# Chapter 7 Neural Networks and Neural Language Models

A modern neural network: a network of small computing units,each of which takes a vector of input values and produces a single output value. In this chapter we use them to classification.

**feedforward network**: computation proceeds iteratively from one layer of units to the next

**deep learning:** the use of modern neural nets

### compared with logistic regression:

more powerful classifier, minimal neural network

**logistic regression**: apply the regression classifier to many tasks by developing many rich kinds of feature templates based on domain knowledge

**neural networks**: build neural networks that take raw words as inputs and learn to induc features as part of the process of learning to classify.

## Units: the building block

- **bias term**: taking a weighted sum of its inputs, with one additional term in the sum, $z = b + \Sigma_i w_i x_i$
- **activation**: z non-linear function f to z

### three popular non-linear function f() below

- **sigmoid function**: $y = \sigma(z) = \frac{1}{1+e^{-z}}$

  map the output in the range [0,1], differentiable

  the output of a neural unit: $y = \sigma(w \cdot x + b) = \frac{1}{1+e^{-w \cdot x + b}}$

- **tanh function**: $y = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

  map the output in range [-1,+1]

- **ReLU**: $y = max(x, 0)$

In the sigmoid or tanh function,very high values of z result in values of y that are **saturated**.By contrast, the tanh function has the nice properties of being smoothly differentiable and mapping outlier values toward the mean.

## The XOR problem

- **perceptron**: a very simple neural unit that has a binary output and does **not** have a non-linear activation function.

  The output is 0 or 1

$$y = \begin{cases} 0, & if\, w \cdot x + b \leq 0 \\ 1, & if\, w \cdot x + b > 0 \end{cases}$$

  A perceptron is a linear classifier

The XOR problem is not a **linearly separable** function

## Feed-Forward Neural Networks

A feedforward network is a multilayer network in which the units are connected with no cycles; the outputs from units in each layer are passed to units in the next higher layer, and no outputs are passed back to lower layers.

In the standard architecture, each layer is **fully-connected**, meaning that each unit in each layer takes as input the outputs from all the units in the previous layer, and there is a link between every pair of units from two adjacent layers

The weight matrix is multiplied by its input vector h to produce the intermediate output z: $z = Uh$