这份笔记是基于你提供的 220 页（实际编号至 184 页）《神经网络》课程幻灯片整理而成的复习要点。内容涵盖了从神经网络的动力、架构到误差传播及高级二阶优化算法的全面解析。

机器学习笔记：神经网络 (Neural Networks)

一、 神经网络的动机与背景 (Motivation)

1. 线性模型的局限性

- **非线性分类问题**：简单的线性假设无法处理复杂的非线性边界（如 XOR 问题）。
- **特征爆炸**：为了处理非线性，引入高阶多项式特征会导致特征数量 $O(D^M)$ 爆炸式增长（D 为输入维度，M 为阶数）。
- **视觉识别挑战**：图像在计算机眼中是像素矩阵（Pixel Intensity），简单的线性组合难以处理光照、旋转、平移等变化。

2. 维度灾难 (The Curse of Dimensionality)

- **空间分割**：随着维度 D 增加，填充空间的单元格数量呈指数增长，所需的训练数据量也呈指数增长。
- **高维空间特性**：在高维球体中，大部分体积集中在靠近表面（$r = 1$）的薄壳内。
- **数据流形 (Data Manifolds)**：实际数据往往分布在低维流形上，神经网络通过学习这些流形来降维和提取特征。

3. 生物学启发

- **“统一学习算法”假设**：大脑皮层（听觉、躯体感觉等）具有普适的学习机制（如：通过手术将视觉信号输入听觉皮层，听觉皮层也能学会“看”）。
- **传感器表示**：人类可以利用舌头“看”、利用回声定位或植入第三只眼，证明了大脑对传感器信号的自适应能力。

二、 前馈网络函数 (Feed-forward Network Functions)

1. 神经元模型：逻辑单元 (Logistic Unit)

- 每个神经元接收输入 $\mathbf{x}$，加权求和得到激活值 a，再通过非线性激活函数 $h(\cdot)$ 得到输出 z。
- **公式**：$a_j = \sum_{i=0}^{D} w_{ji}^{(1)} x_i$，$z_j = h(a_j)$。

2. 网络架构

- **多层感知机 (MLP)**：由输入层、隐藏层和输出层组成。
- **通用逼近定理 (Universal Approximators)**：具有足够多隐藏单元的双层感知机（带线性输出）可以以任意精度逼近任何连续函数。

3. 激活函数 (Activation Functions)

名称	公式	特点
Sigmoid	$\sigma(a) = \frac{1}{1+e^{-a}}$	输出范围 (0,1)，常用于二分类输出
tanh	$\frac{e^{a}-e^{-a}}{e^{a}+e^{-a}}$	输出范围 (-1,1)，中心化，导数 $h' = 1 - h^2$
ReLU	$\max(0, a)$	解决梯度消失问题，计算简单
Softplus	$\ln(1 + e^{a})$	ReLU 的平滑版本
Leaky ReLU	$\max(0, a) + \alpha \min(0, a)$	解决 ReLU 负半轴“死区”问题

三、 误差函数 (Error Functions)

1. 回归 (Regression)

- **假设**：目标变量 t 服从以 $y(\mathbf{x}, \mathbf{w})$ 为均值的高斯分布。
- **损失函数**：平方误差和 (Sum-of-squares error)。
  - $E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(\mathbf{x}_n, \mathbf{w}) - t_n\}^2$

2. 二分类 (Binary Classification)

- **假设**：目标变量服从伯努利分布 (Bernoulli distribution)。
- **损失函数**：交叉熵 (Cross-entropy error)。
  - $E(\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$

3. 多分类 (Multiclass Classification)

- **1-of-K 编码**：目标 t 是一个仅有一位为 1 的向量。
- **输出层**：使用 **Softmax** 函数，确保输出代表互斥类别的概率。
  - $y_k(\mathbf{x}, \mathbf{w}) = \frac{\exp(a_k(\mathbf{x},\mathbf{w}))}{\sum_j \exp(a_j(\mathbf{x},\mathbf{w}))}$
- **损失函数**：$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$

四、 梯度下降与优化 (Gradient Descent)

1. 误差表面 (Error Surfaces)

- 包含局部最小值 (Local minima) 和全局最小值 (Global minimum)。
- **Hessian 矩阵**：二阶导数矩阵，决定了误差表面的曲率。
- **局部二次逼近**：利用泰勒展开分析误差表面特性。

2. 梯度下降策略

- **批量梯度下降 (Batch GD)**：利用整个训练集计算梯度。稳定但计算量大。
- **随机梯度下降 (SGD)**：每次仅利用一个样本。可以逃离局部最小值，处理数据冗余高效。
- **小批量梯度下降 (Mini-batch GD)**：权衡方案，利用硬件加速（GPU）提高效率。

五、 误差反向传播 (Backpropagation)

1. 核心思想：链式法则

- **第一步：前向传播 (Forward Pass)**：计算各层神经元的激活值 a_j 和输出值 z_j。
- **第二步：计算误差信号 δ (Error Signal)**：
  - 输出层：$\delta_k = y_k - t_k$
  - 隐藏层：$\delta_j = h'(a_j)\sum_k w_{kj}\delta_k$ (将误差从后层向前层加权回传)

- **第三步：计算偏导数**：$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$

2. 效率分析

- **反向传播**：计算梯度的时间复杂度为 $O(W)$（W 为参数总数）。
- **数值微分比较**：使用有限差分 (Finite differences) 计算梯度的复杂度为 $O(W^2)$，常用于**梯度检查 (Gradient checking)** 以验证反向传播代码的正确性。

六、 正则化与泛化 (Regularization)

1. 权重衰减 (Weight Decay)

- 在误差函数中加入二次项 $\frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$。
- 对应于参数的零均值高斯先验。
- **一致性问题**：简单的权重衰减对权重和偏置一视同仁，但在输入/输出线性变换下不具有平移和缩放不变性。建议对不同层、权重与偏置使用不同的超参数。

2. 早停法 (Early Stopping)

- 训练时监控验证集误差。当验证集误差上升时停止训练，防止模型过度拟合噪声。

3. 不变性 (Invariances)

- **数据增强 (Data Augmentation)**：通过平移、旋转、缩放原始图像产生副本。
- **切空间传播 (Tangent Propagation)**：通过正则化惩罚模型在变换方向（流形切向量）上的变化。
- **结构化不变性**：利用 **卷积神经网络 (CNN)**，通过局部感受野和权重共享实现平移不变性。

七、 高级主题：Jacobian 与 Hessian

1. Jacobian 矩阵

- **定义**：网络输出相对于输入的导数 $J_{ki} = \frac{\partial y_k}{\partial x_i}$。
- **作用**：衡量输出对输入变化的敏感度，用于误差传播分析。

2. Hessian 矩阵

- **作用**：用于二阶优化算法、网络剪枝（Pruning）、拉普拉斯近似（Bayesian NN）。
- **近似方法**：
  - **对角近似**：忽略非对角元素，求逆简单。
  - **外积近似 (Levenberg-Marquardt)**：$H \approx \sum \nabla y_n \nabla y_n^T$，仅需一阶导数。
- **快速乘法**：利用 $\mathcal{R}$ 算子 (Pearlmutter, 1994)，可以在 $O(W)$ 时间内直接计算 $v^T H$，而无需显式求出巨大的 Hessian 矩阵。

复习提示：重点记忆反向传播中 δ 的传递公式，以及各种损失函数与输出层激活函数的配对关系（回归-线性、分类-Sigmoid/Softmax）。