线性模型研究报告

作者：邵芃 学号：22371480

一、实验背景与目标

本实验基于一组具有显著非线性特征的二维数据（含训练集与测试集），旨在探究以下问题：

1. **线性模型分析**：通过最小二乘法（OLS）、梯度下降法（GD）和牛顿法实现线性回归，对比不同算法在训练与测试误差上的表现；
2. **非线性改进策略**：采用多项式回归优化模型，验证非线性模型对性能的提升效果。

二、方法设计与理论推导

2.1 线性回归模型

最小二乘法（OLS）

核心思想是通过最小化残差平方和确定最优参数，数学表达式为：

$$\theta = (X^T X)^{-1} X^T Y$$

其优势在于解析解的精确性，但计算复杂度随数据维度升高显著增加。

梯度下降法（GD）

通过迭代优化损失函数 $J(\theta)$ 求解参数，更新规则为：

$$\theta_j := \theta_j - \alpha \frac{\partial J}{\partial \theta_j} \quad (j = 0,1)$$

其中学习率 α 需谨慎选择以避免震荡或收敛过慢。

牛顿法

引入二阶导数信息加速收敛，参数更新公式为：

$$\theta^{(k+1)} = \theta^{(k)} - \alpha H^{-1} \nabla J\left(\theta^{(k)}\right)$$

Hessian 矩阵 H 的计算提高了算法复杂度，但收敛速度显著优于梯度下降。

2.2 非线性模型：多项式回归

通过扩展特征维度捕捉非线性关系，模型表达式为：

$$f(x) = \sum_{k=0}^{n} a_k x^k$$

实验中分别测试二次、三次及五次多项式，探究复杂度与泛化能力的平衡。

三、实验结果与深度分析

3.1 线性模型对比

三种方法在相同数据集上表现完全一致：

- **训练误差（MSE）**：0.6134
- **测试误差（MSE）**：0.5950

结论：

- 算法实现正确性得以验证；
- 高误差值表明线性模型难以捕捉数据的非线性模式，模型表达能力受限。

3.2 多项式回归性能

多项式阶数	训练误差（MSE）	测试误差（MSE）
2	0.5654	0.5346
3	0.5653	0.5368
5	0.5252	0.5151

关键发现：

1. **二次与三次模型**：初步捕捉数据趋势，但局部区域拟合不足；
2. **五次模型**：训练误差显著降低，测试误差同步下降，未出现过拟合现象；
3. **复杂度权衡**：实验范围内，高阶模型未导致泛化性能恶化，推测数据内在规律允许适度增加复杂度。

四、实验总结与启示

1. **线性模型局限性**：三种算法理论结果一致，但受限于数据非线性特征，预测能力较弱；
2. **非线性改进有效性**：多项式回归通过特征扩展显著提升性能，五次模型达到最优平衡；
3. **过拟合风险警示**：尽管当前实验未观察到过拟合，但更高阶模型需警惕训练误差与测试误差的分化趋势。

实践建议：

- 针对非线性数据，优先选择多项式回归或核方法；
- 算法选择需综合考虑计算效率（如 OLS 的解析解优势）与迭代优化（如牛顿法的快速收敛）。

附录：完整代码与可视化结果详见实验附件。