Lecture 3

谢丹
清华大学数学系

September 22, 2025

CHAPTER 3:
Non-linear effects

Linear Models: Assumptions and Limitations I

Core Assumptions of Linear Regression

The standard linear model rests on two fundamental probabilistic assumptions:

1. **Gaussian Noise:** The target variable y is assumed to have a Gaussian (normal) distribution around its mean.
2. **Linear Mean:** The mean of this distribution is a linear combination of the input features $\mathbf{x}$:

$$\mathbb{E}[y \mid \mathbf{x}] = \mathbf{w}^T\mathbf{x} = w_0 + w_1x_1 + \ldots + w_Dx_D$$

This leads to the familiar model: $y = \mathbf{w}^T\mathbf{x} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Addressing the Limitations

Linear Models: Assumptions and Limitations II

These assumptions are often too restrictive for real-world data. A powerful solution is to project inputs into a feature space using nonlinear basis functions $\phi_j(\mathbf{x})$:

$$\mathbb{E}[y \mid \mathbf{x}] = \sum_{j=0}^{M} w_j \phi_j(\mathbf{x})$$

The model remains linear in the parameters w_j, allowing for easy optimization, but can capture complex, nonlinear relationships in the data.

Generalized Linear Models & Common Basis Functions

Models of the form $\mathbb{E}[y \mid \mathbf{x}] = \sum_j w_j \phi_j(\mathbf{x})$ are the foundation of generalized linear models and linear basis function models.

A Catalog of Common Basis Functions

Type	**Basis Function $\phi_j(x)$**	**Description**
Polynomial	x^j	Simple curves, global influence.
Gaussian	$\exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right)$	Localized "bumps", universal approximator.
Sigmoidal	$\sigma\left(\frac{x-\mu_j}{s}\right)$	Smooth step functions.
Fourier	$\sin(\omega_j x), \cos(\omega_j x)$	For periodic data.

The art of feature engineering often lies in choosing the right type and parameters (e.g., μ_j, s) for these functions.

The Modern Paradigm: Learned Features

Instead of manual feature engineering, deep learning models *learn* optimal hierarchical features $\phi(\mathbf{x})$ directly from the data, automating this crucial step.

What are Generalized Linear Models?

- **Generalized Linear Models (GLMs)** extend linear regression
- Allow for non-normal response distributions
- Three components:
  1. Random component: Probability distribution of response
  2. Systematic component: Linear predictor $\eta = \mathbf{w}^T \phi(\mathbf{x})$
  3. Link function: $g(\mu) = \eta$

GLM Equation

$$\mathbb{E}[y|\mathbf{x}] = \mu = g^{-1}(\mathbf{w}^T \phi(\mathbf{x}))$$

Training

The loss function can be found from the maximal likelyhood estimation following similar method as the linear model. There is no exact solution and so the SGD estimation method is needed.

Classification: Link Functions

In logistic regression, we model the probability using the sigmoid function $\sigma(z)$.

Generalizing the Model

We can generalize this by using any function $f : \mathbb{R} \to [0, 1]$ that maps real values to valid probabilities:

$$p(t = 1 \mid \mathbf{w}, \mathbf{x}) = f(\mathbf{w}^T \mathbf{x})$$

Such a function f is called a link function or activation function in the context of generalized linear models.

Probabilistic Interpretation

Any valid cumulative distribution function (CDF) F of a continuous random variable defines a proper link function:

$$f(x) = \int_{-\infty}^{x} p(\theta)\, d\theta = F(x)$$

The sigmoid function $\sigma(x)$ corresponds to the CDF of the logistic distribution.

Common Link Functions for Binary Classification

A Catalog of Common Activation Functions

Name	Function $f(z)$	Distribution	Properties
Logistic (Sigmoid)	$\frac{1}{1+e^{-z}}$	Logistic	Smooth, symmetric
Probit	$\Phi(z)$	Standard Normal	Symmetric, S-shaped
Log-Log	$1-\exp(-\exp(z))$	Gumbel (Extreme Value)	Asymmetric
Complementary Log-Log	$\exp(-\exp(-z))$	Gumbel (Max Stable)	Asymmetric

The choice of link function can affect model performance, especially in the tails of the distribution.

Handling Nonlinearity: Basis Function Expansion

To model complex, nonlinear relationships, we can project inputs into a feature space using basis functions $\phi_j(\mathbf{x})$:

$$p(t=1 \mid \mathbf{w}, \mathbf{x}) = f\left(\sum_{j=0}^{M} w_j \phi_j(\mathbf{x})\right)$$

The model remains linear in the parameters w_j but can capture nonlinear decision boundaries in the original input space.

Neural Networks Approach

An alternative approach to encode the mean of random variable y is to use Neural Networks, which have proven to be remarkably powerful function approximators.

Neural Network Fundamentals

Definition

A neural network defines a parameterized function $y(x; w)$ where:

- x: Input variables
- y: Output variables
- w: Network parameters

Basic Components

1. **Nodes** (neurons): Computational units
2. **Edges**: Connections between nodes with weights w

Layered Structure

For a network with layer structure, the function at each node is:

$$z_k^{(i)} = h\left(\sum_j w_{kj}^{(i)} z_j^{(i-1)} + b_k^{(i)}\right)$$

Network Implementation

Simplified Notation

By adding a bias node (fixed at 1) to each layer, we get the compact form:

$$\boxed{z_k^{(i)} = h\left(\sum_j w_{kj}^{(i)} z_j^{(i-1)}\right)}$$

Feedforward Process

- Initialize with input layer ($i = 0$) values
- Recursively compute through hidden layers
- Final output at last layer
- Multiple output nodes produce vector outputs

Network Architecture Choices

1. **Structure Design**:
   - Number of layers and nodes per layer
   - Input layer size determined by data dimensionality
   - Deep Neural Network: Many hidden layers
2. **Activation Functions**:

Table: Common Activation Functions

Function	Formula
Logistic Sigmoid	$\sigma(a) = \frac{1}{1+\exp(-a)}$
Tanh	$\tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}$
Hard Tanh	$h(a) = \max(-1, \min(1, a))$
Softplus	$h(a) = \ln(1 + \exp(a))$
ReLU	$h(a) = \max(0, a)$
Leaky ReLU	$h(a) = \max(0, a) + \alpha \min(0, a)$

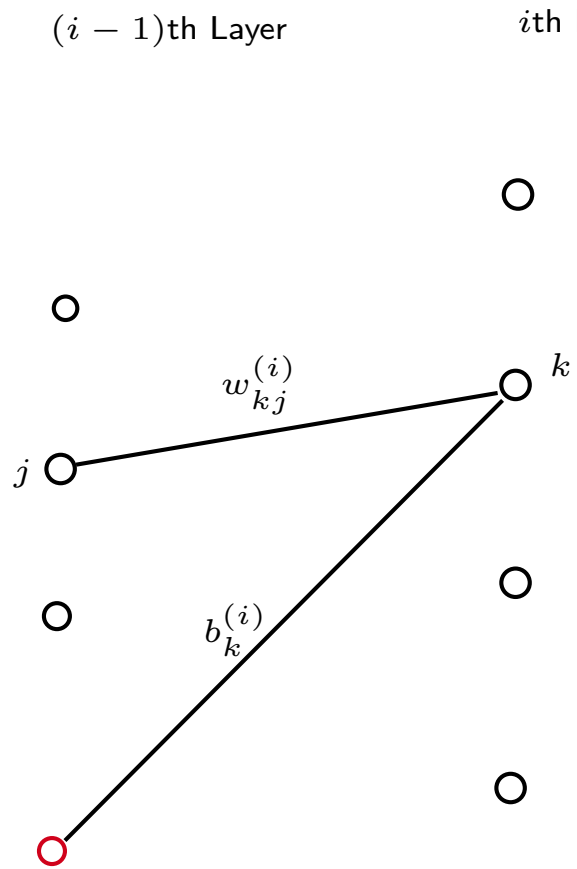
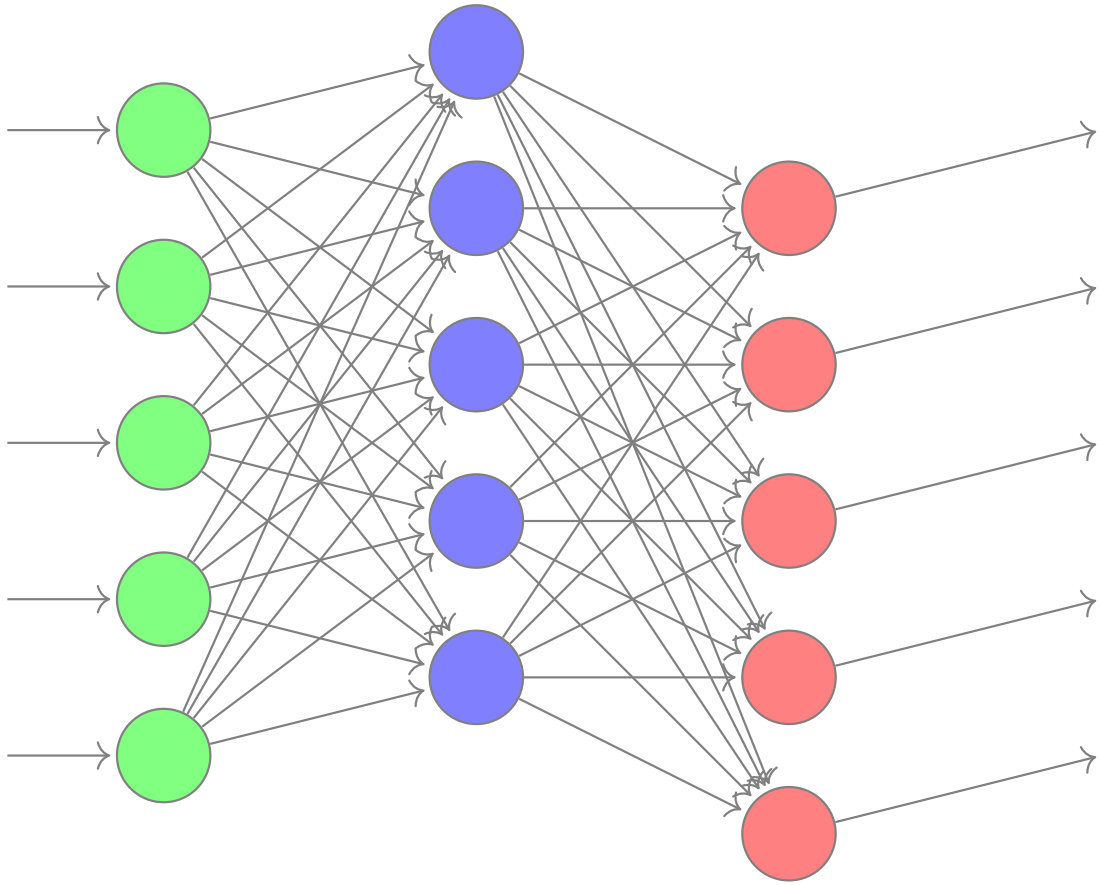


Figure: The building block of neural network.



Here is the probability model for the neural network:

$$P(y|\mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(y|y(x, w), \sigma^2)$$

where $y(x, w)$ is defined by using neural network. Similarly, one can get the probability model for the classification.

Neural Network Loss Function

General Formulation

For both regression and classification problems, the loss function generalizes naturally:

$$E(\mathbf{w}) = \sum_{n=1}^{N} (t_n - y(\mathbf{x}_n, \mathbf{w}))^2 + \lambda \sum_i w_i^2$$

where:

- $y(\mathbf{x}_n, \mathbf{w})$: Neural network output
- λ: Regularization parameter

Per-observation Contribution

The loss decomposes as $E = \sum E_n$ where:

$$E_n = \frac{1}{2}[t_n - y(\mathbf{x}_n, \mathbf{w})]^2 = \frac{1}{2}\left(t_n - h\left(\sum_j w_{kj}^{(f)} z_j^{(f-1)}\right)\right)^2$$

Optimization

The gradient descent method can be applied to minimize $E(\mathbf{w})$, with gradients computed efficiently via:

Backpropagation Algorithm

(based on chain rule differentiation)

对某一个参数 $w_{ji}^{(l)}$ 求导，利用微积分中的链式法则，有

$$\boxed{\frac{\partial E_n}{w_{kj}^{(i)}} = \frac{\partial E_n}{\partial a_k^{(i)}} \frac{\partial a_k^{(i)}}{\partial w_{kj}^{(i)}} = \frac{\partial E_n}{\partial a_k^{(i)}} z_j^{(i-1)} = \delta_k^{(i)} z_j^{(i-1)}}$$

这里我们利用了 $a_k^{(i)} = \sum_j w_{kj}^{(i)} z_j^{(i-1)}$. 这样偏导数就有一个局域的表达形式， 只需要在每一个节点上定义一个新的error项

$$\delta_k^{(i)} = \frac{\partial E_n}{\partial a_k^{(i)}} = \sum_l \frac{\partial E_n}{\partial a_l^{(i+1)}} \frac{\partial a_l^{(i+1)}}{\partial a_k^{(i)}} =$$

$$\sum_l \frac{\partial E_n}{\partial a_l^{(i+1)}} \frac{\partial a_l^{(i+1)}}{\partial a_k^{(i)}} = \sum_l \delta_l^{(i+1)} w_{lk}^{(i+1)} h'(a_k^{(i)})$$

这里我们用到了 $a_l^{(i+1)} = \sum_t w_{lt}^{(i+1)} h(a_t^{(i)})$. 上面的公式给了一个递归的方式来计算error, 但是和原来量不一样的是这个时候 初始的量为输出端。初始条件为

$$\delta_j^{(f)} = \frac{\partial E_n}{\partial a_j^{(f)}} = (t_n - y_n)$$

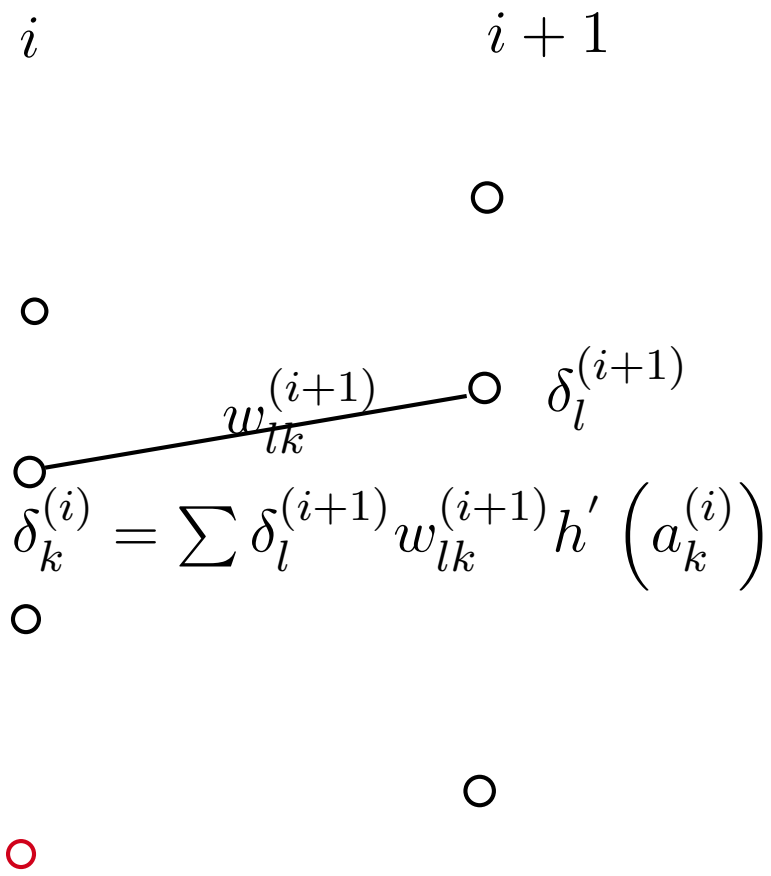


Figure: Backpropagation of gradient

Classification with Neural Networks

Probability Output

For a network with K output nodes $(y_i, i = 1, \ldots, K)$, we can obtain classification probabilities via softmax:

$$p_i = \frac{\exp(y_i)}{\sum_{j=1}^{K} \exp(y_j)}$$

Temperature Parameter

We can introduce temperature T to control output distribution:

$$p_i(T) = \frac{\exp(y_i/T)}{\sum_{j=1}^{K} \exp(y_j/T)}$$

- Higher T: Softer probabilities
- Lower T: More peaked distribution

Generative Applications

The network can parameterize many probability distribution, enabling powerful generative AI models: Text generation, diffusion model for image generation.

Deep Neural Networks

Function Representation

Neural networks provide a geometric framework for representing complex nonlinear functions: universal function approximator.

Depth Advantage

Empirically, deeper networks (more hidden layers) yield better performance for many tasks.

Training Challenges

1. **Gradient Issues**:
   - Vanishing gradients (too small)
   - Exploding gradients (too large)
2. **Computational Complexity**:
   - Millions to billions, trillions of parameters
   - Enormous computational requirements

Solutions

- Architectural innovations address gradient problems
- GPU acceleration enables large-scale training

Computational Considerations

Matrix Operations

- Neural networks primarily perform matrix multiplications
- These operations are highly parallelizable
- GPUs excel at parallel computation of these operations

Efficiency

- Modern GPUs can perform thousands of operations simultaneously
- Specialized tensor cores further accelerate training

Practical Example

Coming Next:

Hands-on Demonstration

of Neural Network Implementation