线性模型拟合

韩琳
hanlin3309@163.com

Abstract

该作业使用题目给出的三种线性方法对数据进行拟合，归一化后得到的训练误差和测试误差都大于0.6，拟合效果不佳，所以本文又尝试使用n次多项式拟合，最终发现项数为11的多项式拟合效果最好，MSE在测试集上显著下降到0.33，尝试使用SVR回归，MSE下降至0.26，体现了数据分布的非线性关系。

Introduction

线性回归（Linear Regression）是一种统计学方法，用于研究一个或多个自变量（特征）与因变量（目标变量）之间的关系。通过建立自变量和因变量之间的线性关系，线性回归能够对因变量进行预测。在简单的线性回归模型中，目标是通过一条直线来拟合数据，使得预测值与真实值之间的误差最小化，即求得$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$的最小化。本作业对训练数据使用最小二乘法、梯度下降法(GD)和牛顿法等常见线性拟合方法进行线性回归拟合，并在测试集上验证拟合效果。本作业给出的数据集分布呈非线性，故在线性方法拟合不佳时通过非线性拟合方式进行回归拟合，在验证集上得到较好结果。

Methodology

1.最小二乘法(Least Squares Method)

最小二乘法是一种通过最小化预测值与实际值之间差异的平方和来估计模型参数的方法，广泛应用于回归分析。

给定数据点 $((x_1, y_1), (x_2, y_2) \dots (x_n, y_n))$，我们希望找到一个线性模型：

$$y = \beta_0 + \beta_1 x$$

β_0 是截距，β_1 是斜率,RSS是误差平方和，$\bar{x}$和 $\bar{y}$ 是 x 和 y的均值。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

最小二乘法的目标是使得 (RSS) 最小化,解为：

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

2.梯度下降法(Gradient Descent)

梯度下降法是一种常用于优化问题的迭代算法，特别是在机器学习和深度学习中。它的目的是通过不断更新模型参数，最小化损失函数.

- 假设我们有一个损失函数 $J(\theta)$，其中 θ 是模型参数。梯度下降的目标是最小化 $J(\theta)$，即找到使得损失函数最小的参数 θ。
- 梯度下降的更新规则为：

$$\theta := \theta - \alpha \nabla J(\theta)$$

θ 是模型的参数，$-\alpha$ 是学习率，控制每次更新的步长，$-\nabla J(\theta)$ 是损失函数 $J(\theta)$ 对 θ 的梯度。

3.牛顿法（Newton's Method）

牛顿法是一种用于求解方程零点（即根）或最优化问题的迭代方法。它基于泰勒级数展开，通过迭代逼近方程的根或函数的最小值。

牛顿法是一个迭代过程，步骤如下：

1. 选择一个初始猜测值 x_0。
2. 根据牛顿法公式计算下一步的 x_{n+1}。
3. 重复步骤 2，直到满足收敛条件（例如，$|x_{n+1} - x_n|$ 小于预设的容忍度）。

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

4.多项式拟合（Polynomial Fitting）

多项式拟合的目标是找到一个多项式函数 $p(x)$，使其通过或尽量接近给定的数据点。一个 n 次多项式的形式为：

$$p(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n$$

其中，$\beta_0, \beta_1, \ldots, \beta_n$ 是多项式的系数，n 是多项式的次数。
在多项式拟合中，常用最小二乘法来求解多项式的系数。最小二乘法的目标是最小化预测值和实际观测值之间的误差平方和。给定数据点 $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$，我们希望找到系数 $\beta_0, \beta_1, \ldots, \beta_n$，使得误差平方和最小化。

$$MSE = \frac{1}{n} \sum_{i=1}^{m} (y_i - p(x_i))^2$$

5.支持向量回归（SVR，Support Vector Regression）

支持向量回归（SVR）是支持向量机（SVM）的一种扩展，用于回归问题。SVR通过找到一个最佳的回归函数来拟合数据，使得大部分的样本点都位于一个特定的容忍度范围内，且尽量减少模型的复杂度。

损失函数

$$L_\epsilon(y, f(x)) = \begin{cases} 0, & \text{if } |y - f(x)| \leq \epsilon \\ |y - f(x)| - \epsilon, & \text{otherwise} \end{cases}$$

- y 是实际值，
- $f(x)$ 是预测值，
- ϵ 是容忍度，控制误差的大小。

优化目标

$$\min_{\mathbf{w}, b, \epsilon} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \$\$$$

- $\mathbf{w}$ 是模型的权重，
- b 是偏差项，
- ξ_i 是松弛变量，用于允许某些点偏离容忍范围，
- C 是正则化参数，控制模型的复杂度。

本实验中正则化参数为1000，表示模型正则化比较高

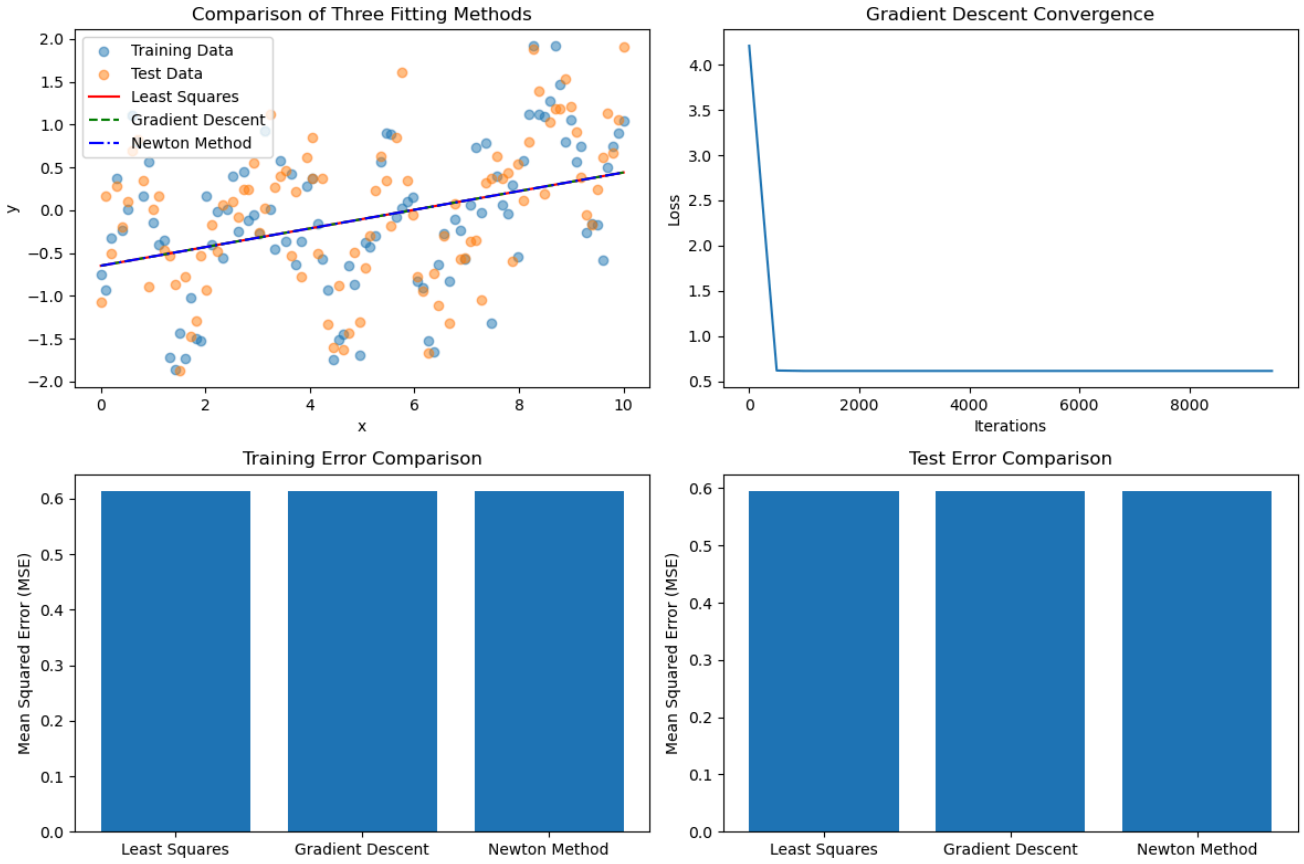
核函数

- 线性核：$K(x, x') = x^T x'$
- 多项式核：$K(x, x') = (x^T x' + c)^d$
- 高斯核（RBF核）：$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$
  本实验中使用高斯核函数，对非线性数据进行拟合

Experimental Studies

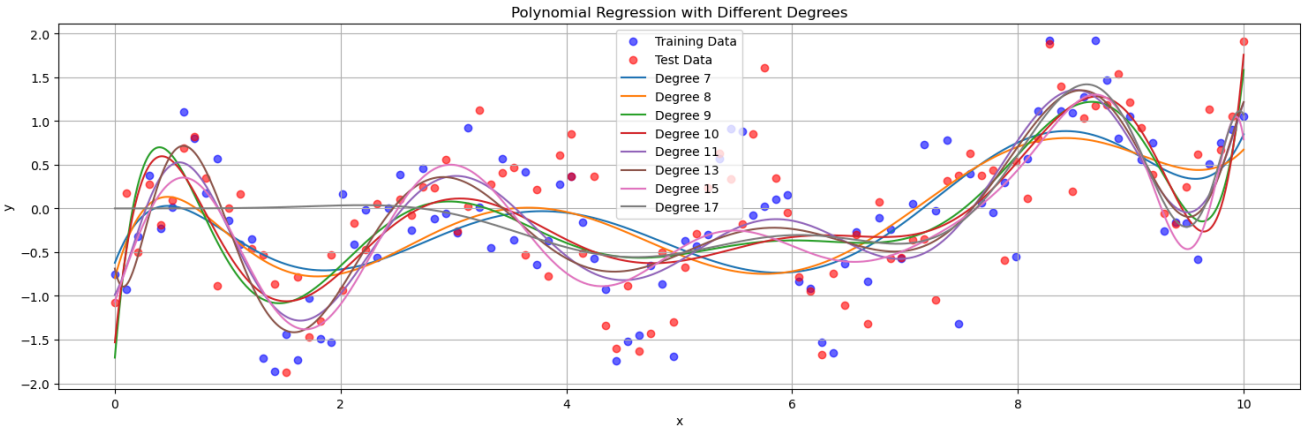
- 三种线性拟合方法得到的拟合直线图和误差值如下

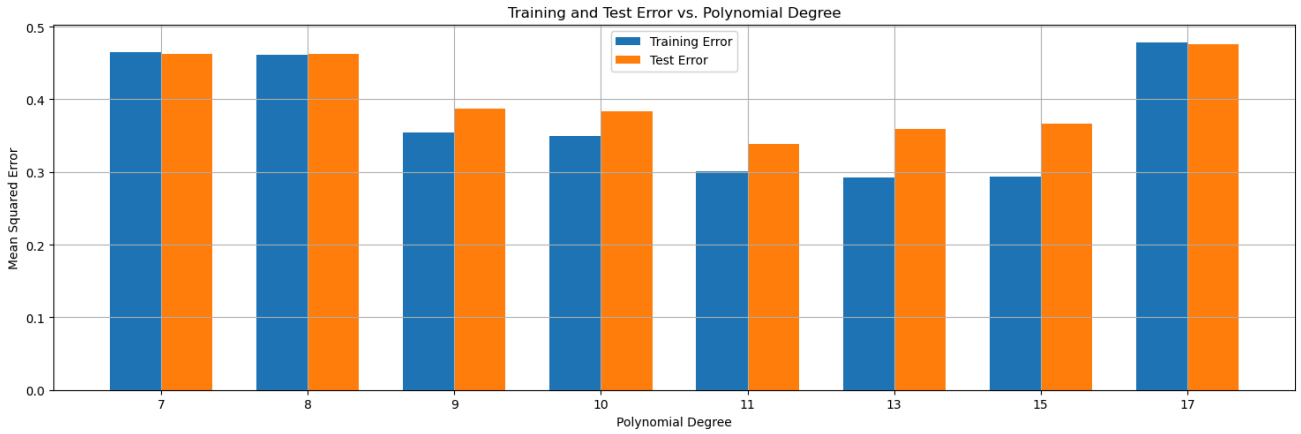
```
Error Comparison of Three Methods:
             Method  Train Error  Test Error
0     Least Squares     0.613402    0.595043
1  Gradient Descent     0.613402    0.595043
2     Newton Method     0.613402    0.595043
```



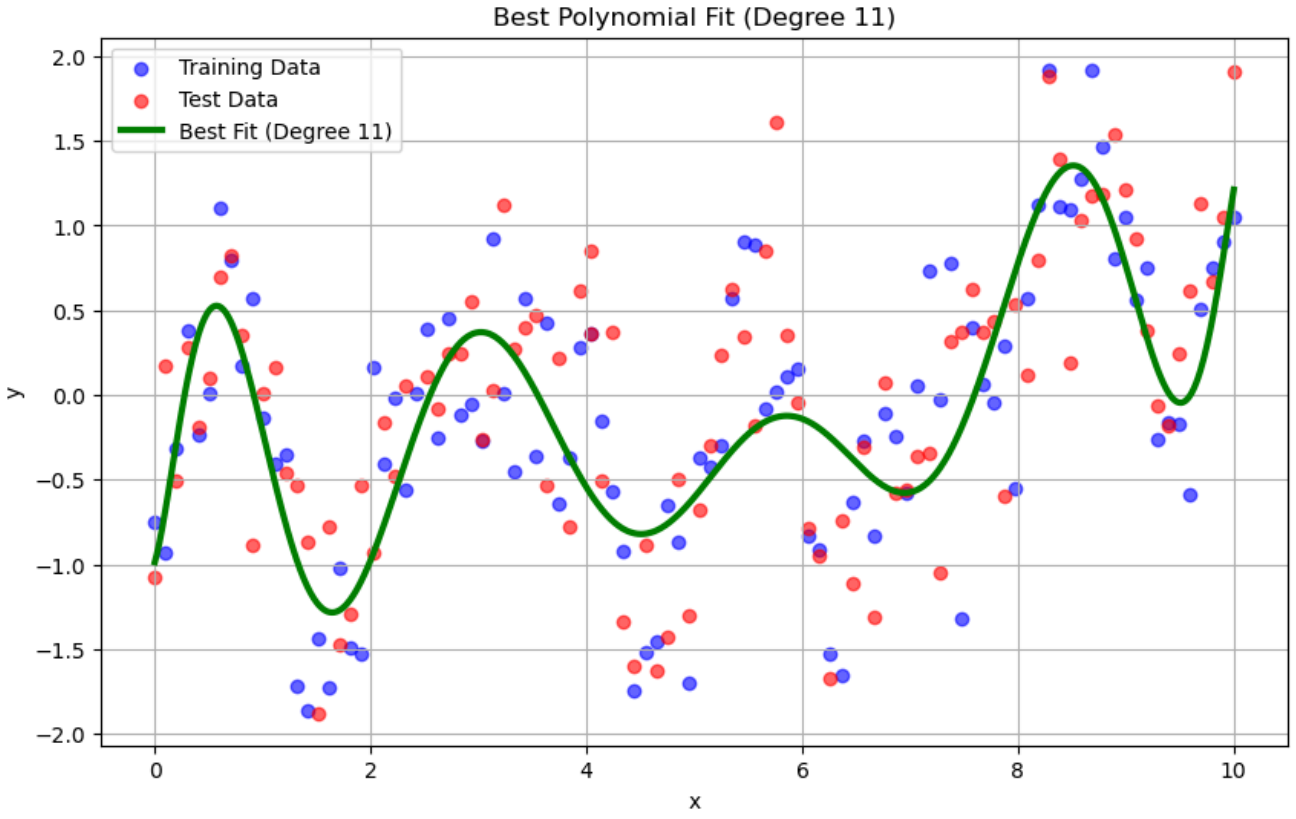

```
Polynomial Degree 7 - Training MSE: 0.465070, Test MSE: 0.463095
Polynomial Degree 8 - Training MSE: 0.461389, Test MSE: 0.461953
Polynomial Degree 9 - Training MSE: 0.354109, Test MSE: 0.387459
Polynomial Degree 10 - Training MSE: 0.349728, Test MSE: 0.383665
Polynomial Degree 11 - Training MSE: 0.301304, Test MSE: 0.339178
Polynomial Degree 13 - Training MSE: 0.292076, Test MSE: 0.358760
Polynomial Degree 15 - Training MSE: 0.294089, Test MSE: 0.366148
Polynomial Degree 17 - Training MSE: 0.478845, Test MSE: 0.475452
```

- **通过多项式拟合得到曲线和误差平方和如下**



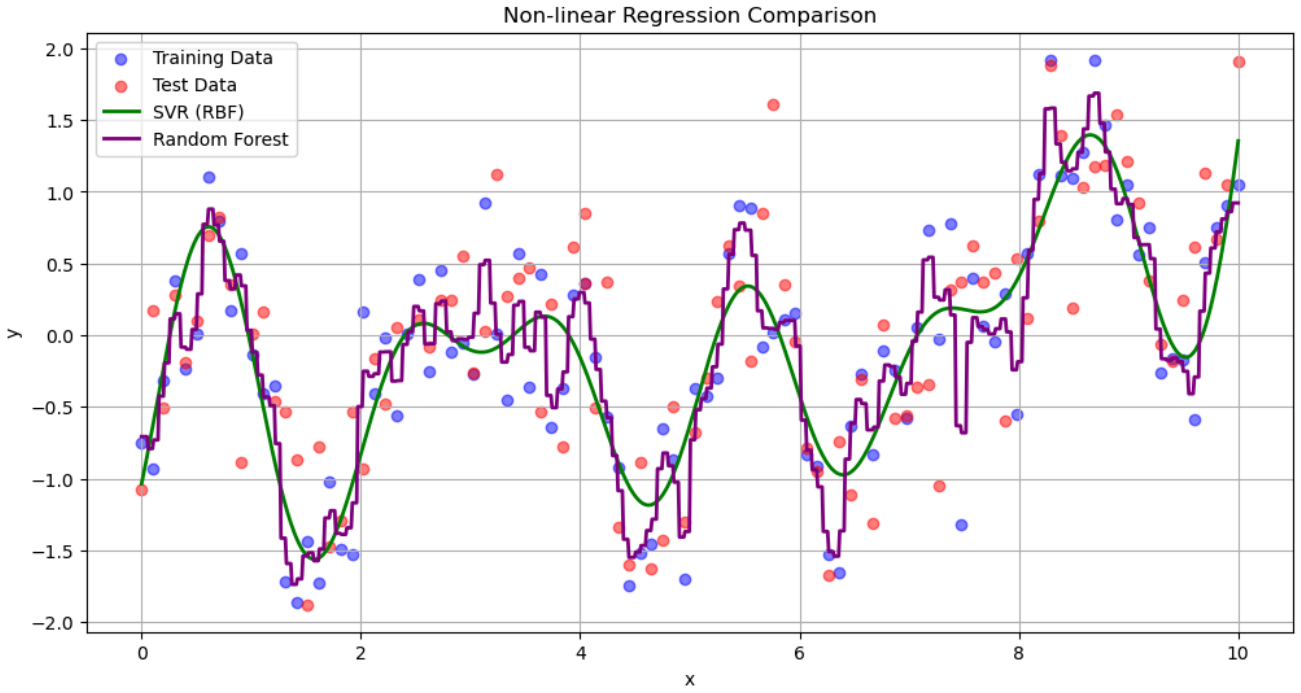


最终得到11项时在训练集和测试集上MSE最小，训练集误差为0.301，测试集误差为0.339



- 通过SVR回归和随机森林得到的训练结果如下

```
SVR (RBF) - Training MSE: 0.191155, Test MSE: 0.268754
Random Forest - Training MSE: 0.045767, Test MSE: 0.276250
```



Conclusions

三种线性拟合效果一致，都不能很好拟合数据，多项式拟合得到的MSE显著小于线性拟合，SVR曲线拟合得到的MSE最小，随机森林虽然在训练集上得到的MSE较小，但是在测试集上泛化效果不佳，拟合曲线离散化，只能显示出数据的变化趋势。