CSCE 790: Neural Networks and Their Applications

AIISC and Dept. Computer Science and Engineering
Email: vignar@sc.edu

Dr. Vignesh Narayanan

August 31, 2023

UNIVERSITY OF
South Carolina

Projects - Start Early!

Your overall final course letter grade will be determined by your grades on the following assessments.

Homework Assignment	15%
Presentation	15%
Midterm Exam (Take home)	15%
Final Project	55%

Linear-in-the-Parameter Networks

- Consider the two-layer NN

$$y = W\phi(vx), \quad \text{output layer activation function is linear.}$$

- If the first layer weights v are predetermined by some apriori technique, then only the second layer weights W and threshold are to be trained
- In this case, we can define $\sigma(x) = \phi(vx)$ so that $y = w\sigma(x)$, where $x \in \mathbb{R}^n$ and $y = \mathbb{R}^m$, $\sigma : \mathbb{R}^n \to \mathbb{R}^L$, L is the number of hidden layer neurons
- $y = W\sigma(x)$ is called *function-link neural network (FLNN, Sadegh, 1993)*
- Here $\sigma(x)$ is allowed to be a general function from $\mathbb{R}^n \to \mathbb{R}^L$ and it is not diagonal.
- RVFL - Random vector functional-link neural network - Stochastic Basis (Igelnik and Pao 1995)

Activation Functions

- The activation function $\sigma(\cdot)$ is selected on a case-by-case basis
- The role of the activation function is to model the behavior of the nerve cell, where there is no o/p below a certain value of the argument of $\sigma(\cdot)$ and it takes a specific magnitude above the value of the argument.
- A general class of monotonically nondecreasing function taking on bounded values at $-\infty$ to ∞ is the sigmoid functions.
- Typically, the normalized amplitude range of the output of a neuron is written as the closed unit interval (e.g., $[0, 1]$, $[-1, 1]$).

Examples of Activation Function

Example (Threshold Function - Heaviside Function)

Let $\alpha = \sum_{j=1}^{n} v_j x_j + v_0$. The threshold function can be defined as $\sigma(\alpha) = \begin{cases} 1 & \text{if} \quad \alpha > 0 \\ 0 & \text{if} \quad \alpha \leq 0 \end{cases}$

Example (Piecewise Linear Function)

$$\sigma(\alpha) = \begin{cases} 1 & \text{if} \quad \alpha \geq \frac{1}{2} \\ \alpha & \text{if} \quad \frac{1}{2} > \alpha > \frac{-1}{2} \\ 0 & \text{if} \quad \alpha \leq \frac{-1}{2} \end{cases}$$

Example (Sigmoid Function)

- $\sigma(\alpha) = \frac{1}{1+\exp^{-\beta\alpha}}$
- β determines slope
- slope at origin is $\beta/4$ and as $\beta \to \infty$, sigmoid $\to$ threshold

Example: Gaussian or Radial Basis Function Network

- An NN activation often used is the Gaussian or RBF (Sanner and Slotine, 1991)
- Given, when $x \in \mathbb{R}$ (is a scalar),

$$\sigma(x) = \exp^{-(x-\mu)^2/2p},$$

  where μ is the mean and p is the variance
- When $x \in \mathbb{R}^n$, $\mu = (\mu_1, \ldots, \mu_n)' \in \mathbb{R}^n$, then $\sigma_j(x) = \exp^{-\frac{1}{2}(x-\mu_j)' P_j^{-1}(x-\mu_j)}$, P_j is an $n \times n$ matrix
- Let $\sigma(x) = (\sigma_1(x), \ldots, \sigma_n(x))'$, then $y = W\sigma(x)$
- Typically, μ, p or P are pre-selected and fixed and only the weights of the o/p layer are trained

Radial Basis Function Network

- The RBF network has a feedforward structure consisting of a single hidden layer of L locally-tuned units which are fully interconnected to an output layer of m linear units
- All hidden units simultaneously receive the n-dimensional real-valued input vector x
- Hidden unit outputs are not calculated using the weighted-sum/sigmoidal activation mechanism
- Output of each hidden layer units z_j is obtained by calculating the "closeness" of the input x to an n-dimensional parameter vector j associated with the j^{th} hidden unit.

$$z_j(x) = \exp^{-\frac{1}{2}(x-\mu_j)' P_j^{-1}(x-\mu_j)}$$

- Output of the network is computed directly as the weighted-sum of the hidden layer outputs

Example: Cerebellar Model Arithmetic Controller (CMAC) Network

- These were introduced by James Albus, 1975
- Instead of RBF, they are made up of spline functions (e.g., 2^{nd} order splines are triangular functions)
- The activation function of CMAC network is called receptive field functions (analogous to the optical receptive fields in the eye)

**Approximation by Superpositions of a Sigmoidal Function

Quick Recap

- Artificial neural networks (a brief evolutionary history)
- ML - Functional view - Models - Parametric models
- McCulloch-Pitts model and its probabilistic interpretation
- Perceptron model
- Multi-layer feedforward neural network
- Role of bias/threshold function
- Some types of activation functions
- Special types of feedforward network architectures - Linear-in-the parameter (FLNN), RBF, CMAC

Learning Paradigms

- **Supervised Learning:** The model is provided with a set of examples of proper behavior (inputs/targets)
- **Unsupervised Learning:** Only inputs are available to the learning model. The model learns to categorize (cluster) the inputs
- **Reinforcement Learning:** The model is only provided with a grade, or score, which indicates performance, and the objective is to maximize the reward over a long-time interval
- Semi-supervised learning – check this out!

Supervised Learning

- Input and target outputs are given for training
- Learning relationship between the input output pairs
- Types:
  - **Regression:** Covers situations where Y is continuous (quantitative)
  - Example: predicting the value of the Dow in 6 months, predicting the value of a given house based on various inputs, etc.
  - **Classification:** Covers situations where Y is categorical (qualitative)
  - Example: Will the Dow be up or down in 6 months? Is this email spam or not?

Revisiting Parametric Models

- Given data:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

- Model Choice:

$$\hat{y}_i = f_\theta(x_i) = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_n x_{ip}$$

- Let $x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$
- Let $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}$
- n, p are number of samples and number of features per sample
- $f_\theta(x)$ is a linear model

Cost Function

- We can measure the accuracy of our hypothesis function by using a cost function

$$J(\theta) = \frac{1}{n} \sum_{i=1,\ldots,n} (\hat{y}_i - y_i)^2 = \frac{1}{n} \sum_{i=1,\ldots,n} (f_\theta(x_i) - y_i)^2$$

- Find θ such that the predicted output is close to the actual output

$$\min_{\theta \in \mathbb{R}^p} J(\theta)$$

Example 2-Parameter Model

- For a fixed θ, $f_\theta(x)$ is a function of x
- Example:

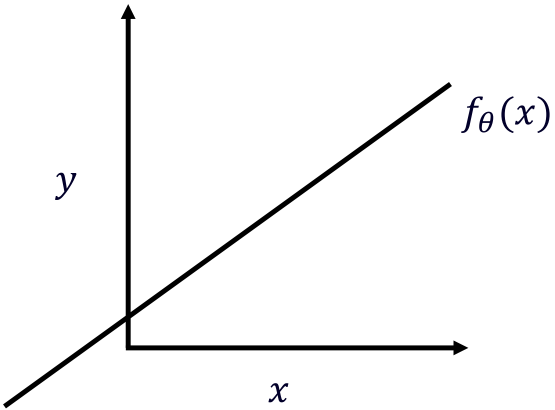

Figure: Example linear function for a fixed θ_0, θ_1

Cost Function

- The cost/objective/loss function is supported on the parameter space
- Example

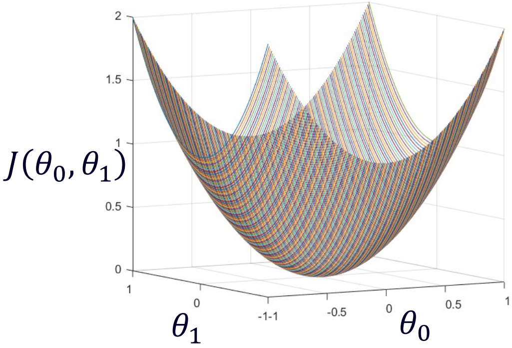

Figure: Example cost function supported in the two-dimensional parameter space (with θ_0, θ_1)

Complex Models

- From MLP to DNN - Parameter space is extremely large
- Example

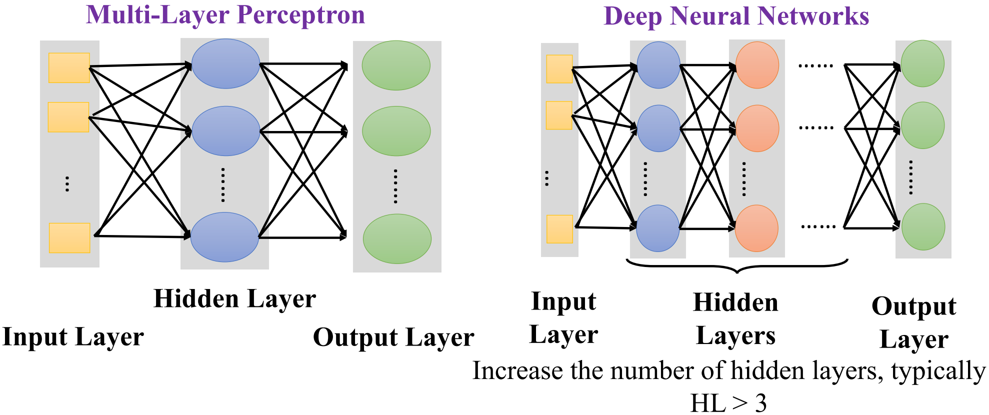

Figure: Feed-forward (Static) NN Models

Second Detour - Linear Algebra - Review

Definition

A tensor is an array of numbers, that may have

- zero dimensions, and be a scalar
- one dimension, and be a vector
- two dimensions, and be a matrix
- or more dimensions.

Second Detour - Linear Algebra - Spaces

Definition (Vector Space)

A vector space V over a field $\mathbb{F}$ is a set of elements called vectors, together with two operations, addition: $V \times V \to V$, $x, y \in V \mapsto x + y \in V$, and scalar multiplication: $\mathbb{F} \times V \to V$, $\alpha \in \mathbb{F}$, $x \in V \mapsto \alpha x \in V$, satisfying for $\forall x, y, z \in V$ and $\forall \alpha, \beta \in \mathbb{F}$:

1. $x + y = y + x$ (additive commutativity)
2. $(x + y) + z = x + (y + z)$ (additive associativity)
3. $\exists 0 \in V : x + 0 = 0 + x = x$ (additive identity)
4. $\exists (-x) \in V : x + (-x) = 0$ (additive inverse)

Vector Space

Definition

5. $\alpha(x + y) = \alpha x + \alpha y$ (scalar distributivity)
6. $(\alpha + \beta)x = \alpha x + \beta x$ (vector distributivity)
7. $(\alpha\beta)x = \alpha(\beta x)$ (multiplicative associativity)
8. $\exists 1 \in \mathbb{F} : 1x = x$ (multiplicative identity)

Example (Vector spaces)

(i) $\{0\}$, the trivial space.

(ii) $\mathbb{R}$ over $\mathbb{R}$.

(iii) $\mathbb{R}^n$ over $\mathbb{R}$.

(iv) Space of $m \times n$ matrices, $\mathbb{R}^{m\times n}$ over $\mathbb{R}$.

Subspace

Definition (Subspace)

A nonempty subset S of a vector space V is called a subspace of V if $\alpha x + \beta y \in S$ for every $x, y \in S$ and every $\alpha, \beta \in \mathbb{R}$.

Remarks

1. By definition, a subspace must contain the null vector 0.
2. V is itself a subspace of V.
3. A subspace not equal to the entire space is said to be a proper subspace.

Span

Definition (Span)

Let V be a vector space. Given $x_1, \ldots, x_m \in V$, the span of $x_1, \ldots, x_m$, denoted by $\mathrm{span}\{x_1, \ldots, x_m\}$, is the set of all vectors v that can be written as $v = \sum_{i=1}^{m} \alpha_i x_i$ for some $\alpha_i \in \mathbb{R}$. That is,

$$\mathrm{span}\{x_1, \ldots, x_m\} = \{v \in V : v = \sum_{i=1}^{m} \alpha_i x_i \text{ for some} \alpha_i \in \mathbb{R}\}.$$

We say v can be written as a linear combination of the vectors $x_1, \ldots, x_m$.

Linear Independence

Definition (Linear Independence)

A set of vectors $x_1, \ldots, x_k$ in a vector space V is said to be linearly independent if $\sum_{i=1}^{m} \alpha_i x_i = 0$, where $\alpha_1, \ldots, \alpha_m$ are constants, implies that $\alpha_i = 0$ for all $i = 1, \ldots, m$. That is,

$$\sum_{i=1}^{m} \alpha_i x_i = 0 \implies \alpha_i = 0, \forall i = 1, \ldots m.$$

Definition (Basis)

If $\mathrm{span}\{x_1, \ldots, x_n\} = V$ and $\{x_1, \ldots, x_n\}$ is a linearly independent set, it is said to be a basis of V.