets**: Works well for large datasets where loading the entire data into memory is impractical.
   * **Escapes local minima**: Due to the noise introduced by using individual data points, it is more likely to escape local minima and saddle points compared to batch gradient descent.
2. **Disadvantages of SGD**:
   * **Noisy updates**: Since SGD relies on single data points, the gradient can fluctuate a lot, making the convergence process noisy and less stable.
   * **May require more iterations**: Although the updates are frequent, it may take more iterations to converge due to the noise introduced by single-sample updates.
   * **Tuning learning rate**: Choosing the right learning rate is critical for achieving convergence. If the learning rate is too high, the algorithm may overshoot the minimum.

**Improvements to SGD:**

Several variants of SGD have been developed to address its weaknesses:

* **Mini-batch Gradient Descent**: Combines the benefits of both batch and stochastic gradient descent by using small batches (e.g., 32 or 64 samples) instead of a single sample or the entire dataset.
* **Momentum**: Adds a velocity term to smooth out the noisy updates and speed up convergence.
* **Adam (Adaptive Moment Estimation)**: A popular optimization algorithm that adapts the learning rate for each parameter and combines the benefits of both momentum and RMSprop.

**Example Use:**

SGD is often used in training large-scale models, including deep learning models, where full-batch gradient descent would be too slow or computationally impractical.

## BGD vs SGD

* **Batch Gradient Descent** computes the gradient based on the entire dataset in each iteration.
* **Stochastic Gradient Descent** computes the gradient based on a single data point, leading to more frequent updates but noisier convergence.
* **Mini-batch Gradient Descent** is a compromise between the two, updating parameters based on a small random subset (mini-batch) of the dataset.

**In Summary:**

* **BGD** is accurate and stable but slow for large datasets, whereas **SGD** is faster but less stable.
* **BGD** is used when you can afford to process the entire dataset in each iteration and need stable convergence.

## BGD - Batch Gradient Descent

**BGD** stands for **Batch Gradient Descent**, which is one of the most common optimization algorithms used in machine learning to minimize a cost (or loss) function by updating model parameters (like weights) in the direction of the steepest descent (i.e., negative gradient).

**How BGD Works:**

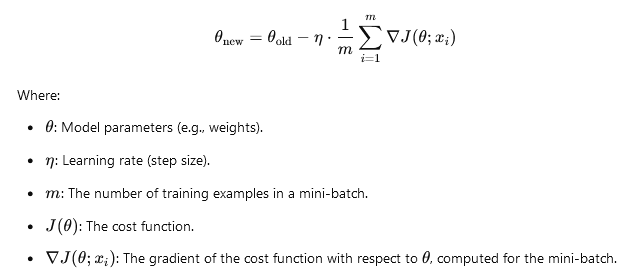
In **Batch Gradient Descent**, the gradient of the cost function is computed using the **entire dataset** at once. The algorithm then updates the model parameters in one step based on this cumulative gradient.

**Steps of Batch Gradient Descent:**

1. **Calculate the gradient** of the cost function for all the training data. This means you look at how much the model's error changes as you slightly adjust the model's parameters (weights and biases).
2. **Update the model parameters** by moving them in the opposite direction of the gradient. This process is repeated iteratively until the algorithm converges to the minimum of the cost function.

**Update Rule:**

The update rule for the model parameters θ\thetaθ in Batch Gradient Descent is:



**Key Features of BGD:**

1. **Uses the entire dataset**: At each iteration, BGD calculates the gradient based on all data points, ensuring that each step is a true representation of the entire dataset.
2. **Stable convergence**: Since it considers all the data at once, the updates are typically more stable, and it ensures a smooth path toward the minimum of the cost function.
3. **Global minima**: Batch Gradient Descent is less prone to being stuck in local minima (for convex problems) compared to Stochastic Gradient Descent, which has more noise due to the single-sample updates.

**Advantages:**

* **Accurate gradient estimation**: Since the gradient is computed using the entire dataset, the updates are precise, leading to more stable convergence.
* **Suitable for convex problems**: For convex functions, BGD is guaranteed to converge to the global minimum.

**Disadvantages:**

* **Slow for large datasets**: Since BGD requires calculating the gradient over the entire dataset before updating the parameters, it can be slow and computationally expensive, especially for large datasets.
* **Memory-intensive**: It needs to load the entire dataset into memory at once, which may not be feasible for very large datasets.
* **Slower updates**: The model parameters are updated less frequently compared to Stochastic Gradient Descent (SGD), which updates parameters for each individual sample.

**Example Use:**

In smaller datasets, where the entire dataset can fit into memory and the computational load is not too heavy, BGD is a practical and effective optimization algorithm. However, for larger datasets, more efficient alternatives like **Stochastic Gradient Descent (SGD)** or **Mini-batch Gradient Descent** are preferred.

# 5th October

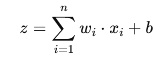
## Perceptron or Single Layer Perceptron

A **perceptron** is a type of artificial neuron used in machine learning, specifically in the context of supervised learning. It is the simplest form of a neural network and serves as a fundamental building block for more complex models like **multi-layer perceptrons (MLPs)** or **deep neural networks**.

**Components of a Perceptron:**

A perceptron takes multiple input features, applies weights to them, computes a weighted sum, and passes the result through an activation function to produce an output (usually a binary decision). Here’s how it works step-by-step:

1. **Input features** x1,x2,...,xn ​:
   * These are the features of the data. Each feature corresponds to one input for the perceptron.
2. **Weights** w1,w2,..., wn​:
   * Each input is associated with a weight, which determines the importance of that input in making the prediction.
3. **Bias** b:
   * The bias term allows the perceptron to shift the decision boundary, enabling it to make predictions even when all input values are zero.
4. **Weighted Sum**:
   * The perceptron calculates the weighted sum of the inputs:

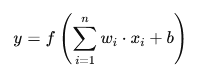


1. **Activation Function**: The weighted sum zzz is passed through an activation function, which produces the final output. The classic perceptron uses a **step function** as the activation function:

****

* + The output yyy is binary (either 0 or 1), making the perceptron a binary classifier.

**Perceptron Model:**

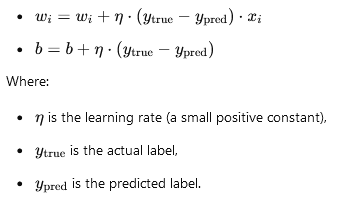


Where f is the activation function, often the step function in basic perceptrons.

### Perceptron Learning Rule

The perceptron adjusts its weights based on the errors it makes during the learning process. This is done using the **Perceptron Learning Algorithm**:

1. For each training example, the perceptron computes the output.
2. If the output is correct, no changes are made.
3. If the output is incorrect, the weights and bias are updated as follows:



### Example of Perceptron

Consider a perceptron used to classify whether a student will pass or fail based on two input features: study hours (x1x\_1x1​) and sleep hours (x2x\_2x2​):

* Inputs: x1=study hours, x2=sleep hoursx
* Weights: w1=0.6,w2=0.4
* Bias: b=−0.8
* Step function: Outputs 1 if the weighted sum z≥0, otherwise outputs 0.

The perceptron calculates:

* z=(0.6×study hours)+(0.4×sleep hours)−0.8

Based on the result, the perceptron predicts whether the student will pass (output 1) or fail (output 0).

### Limitations of the Perceptron

1. **Linearly Separable Data**: A perceptron can only classify data that is **linearly separable**. This means the data must be separable by a straight line (in 2D) or a hyperplane (in higher dimensions). It cannot solve problems where the decision boundary is nonlinear (e.g., XOR problem).
2. **No Hidden Layers**: The basic perceptron only has one layer of computation (the input layer and output). More complex tasks require multiple layers of neurons, which leads to **multi-layer perceptrons (MLPs)** and **deep learning models**.

# 6th September

## Multi-Layer Perception

A **Multilayer Perceptron (MLP)** is a class of feedforward artificial neural network that consists of multiple layers of neurons, including one or more hidden layers between the input and output layers. It is a foundational model in **deep learning** and is capable of solving both **linear** and **non-linear problems**, unlike the basic perceptron, which can only handle linearly separable problems.

### Structure of a Multilayer Perceptron (MLP)

1. **Input Layer**:
   * The first layer of the MLP that receives the input data. Each neuron in this layer represents one input feature.
2. **Hidden Layer(s)**:
   * One or more intermediate layers of neurons between the input and output layers. These layers introduce complexity, allowing the MLP to learn and represent non-linear relationships in the data.
   * Each hidden layer is fully connected to the next layer, meaning each neuron in one layer is connected to every neuron in the next layer.
3. **Output Layer**:
   * The final layer of the MLP that produces the model’s prediction. For classification tasks, the output is typically transformed using a **softmax** (for multi-class classification) or **sigmoid** function (for binary classification). For regression tasks, the output is typically a linear function.

### Activation Functions in MLP

The neurons in hidden layers use **non-linear activation functions** to introduce non-linearity, enabling the MLP to solve complex problems. Common activation functions include:

* **Sigmoid**



* **ReLU (Rectified Linear Unit)**



* **Tanh**

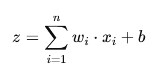


The activation function is applied to the weighted sum of the inputs for each neuron, which allows the network to model complex, non-linear relationships in the data.

### Forward Propagation in MLP

In an MLP, data flows from the input layer, through the hidden layer(s), to the output layer in a process known as **forward propagation**. The steps involved are:

1. **Weighted Sum**: Each neuron calculates the weighted sum of the inputs:



1. Where wi are the weights, xi are the input values, and b is the bias.
2. **Activation Function**: The weighted sum zzz is passed through an activation function (such as ReLU or Sigmoid), which produces the neuron’s output.
3. **Propagation to Next Layer**: The output from each neuron in the current layer becomes the input to neurons in the next layer.

This process continues until the output layer produces a final prediction.

### Training a Multilayer Perceptron

MLPs are trained using a process called **backpropagation**, combined with an optimization algorithm like **Gradient Descent** or **Stochastic Gradient Descent (SGD)**. Training involves the following key steps:

1. **Initialization**: The weights and biases are initialized, usually with small random values.
2. **Forward Propagation**: The input data is passed through the network layer by layer, and the final output is calculated.
3. **Loss Function**: A loss (or cost) function is used to measure the difference between the predicted output and the true output. Common loss functions are:
   * **Mean Squared Error (MSE)**: For regression tasks.
   * **Cross-Entropy Loss**: For classification tasks.
4. **Backward Propagation (Backpropagation)**:
   * The error (difference between the actual and predicted outputs) is propagated backward through the network.
   * The gradients of the loss function with respect to the weights are computed using the **chain rule** of calculus.
5. **Weight Update**: The weights are updated using an optimization algorithm like gradient descent, which adjusts them to minimize the loss function:



Where:

* w are the weights,
* η is the learning rate,
* ∇J(w) is the gradient of the loss function with respect to the weights.

1. Repeat: Forward propagation and backpropagation are repeated for multiple epochs (passes over the dataset) until the model converges and the loss is minimized.

### Key Features of MLP

* **Fully Connected**: Every neuron in one layer is connected to every neuron in the next layer.
* **Multiple Hidden Layers**: The presence of multiple hidden layers allows MLPs to model complex patterns and non-linear relationships in data.
* **Non-linear Activation Functions**: These introduce the ability to solve non-linear problems, which cannot be handled by a simple perceptron.

### Advantages of MLP

1. **Capability of Learning Non-linear Patterns**: Unlike the basic perceptron, MLPs can solve complex problems with non-linear decision boundaries.
2. **Versatile**: MLPs can be applied to various machine learning tasks, including classification, regression, and even time-series prediction.
3. **Expressive Power**: MLPs can approximate any continuous function given enough hidden units, making them universal approximators.

### Disadvantages of MLP

1. **Requires Tuning**: Choosing the right number of hidden layers, neurons, and hyperparameters like the learning rate can be difficult and time-consuming.
2. **Computationally Intensive**: Training MLPs can be slow, especially on large datasets, due to the full connection of layers and the need for iterative updates.
3. **Sensitive to Hyperparameters**: MLP performance depends on choosing appropriate learning rates, activation functions, and optimizers, and can be prone to overfitting.

### Applications of MLP

* **Image Recognition**: Early image recognition models, before convolutional neural networks (CNNs), used MLPs to classify images.
* **Natural Language Processing**: MLPs can be used for text classification, sentiment analysis, and other tasks.
* **Regression Tasks**: MLPs are widely used in regression problems to predict continuous values.

### Example of MLP

Consider a simple MLP for classifying whether an email is spam or not (binary classification). The MLP could have:

* **Input Layer**: Features like email length, number of links, certain keywords, etc.
* **Hidden Layer**: One or more layers with neurons that learn complex relationships between the features.
* **Output Layer**: A single neuron with a sigmoid activation function that outputs a probability between 0 and 1, where values close to 1 indicate "spam" and values close to 0 indicate "not spam."

### Summary

A **Multilayer Perceptron (MLP)** is a type of neural network with multiple layers that can solve both linear and non-linear problems. It is trained using forward and backward propagation, and it plays a key role in machine learning and deep learning models. While powerful, it requires careful tuning of hyperparameters and computational resources for training.

## Neuron and Perceptron

The terms **neuron** and **perceptron** are related but have distinct meanings in the context of artificial neural networks (ANNs):

**1. Neuron (Artificial Neuron):**

* **Definition**: In the context of artificial neural networks, a **neuron** is a basic computational unit that takes multiple inputs, processes them, and produces an output. It is inspired by the biological neurons in the human brain.
* **Components**:
  + **Inputs**: Neurons receive inputs (which can be features from data or outputs from other neurons).
  + **Weights**: Each input is multiplied by a corresponding weight to signify its importance.
  + **Bias**: A bias term is added to the weighted sum to shift the decision boundary.
  + **Activation Function**: The weighted sum is passed through an activation function (such as sigmoid, ReLU, or tanh) to introduce non-linearity.
* **Output**: The neuron produces an output that can either be passed to another neuron (in a hidden layer or output layer) or be the final prediction.
* **Role**: A neuron can be part of a single-layer or multi-layer neural network. It's a generic term that applies to any node in a neural network, whether in the input, hidden, or output layer.
* **Activation Function**: Neurons typically use non-linear activation functions like sigmoid, ReLU, or tanh to learn complex relationships.

**2. Perceptron:**

* **Definition**: A **perceptron** is a specific type of artificial neuron used in machine learning, originally introduced by Frank Rosenblatt in 1958. It is a simple linear classifier that forms the foundation of modern neural networks. The perceptron is the simplest kind of neural network model and can be used for binary classification.
* **Components**: Similar to an artificial neuron, the perceptron takes multiple inputs, weights them, adds a bias, and computes a weighted sum.
* **Activation Function**: Unlike neurons in modern neural networks, the perceptron uses a **step function** (or threshold function) as its activation function. The step function outputs a binary result:
  + y=1 if the weighted sum z≥0
  + y=0 if z < 0
* **Role**: The perceptron is primarily a **linear classifier**, meaning it can only classify data that is linearly separable. If the data can be divided by a straight line (or hyperplane in higher dimensions), a perceptron will correctly classify the points. However, it fails when the data is non-linearly separable, like the XOR problem.

**Comparison of Neuron vs. Perceptron:**

| **Feature** | **Neuron** | **Perceptron** |
| --- | --- | --- |
| **General Definition** | A computational unit in a neural network, part of a larger network architecture. | A type of artificial neuron designed to classify binary data. |
| **Role** | Can be part of input, hidden, or output layers in deep learning models. | Primarily used for linear classification tasks in early neural network models. |
| **Activation Function** | Uses modern, non-linear functions like ReLU, sigmoid, or tanh. | Uses a step function, producing binary outputs (0 or 1). |
| **Complexity** | Used in complex, multi-layer architectures to learn non-linear patterns. | Simple, single-layer linear classifier (one output layer). |
| **Learning Capacity** | Can handle non-linear separable problems (with multiple layers and non-linear activations). | Can only handle linearly separable problems. |
| **Historical Significance** | A building block in modern deep learning (e.g., in MLPs, CNNs, RNNs). | The original neuron model in machine learning, foundational for later advancements. |
| **Limitation** | Can solve complex problems using backpropagation and multi-layer networks. | Limited to simple, linearly separable problems. |

**Summary:**

* A **neuron** is a general concept used in artificial neural networks, often part of complex architectures like multi-layer perceptrons or deep neural networks, and can solve non-linear problems using various activation functions.
* A **perceptron** is a specific, simpler version of an artificial neuron that only performs linear classification and uses a step function as its activation function.

While perceptrons are important for understanding the basics of neural networks, modern neural networks typically involve neurons with more advanced functionality, allowing them to solve complex, non-linear problems.