Smarter education for smarter planet
IBM Innovation Centre for Education

Welcome to:

Neural networks, Natural language understanding

Unit objectives

After completing this unit, you should be able to:

- Understand the basics of neural network
- Gain knowledge on the applications of neural networks
- Learn about the different types of neural network based on applications
- Gain an insight into the role of neural network in natural language processing

Introduction

- Neural network learning methods provide a robust approach to approximating real-valued, discrete-valued, and vector-valued target functions.

- For certain types of problems, such as learning to interpret complex real-world sensor data, artificial neural networks are among the most effective learning methods currently known.

- For example, the Back-propagation algorithm described in this module has proven surprisingly successful in many practical problems such as learning to recognize handwritten characters, learning to recognize spoken words and learning to recognize faces.

- The study of artificial neural networks (ANNs) has been inspired in part by the observation that biological learning systems are built of very complex webs of interconnected neurons.

- In rough analogy, artificial neural networks are built out of a densely interconnected set of simple units, where each unit takes a number of real-valued inputs (possibly the outputs of other units) and produces a single real-valued output (which may become the input to many other units).

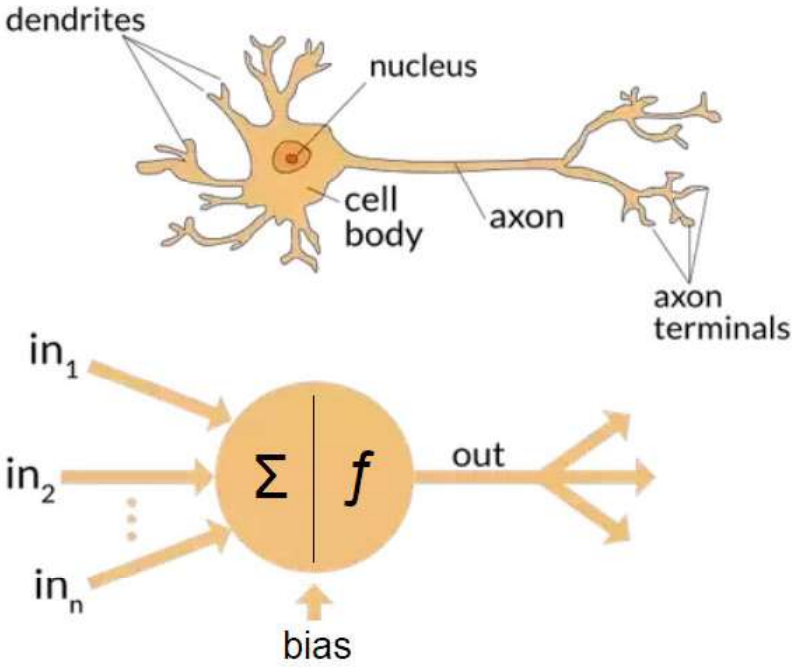
Artificial Neural Network

- It is a computational model inspired by the way biological neural networks in the human brain process information

- It has a lot of excitement and breakthrough in
  - Machine Learning research and industry
  - Speech recognition
  - Computer vision
  - Text processing

- Used for regression and classification problems

- It combines hundreds of algorithms and their variants



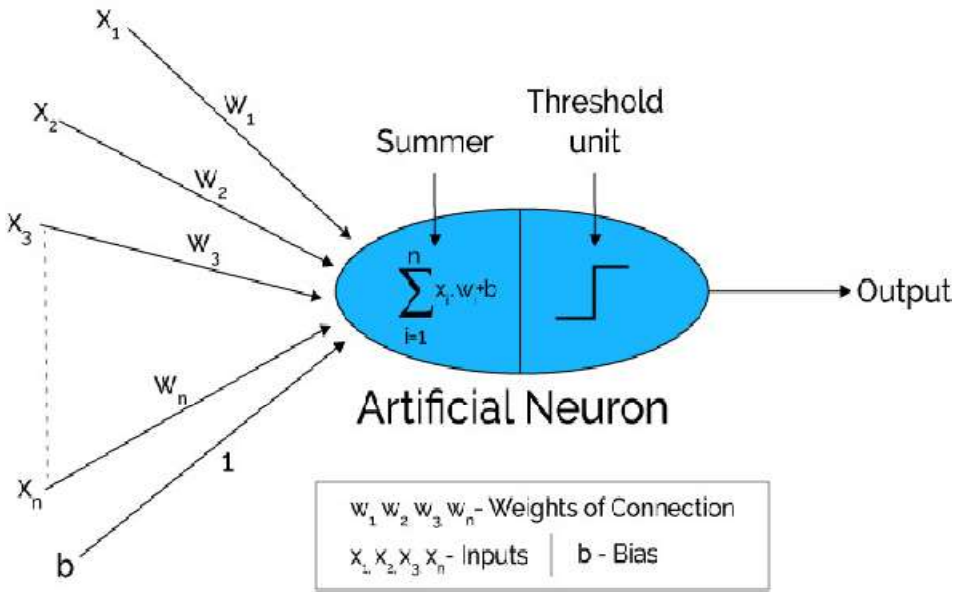
Appropriate problems for neural network learning

- ANN learning is well-suited to problems in which the training data corresponds to noisy, complex sensor data, such as inputs from cameras and microphones.

- It is also applicable to problems for which more symbolic representations are often used, such as the decision tree learning tasks.

- In these cases, ANN and decision tree learning often produce results of comparable accuracy.

- The back-propagation algorithm is the most commonly used ANN learning technique.



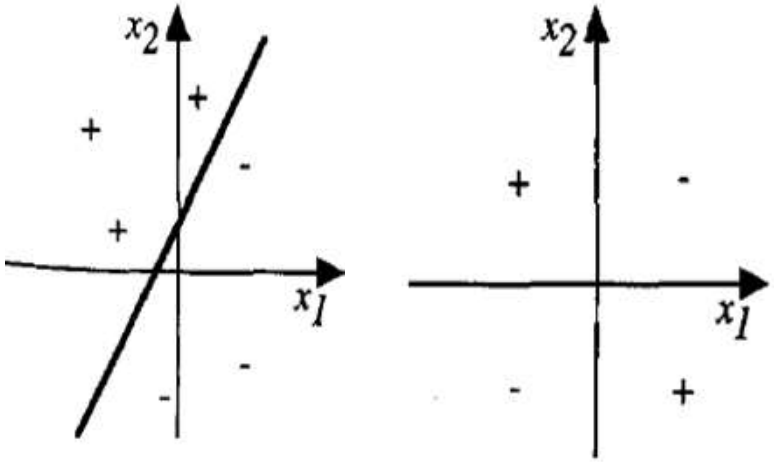
Characteristics of the problems

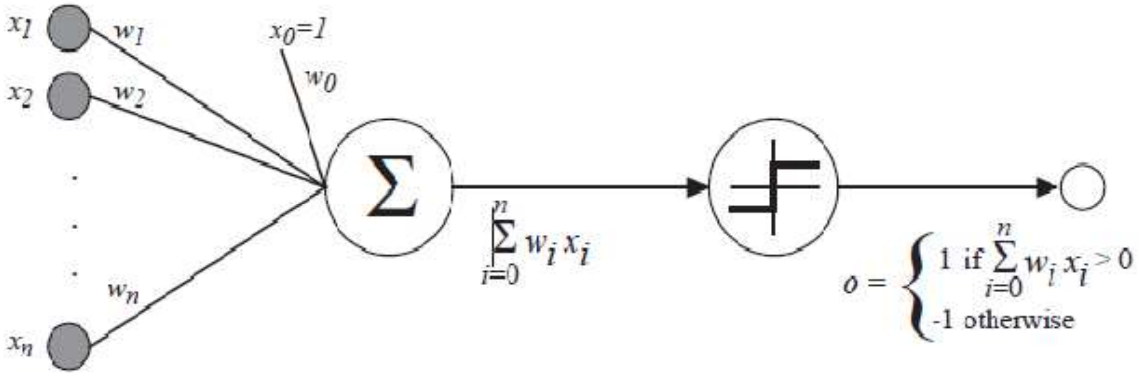
- Instances are represented by many attribute-value pairs.
- The target function output may be discrete-valued, real-valued, or a vector of several real- or discrete-valued attributes.
- The training examples may contain errors.
- Long training times are acceptable
- Fast evaluation of the learned target function may be required.
- The ability of humans to understand the learned target function is not important.

Basic understanding of neural networks

- Perceptron is the basic operational unit of artificial neural networks. It employs supervised learning rule and is able to classify the data into two classes.

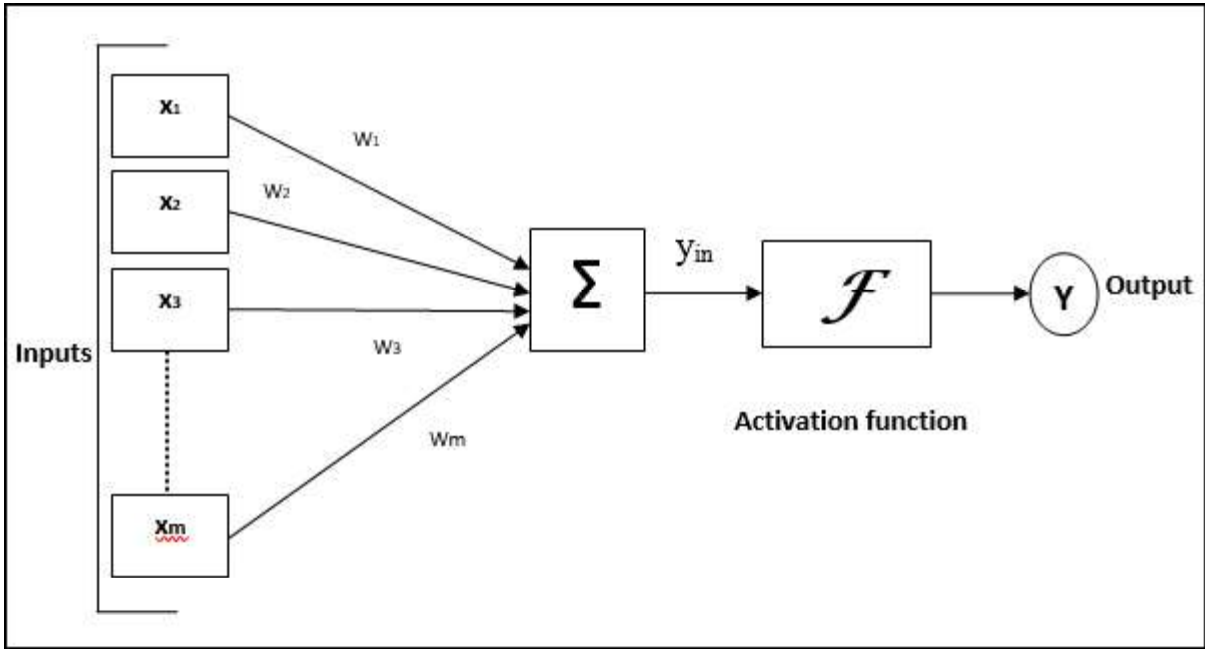
- We can view the perceptron as representing a hyperplane decision surface in the n-dimensional space of instances (i.e., points). The perceptron outputs a 1 for instances lying on one side of the hyperplane and outputs a -1 for instances lying on the other side, as illustrated in Figure given below.





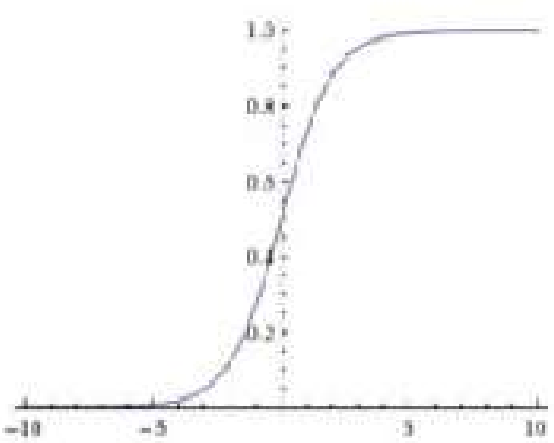
A single neuron

- The basic unit of computation in a neural network is the neuron, often called a node or unit
- Receives input from some other nodes, or from an external source and computes an output
- Each input has an associated weight (w) which is assigned on the basis of its relative importance to other inputs
- The node applies a function f to the weighted sum of its inputs as shown below

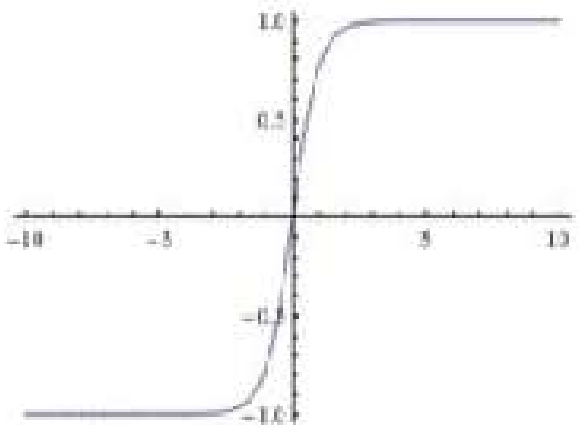


Activation functions

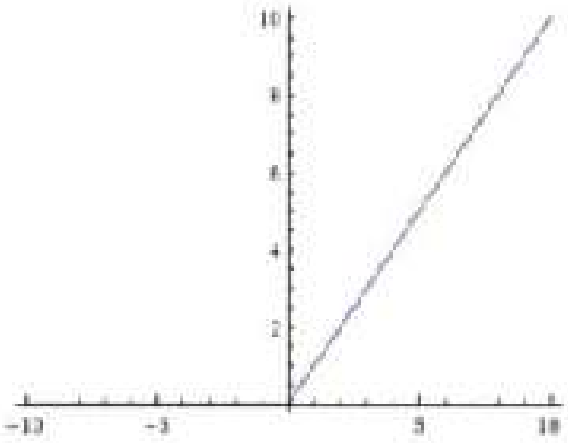
- Sigmoid: Takes a real-valued input and squashes it to range between [0,1]
- tanh: Takes a real-valued input and squashes it to the range [-1, 1]
- ReLU: It stands for Rectified Linear Unit
- It takes a real-valued input and thresholds it at zero (replaces negative values with zero)



Sigmoid



tanh



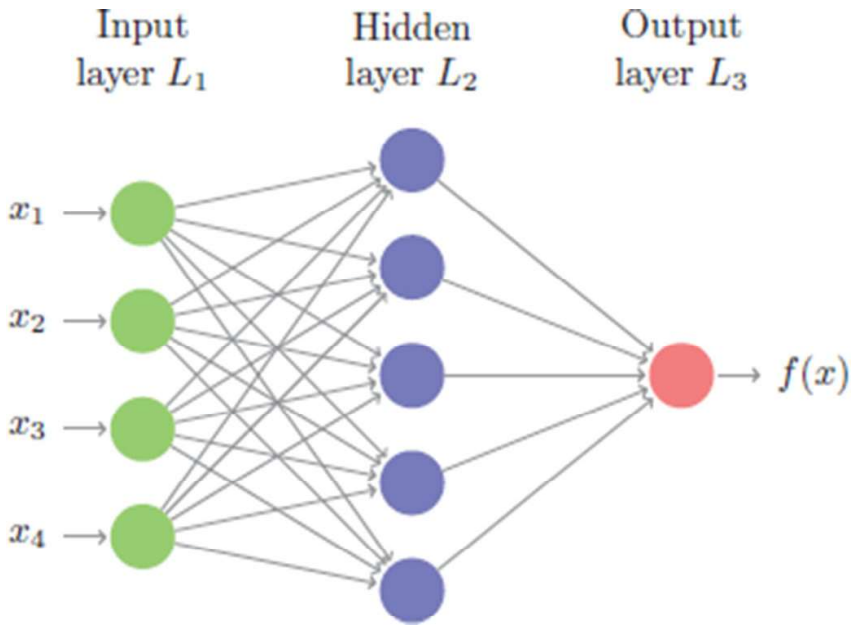
ReLU

Architectures of neural networks

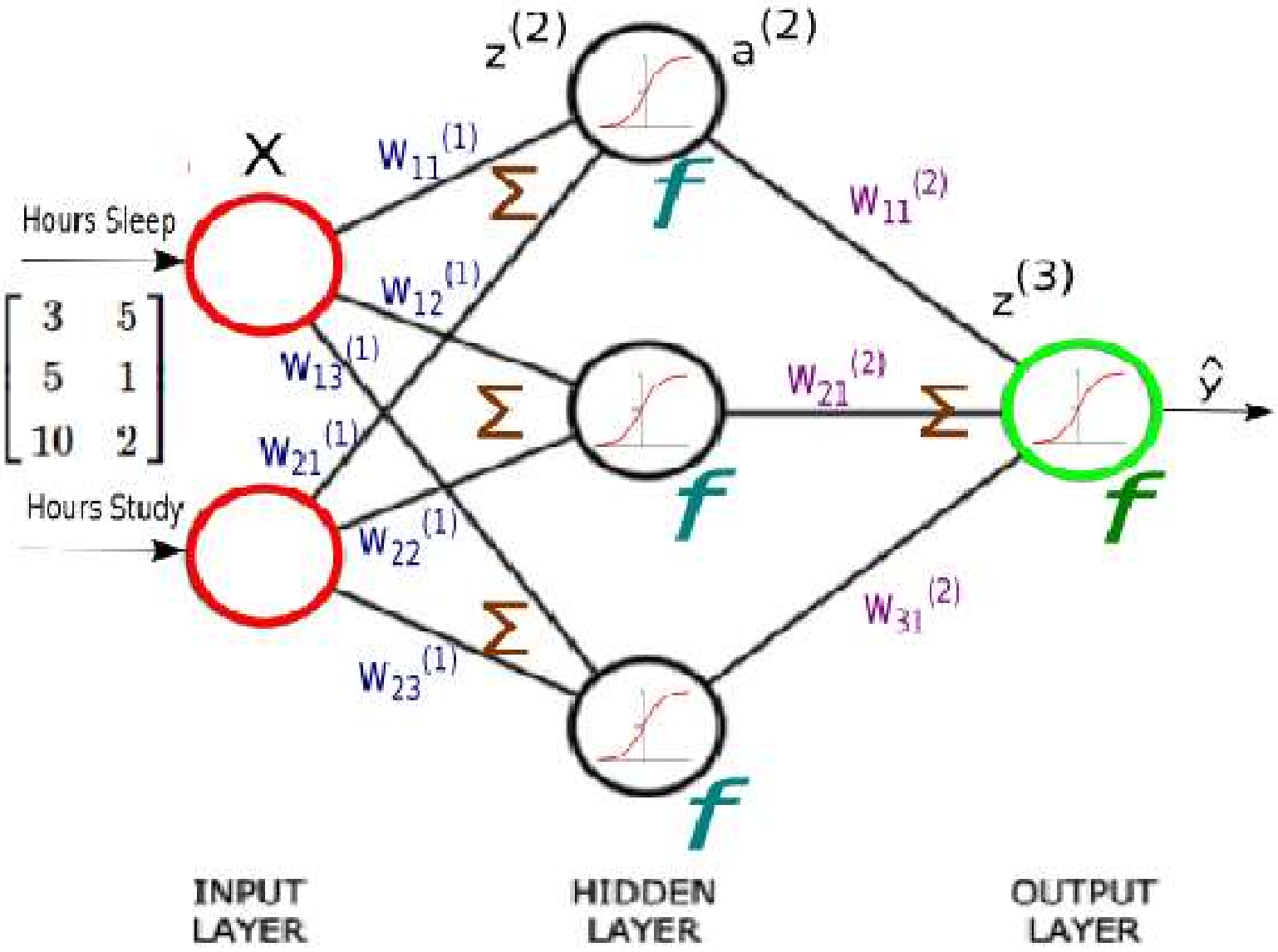
- The main architectures of artificial neural networks, considering the neuron disposition, as well as how they are interconnected and how its layers are composed, classified as follows:
  - single-layer feedforward network,
  - multilayer feedforward networks,
  - recurrent networks and
  - mesh networks.

Feedforward neural network

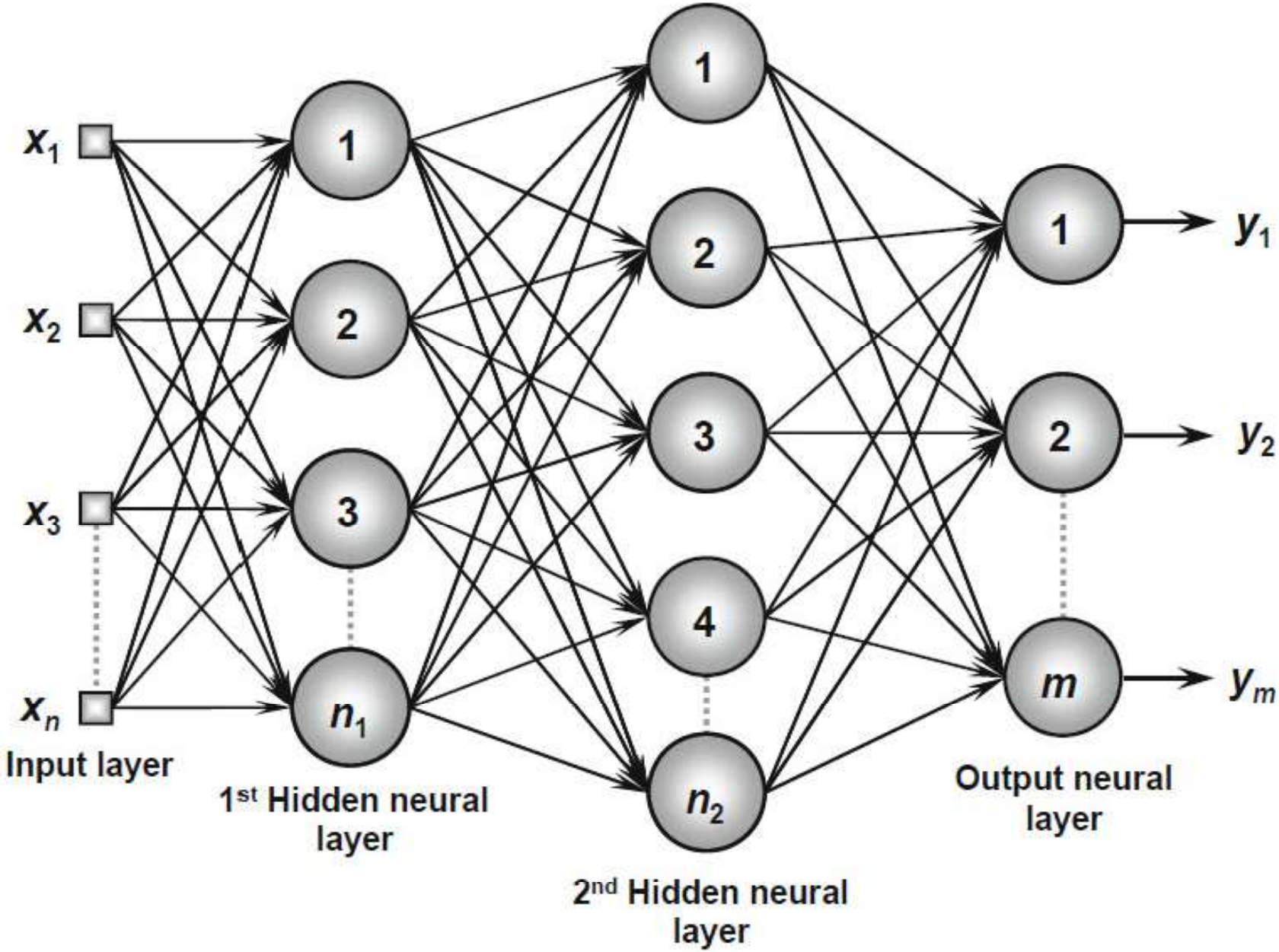
- The simplest type of artificial neural network
- Contains multiple neurons (nodes) arranged in layers
- Nodes from adjacent layers have connections or edges between them
- All these connections have associated weights
- A feedforward neural network consists of three types of layers:



Single-Layer feedforward architecture



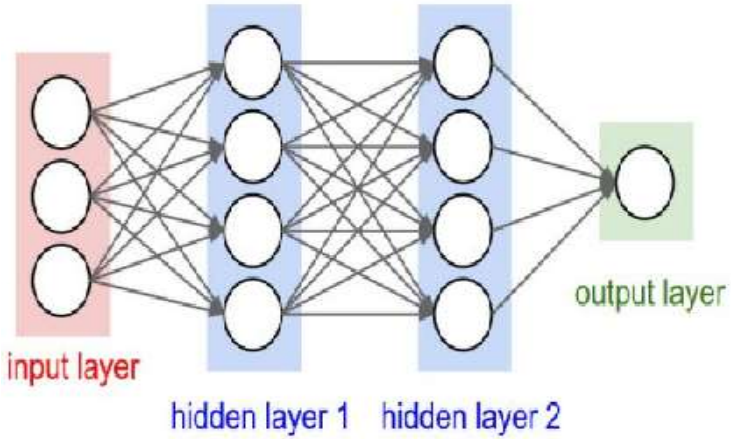
Multiple-Layer feedforward architecture

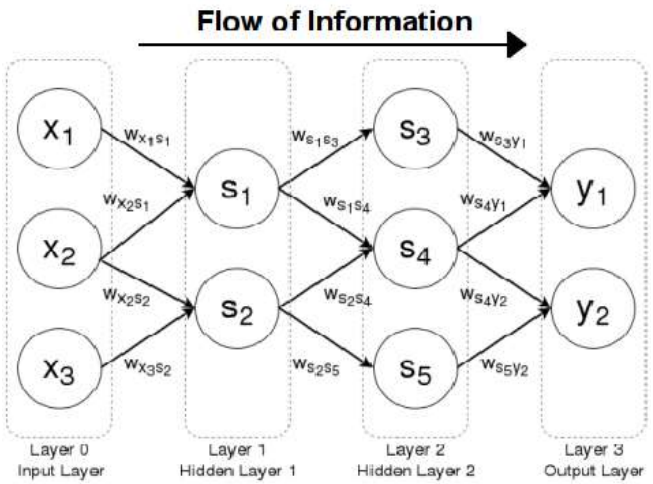


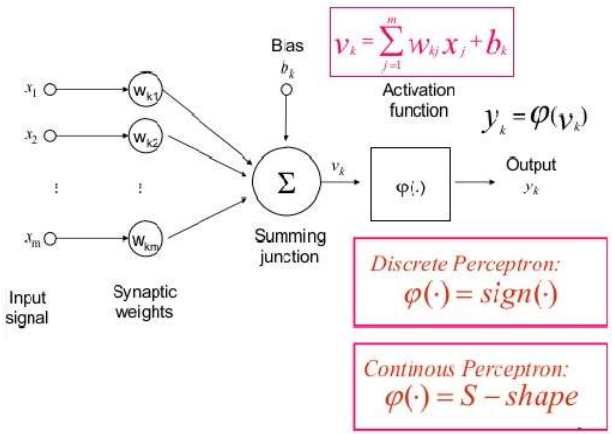
Types of feedforward networks

- Single Layer Perceptron(SLP):
  - The simplest feedforward neural network with no hidden layers
  - Not as such applicable in problem solving

- Multi Layer Perceptron (MLP):
  - Has one or more hidden layers
  - MLP is more applicable than single layer perceptron as many of the problems in the real world deserves more number of computing units
  - MLP can learn non-linear functions (non-linearity)







Multi layer perceptron

- Suppose we have the following student-marks dataset

Hours Studied	Mid Term Marks	Final Term Result
35	67	1 (Pass)
12	75	0 (Fail)
16	89	1 (Pass)
45	56	1 (Pass)
10	90	0 (Fail)

- The two input columns show the number of hours the student has studied and the mid term marks obtained by the student
- The Final Result column can have two values 1 or 0 indicating whether the student passed in the final term
- It is observed that if the student studied 35 hours and had obtained 67 marks in the mid term, the student will pass the final term
- Question: Suppose, Will a student studying 25 hours and having 70 marks in the mid term pass the final term?

Hours Studied	Mid Term Marks	Final Term Result
25	70	?

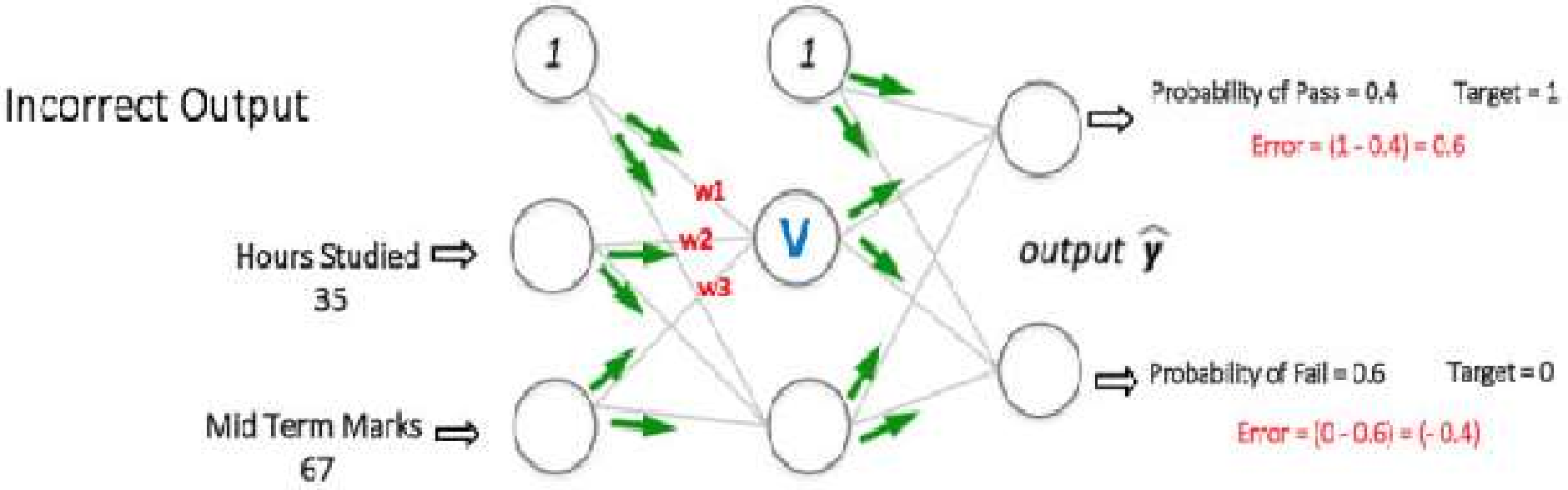
- This is a binary classification problem where the MLP can learn from the given examples (training data) and make prediction given a new data point (test data)

Training MLP: The back-propagation algorithm

- The process by which the MLP learns is called the Backpropagation algorithm
- Backward Propagation of Errors is one of the several ways in which an artificial neural network (ANN) can be trained
- It is a supervised training scheme, that learns from labeled training data
- BackProp is like “learning from mistakes“
- The supervisor corrects the ANN whenever it makes mistakes

Step 1: Forward propagation

- All weights in the network are randomly assigned
- Lets consider the hidden layer node marked V
- Assume the weights of the connections from the inputs to that node are w1, w2 and w3 (as shown)



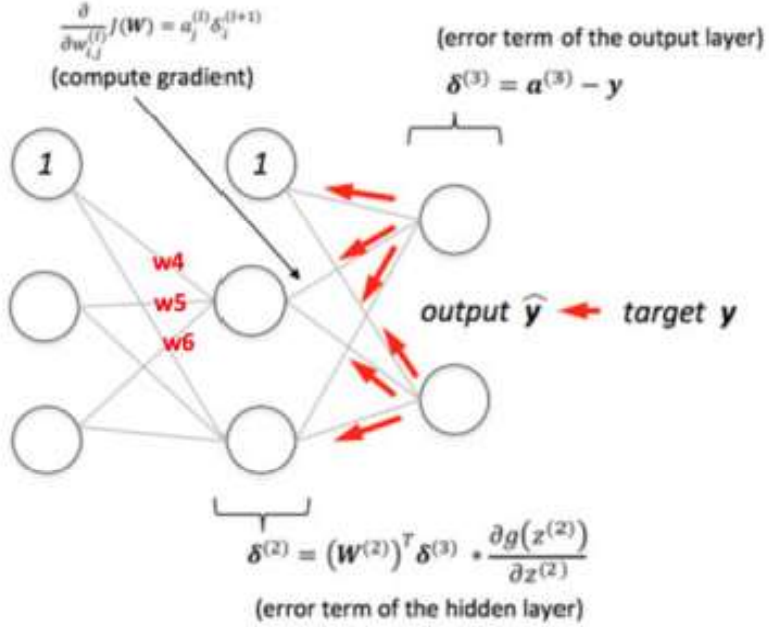
Step 2: Back propagation and weight updation

- Calculate the total error at the output nodes and propagate these errors back through the network using Backpropagation to calculate the gradients

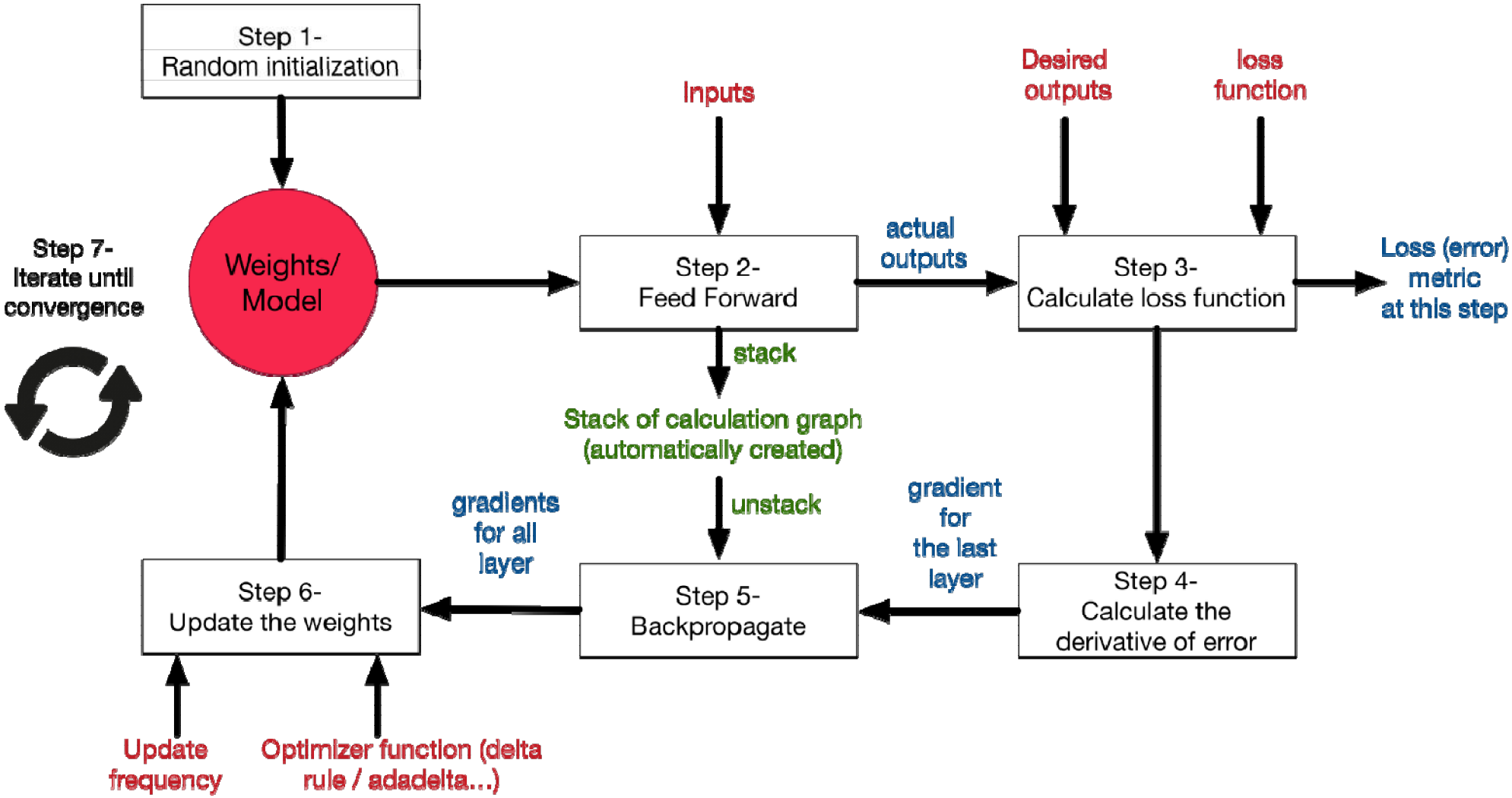
- Then, use an optimization method such as Gradient Descent to ‘adjust’ all weights in the network with an aim of reducing the error at the output layer

- Suppose that the new weights associated with the node in consideration are w4, w5 and w6 (after Backpropagation and adjusting weights)

Backpropagation
+
Weights Adjusted



Process of learning in neural network

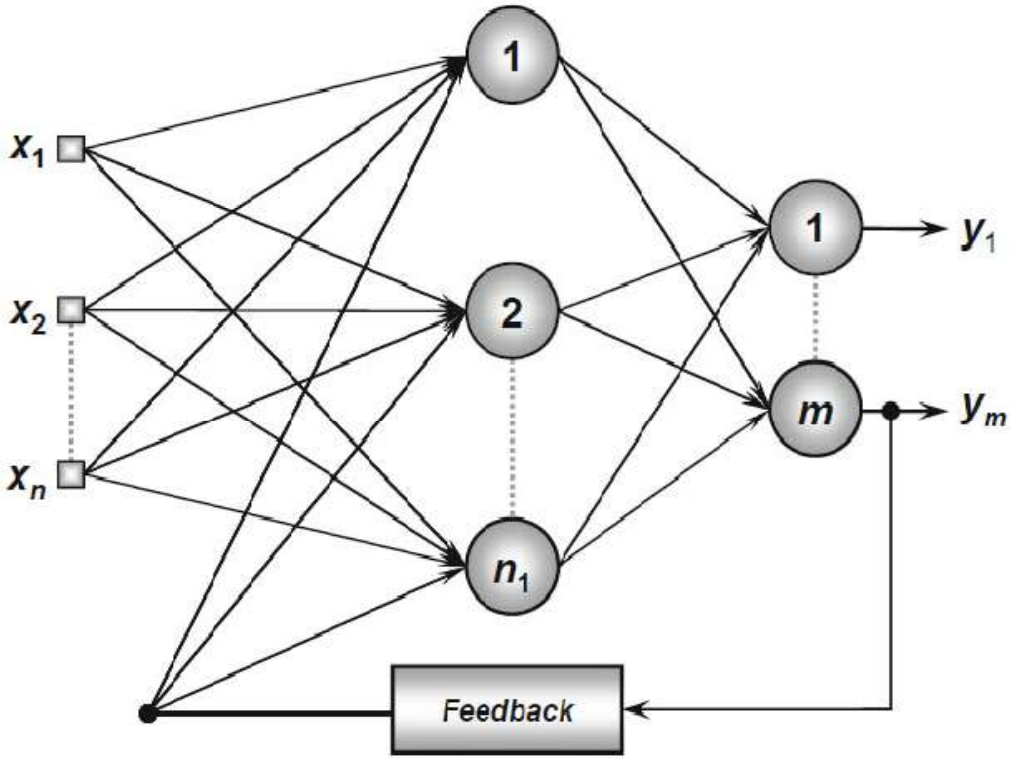


Courtesy: https://medium.com

Recurrent or feedback architecture

- Among the main feedback networks are the Hopfield and the Perceptron with feedback between neurons from distinct layers, whose learning algorithms used in their training processes are respectively based on energy function minimization and generalized delta rule, as will be investigated in the next chapters.

- Thus, using the feedback process, the networks with this architecture produce current outputs also taking into consideration the previous output values.



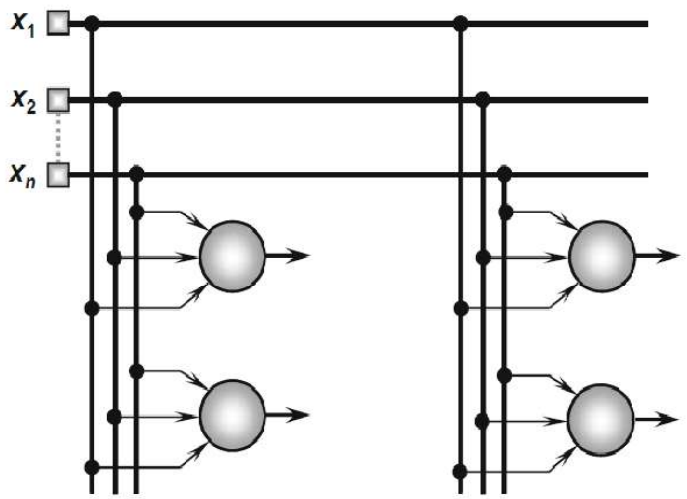
Mesh architectures

- The main features of networks with mesh structures reside in considering the spatial arrangement of neurons for pattern extraction purposes, that is, the spatial localization of the neurons is directly related to the process of adjusting their synaptic weights and thresholds.

- Although we are interested in learning networks of many interconnected units, let us begin by understanding how to learn the weights for a single perceptron.

- Here the precise learning problem is to determine a weight vector that causes the perceptron to produce the correct ±1 output for each of the given training examples.

- Several algorithms are known to solve this learning problem. Here we consider two:
  - The perceptron rule and
  - The delta rule



Gradient-descent (training_examples, η)

- Each training example is a pair of the form <x, t > , where x is the vector of input values, and t is the target output value. η is the learning rate.

Method

- Initialize each wi to some small random value
- Until the termination condition is met, Do
  - Initialize each Δwi to zero
  - For each < x, t > in training_examples, Do
    - Input the instance x to the unit and compute the output o
    - For each linear unit weight wi, Do Δwi = Δwi + η(t − o)xi
- For each linear unit weight wi, Do wi = wi + Δwi




Stochastic gradient-descent (training_examples, η)

- Each training example is a pair of the form <x, t > , where x is the vector of input values, and t is the target output value. η is the learning rate.

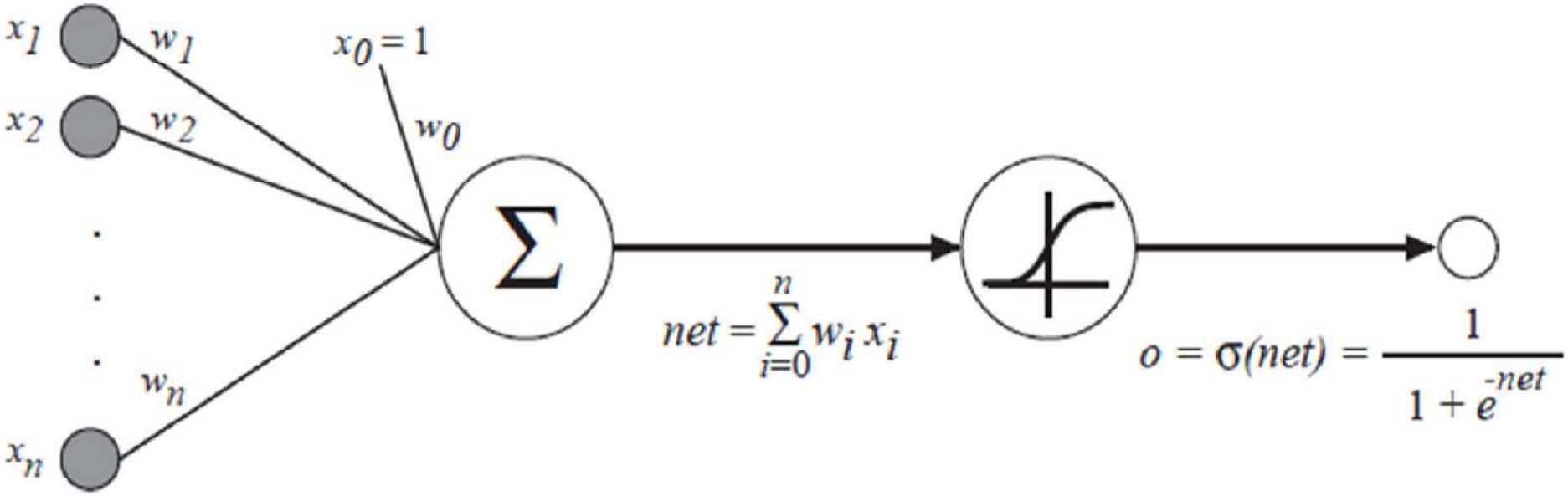
Method

- Initialize each wi to some small random value
- Until the termination condition is met, Do
- Initialize each Δwi to zero
- For each < x, t > in training_examples, Do
  - Input the instance x to the unit and compute the output o
  - For each linear unit weight wi, Do wi = wi + η(t − o)xi

Multilayer networks and backpropagation algorithm

- Single perceptrons can only express linear decision surfaces.

- In contrast, the kind of multilayer networks learned by the backpropagation algorithm are capable of expressing a rich variety of nonlinear decision surfaces.



The backpropagation algorithm

- Because we are considering networks with multiple output units rather than single units as before, we begin by redefining E to sum the errors over all of the network output units

$$E(\vec{w}) \equiv \frac{1}{2} \sum_{d \in D} \sum_{k \in outputs} (t_{kd} - o_{kd})^2$$

- Where outputs are the set of output units in the network, and TKD and OKD are the target and output values associated with the kth output unit and training example d.

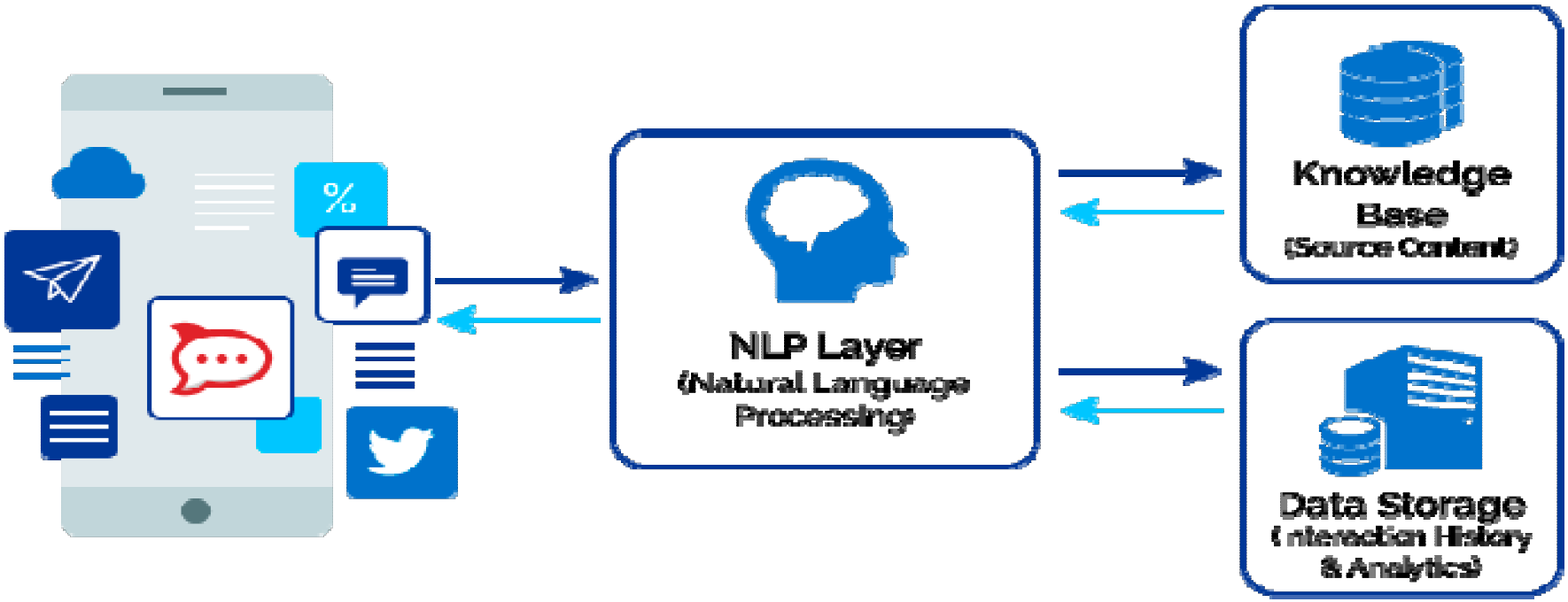
- The learning problem faced by Backpropagation search a large hypothesis space defined by all possible weight values for all the units in the network.

Natural language processing

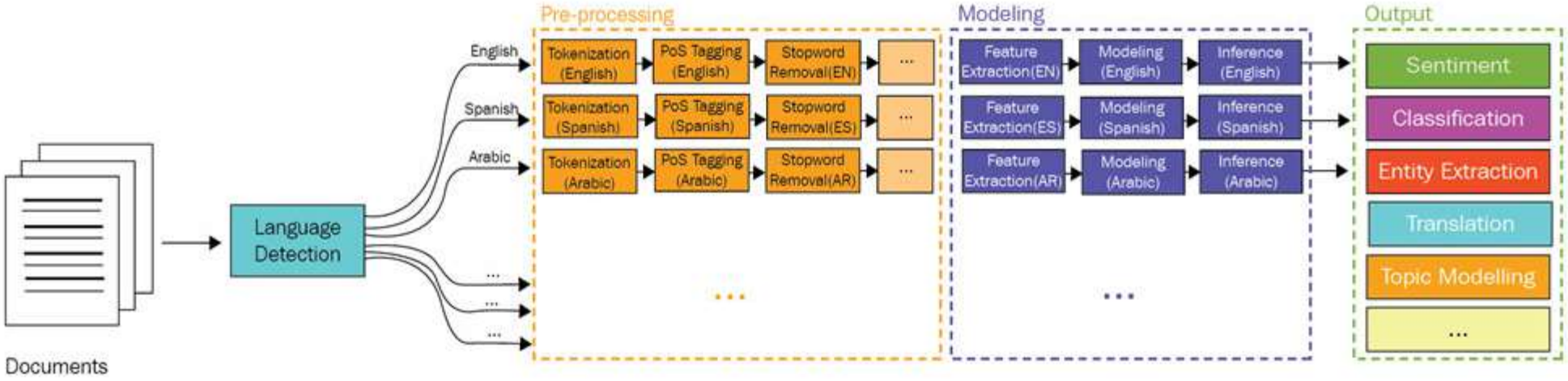
- Natural language processing (NLP) is the field of designing methods and algorithms that take as input or produce as output unstructured, natural language data.

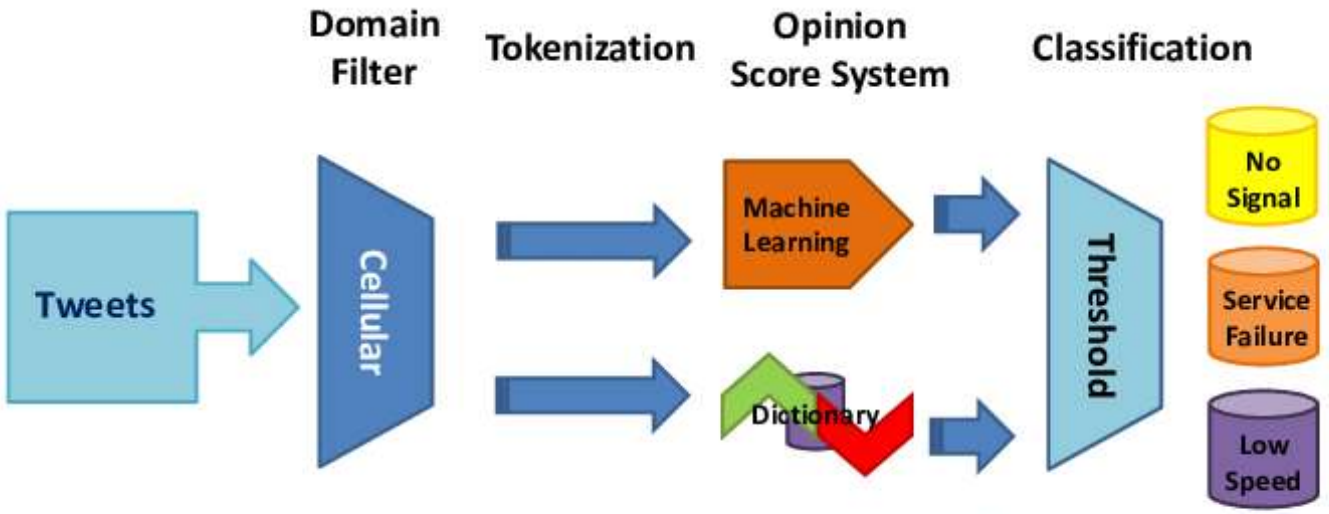
- People are great at producing language and understanding language, and are capable of expressing, perceiving, and interpreting very elaborate and nuanced meanings.

- At the same time, while we humans are great users of language, we are also very poor at formally understanding and describing the rules that govern language.



Classical NLP





Lexicon Example:
@AVOID_TalkTalk we've had no service for 4 days. Is this the worst company in the UK?

Domain Keywords:
[TalkTalk] + [no service]

Tokenization:
avoid talktalk we have had no service for 4 days is this the worst company in the UK

Opinion Score (-3):
avoid
no
worst

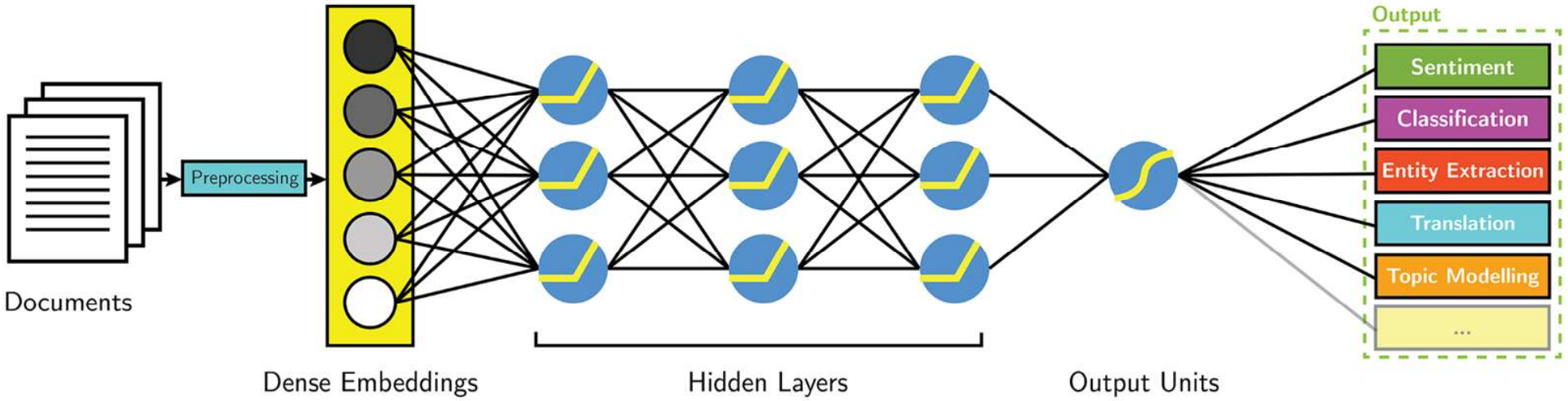
Classification:
Type: No Service
Opinion Score: -3
GPS: 51.54, -0.11
Location: St Mary Magdalene Academy, N7 8PG, London

Feed-forward networks

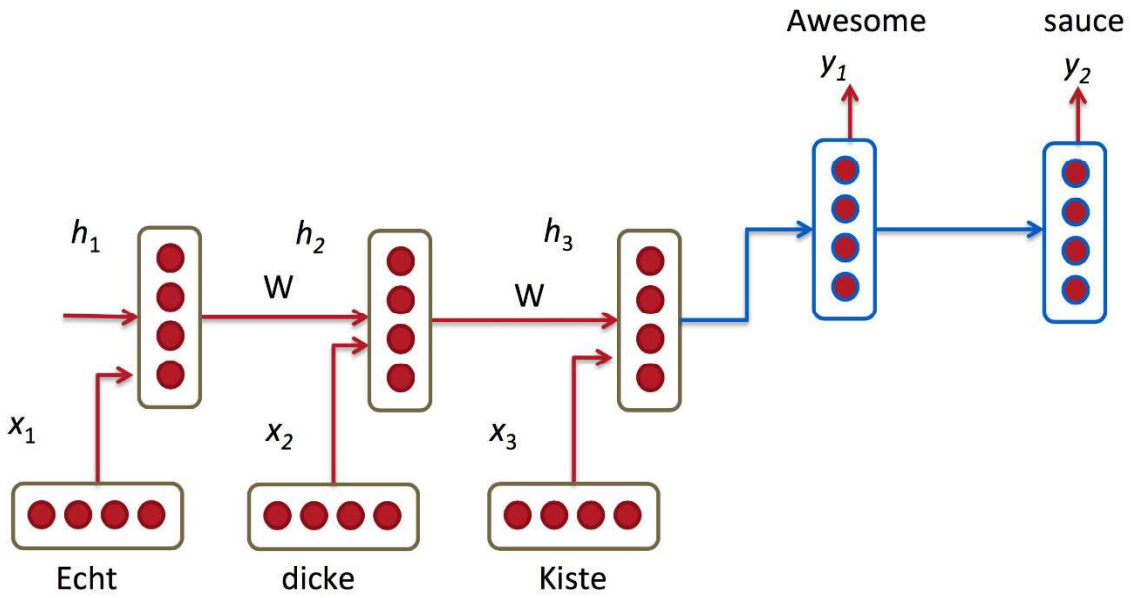
- Feed-forward networks, in particular multi-layer perceptrons (MLPs), allow to work with fixed sized inputs, or with variable length inputs in which we can disregard the order of the elements.

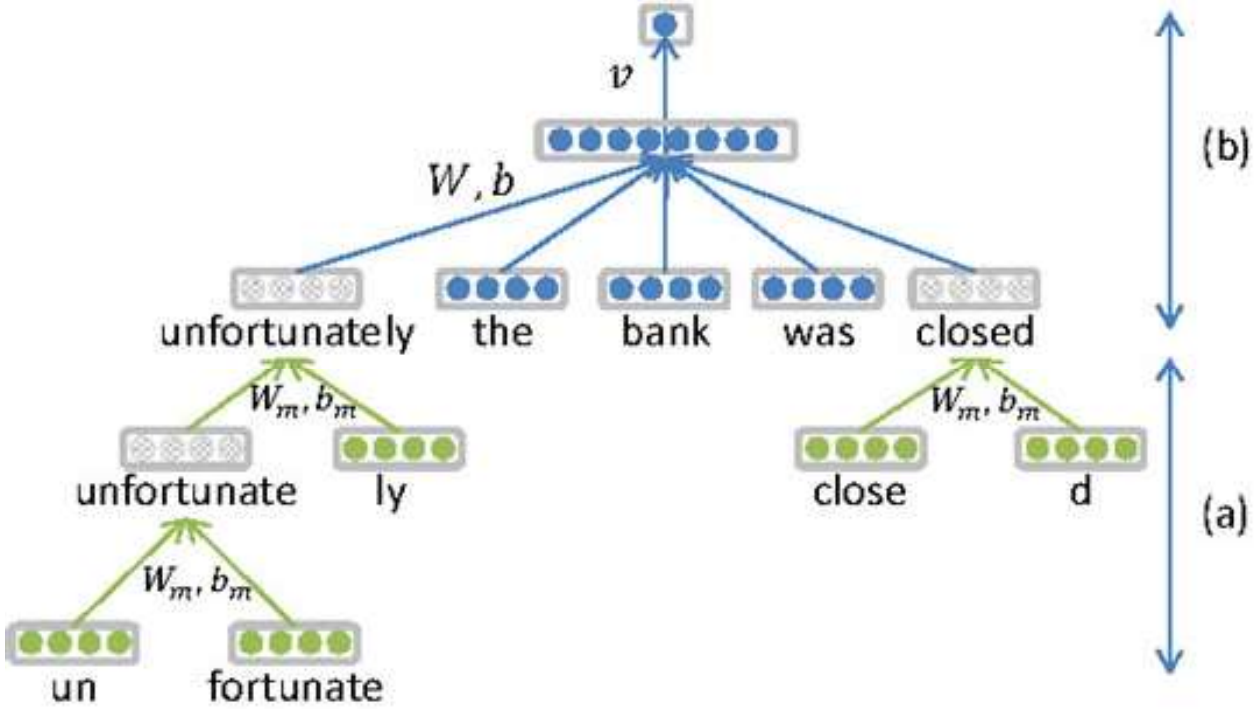
- When feeding the network with a set of input components, it learns to combine them in a meaningful way. MLPs can be used whenever a linear model was previously used.

- The nonlinearity of the network, as well as the ability to easily integrate pre-trained word embeddings, often leads to superior classification accuracy.



Recurrent neural networks and recursive networks





Features for NLP problems

- As ‘words’ and ‘letters’ are discrete items, our features often take the form of indicators or counts.

- An indicator feature takes a value of 0 or 1, depending on the existence of a condition (e.g., a feature taking the value of 1 if the word ‘dog’ appeared at least once in the document, and 0 otherwise).

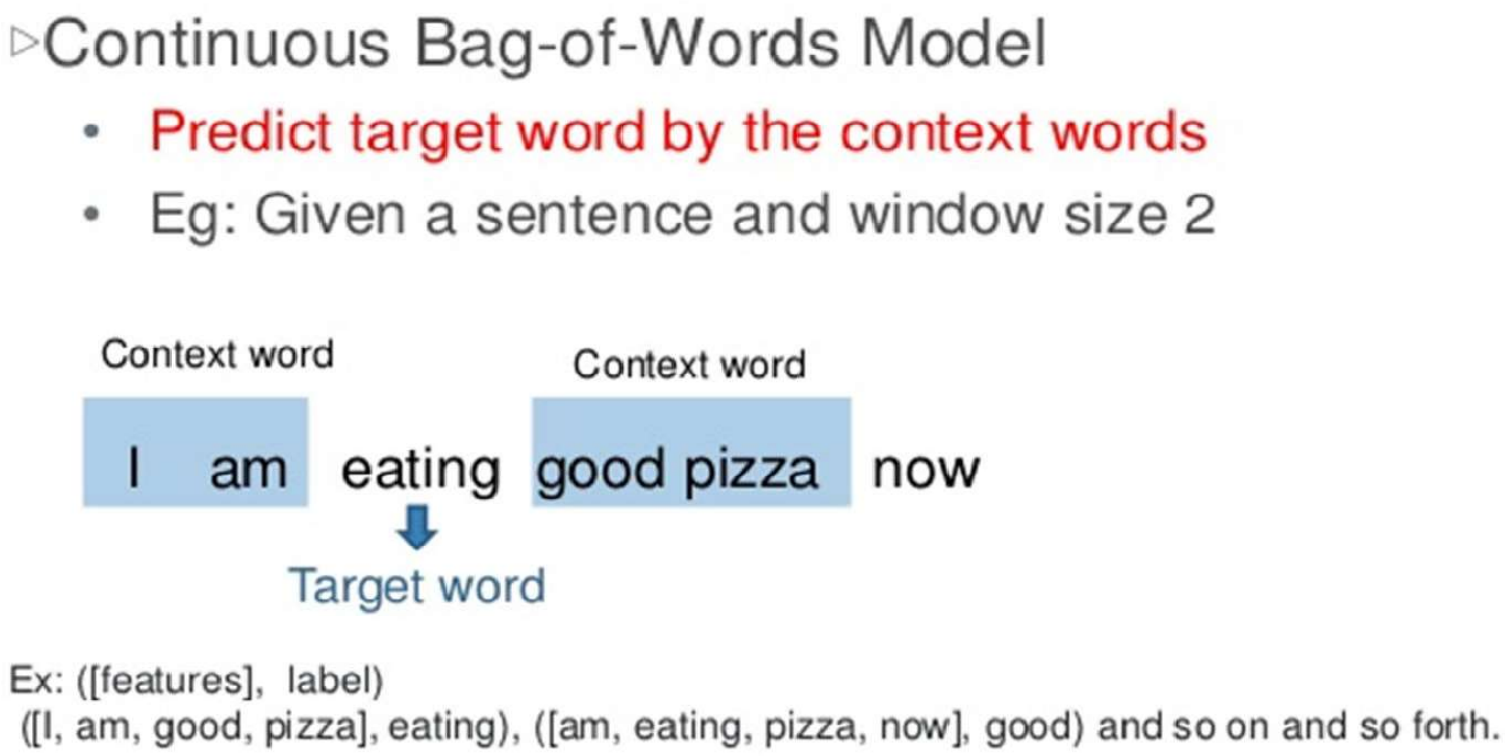
- A count takes a value depending on the number of times some event occurred, e.g., a feature indicating the number of times the word dog appears in the text.

Framenet vs. Wordnet

FrameNet	WordNet
Frame: Typicality **Definition:** Unorthodox or unexpected	**Syn. Set:** Curious, funny, odd, peculiar, etc. **Definition:** Beyond or deviating from the usual or expected
Frame: Mental_stimulus_exp_focus **Definition:** Interested or inquisitive (about something)	**Syn. Set:** Curious **Definition:** Eager to investigate and learn or learn more; having curiosity aroused; eagerly interested in learning more
Frame: Mental_property **Definition:** Driven to investigate and learn	

Features for text

- When we consider a sentence, a paragraph, or a document, the observable features are the counts and the order of the letters and the words within the text.
  - Bag of words
  - Weighting
  - Windows



Ex: ([features], label)
([I, am, good, pizza], eating), ([am, eating, pizza, now], good) and so on and so forth.

Features for word relations

- When considering two words in context, we can also look at the distance between the words.
- We are interested in a conjunction of features occurring together.
- Linear models cannot assign a score to a conjunction of events
- When designing features for a linear model, we must define many ‘combination features
- Neural networks provide nonlinear models.
- This greatly simplifies the work of the model designer

NGRAM features

- NGRAMS—consecutive word sequences of a given length.
- It should be intuitively clear why word-bigrams are more informative than individual words
- A bag-of-bigrams representation proves very hard to beat.
- Of course, not all bigrams are equally informative
- Some form of generalization is achieved across word types by mapping them to coarser grained categories
- These solutions are quite limited
- We are interested in making predictions based on ordered sets of items
- Some of the sentence words are very informative of the sentiment and others less informative

Some terminologies

- Rules written A → B c
  - Terminal vs. non-terminal symbols
  - Left-hand side (head): always non-terminal
  - Right-hand side (body): can be mix of terminal and non-terminal, any number of them
  - Unique start symbol (usually S)
  - '→' "rewrites as", but is not directional (an "=" sign would be better)