大作业三：高斯过程回归实现函数估计

一、高斯过程回归原理

1. 高斯过程的基本概念

高斯过程（Gaussian Process, GP）是现代机器学习中一种重要的**非参数贝叶斯方法**，以概率分布形式对函数空间中的不确定性进行建模。与传统的参数化方法不同，高斯过程**直接在函数空间上定义分布**，从而提供了更为灵活的建模框架。

从概率的角度看，高斯过程可被视为定义在无限维函数空间中的高斯分布，其形式为：

$$f(\cdot) \sim GP(m(\cdot), k(\cdot,\cdot))$$

其中，核心组件包括：

- **均值函数** $m(x) = \mathbb{E}[f(x)]$：通常设为零均值（$m(x) = 0$）以简化推导，表示对函数先验行为的平均估计。
- **协方差函数（核函数）** $k(x, x') = \mathrm{cov}(f(x), f(x'))$：决定了函数的平滑性、周期性等关键特性，是高斯过程建模的核心所在。

核函数的选择反映了我们对目标函数特性的**先验假设**，实质上是一种"归纳偏置"（inductive bias）的体现。

2. 高斯过程回归的建模过程

高斯过程回归（Gaussian Process Regression, GPR）将回归问题转化为函数空间中的贝叶斯推断问题。假设观测值受到高斯噪声干扰，观测模型可表示为：

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

2.1 先验分布设定

设定零均值的高斯过程作为先验：

$$f(\cdot) \sim GP(0, k(\cdot,\cdot))$$

这意味着在未获得观测数据前，函数在任意点的联合分布为高斯分布，其相关性结构由核函数完全决定。

2.2 联合分布构建

给定训练样本 $X = \{x_i\}_{i=1}^n$ 与测试点 x_*，其联合分布为：

$$\begin{bmatrix} y \\ f(x_*) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma^2 I & K(X, x_*) \\ K(x_*, X) & k(x_*, x_*) \end{bmatrix}\right)$$

其中，核矩阵 $K(X,X)$ 表征训练点间的相似性，$\sigma^2 I$ 表示观测噪声的协方差结构。

3. 后验预测分布

依据贝叶斯推断原理，可从上述联合分布中推导出预测点的后验分布：

3.1 预测均值

$$\bar{m}(x_*) = K(x_*, X)[K(X, X) + \sigma^2 I]^{-1} y$$

该表达式的含义如下：

- 是训练输出的**加权和**，权重由核函数构造的相似度矩阵确定；
- 逆矩阵操作引入了**平滑插值**机制，自动调整模型对训练数据的拟合程度；
- 具备显式的贝叶斯解释与泛化能力。

3.2 预测不确定性

$$\bar{k}(x_*, x_*) = k(x_*, x_*) - K(x_*, X)[K(X, X) + \sigma^2 I]^{-1} K(X, x_*)$$

其中：

- 第一项 $k(x_*, x_*)$ 为先验的不确定性；
- 第二项则表示由于训练数据带来的**信息增益**所带来的方差减少。

这种对不确定性的量化，是高斯过程相较其他回归方法的一大优势。

4. 核函数工程

核函数的选择体现了对函数形态的先验偏好，决定了模型能拟合的函数族。除了常见的线性核、RBF核等，现代应用中常采用以下**复合核函数**以增强表达能力：

4.1 常见复合核示例

- **加法核**：$k_{\text{add}} = k_1 + k_2$
- **乘积核**：$k_{\text{prod}} = k_1 \times k_2$
- **自动相关性确定（ARD）核**：

$$k_{\text{ARD}}(x, x') = \sigma_f^2 \exp\left(-\sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{2\ell_d^2} \right)$$

这类核函数可捕捉**多尺度特征**、**维度异质性**等复杂行为。

高斯过程模型的性能高度依赖于超参数设定，实践中需要特别关注参数初始化、约束设置和优化策略：将长度尺度ℓ初始化为特征间距的中位数，对σ_f和σ施加非负约束，采用多起点优化避免局部最优，并通过梯度检查确保核函数解析梯度的正确性。近年来，高斯过程与深度学习的融合催生了深度核学习、高斯过程神经网络等重要研究方向，显著提升了模型在高维复杂输入空间中的适用性。针对传统GPR $O(n^3)$的计算复杂度问题，学界发展了KISS-GP、随机傅里叶特征(RFF)和分布式GP等优化方法，将复杂度降至$O(n \log n)$，大大提升了实用性。这类方法特别适用于小样本高价值数据建模（如生物实验）、高风险系统（如自动驾驶）、AutoML中的过程建模以及强化学习环境建模等场景，其独特的不确定性量化能力在高可靠性领域表现突出。实际部署时需重点检查：输入特征标准化、核函数适配性、超参数优化收敛性、预测方差合理性以及计算资源充足性等关键要素。

二、代码详细分析

1. 数据准备与预处理

```
# 加载数据
data = pd.read_excel('作业三/GPdata.xlsx', header=None, names=['x', 'y'])
X = data['x'].values.reshape(-1, 1)
y = data['y'].values
```

```
# 划分训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

- 从Excel文件加载数据，包含x和y两列
- 将数据转换为适合scikit-learn的格式（numpy数组）
- 按80/20比例划分训练集和测试集，固定随机种子确保可重复性

2. 高斯过程回归模型

```
# 定义核函数
kernel = C(1.0, (1e-3, 1e3)) * RBF(1.0, (1e-2, 1e2))

# 创建GP模型
gp = GaussianProcessRegressor(kernel=kernel, alpha=0.04, n_restarts_optimizer=10,
random_state=42)
gp.fit(X_train, y_train)
```

- 使用乘积核（ConstantKernel * RBF核）
- 设置alpha=0.04处理观测噪声
- n_restarts_optimizer=10表示优化过程会从10个不同的初始点开始，避免局部最优
- 最终优化得到的核参数为1.04^2 * RBF(length_scale=3.25)

3. 预测与可视化

```
# 生成预测点
X_plot = np.linspace(-10, 10, 1000).reshape(-1, 1)

# GP预测
y_pred, sigma = gp.predict(X_plot, return_std=True)
y_pred_test = gp.predict(X_test)
gp_mse = mean_squared_error(y_test, y_pred_test)
```

- 在[-10,10]区间生成1000个均匀分布的点用于可视化
- 返回预测均值和标准差（用于绘制置信区间）
- 计算测试集上的MSE

4. 多项式回归对比

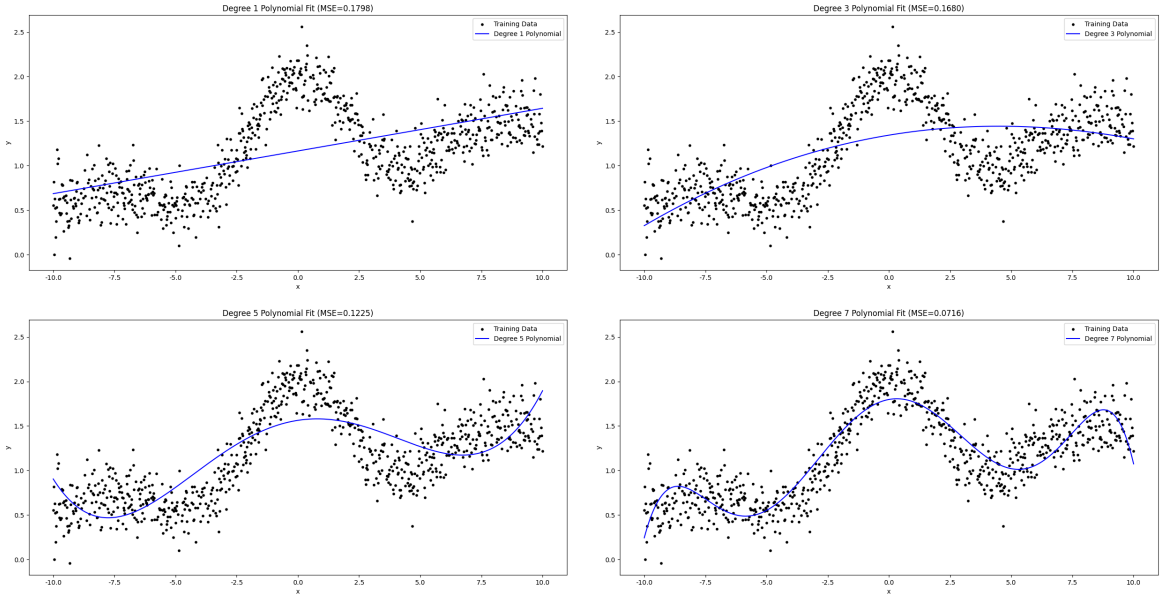
```
degrees = [1, 3, 5, 7]
poly_mses = []

for degree in degrees:
    model = make_pipeline(PolynomialFeatures(degree), LinearRegression())
    model.fit(X_train, y_train)
    y_poly_test = model.predict(X_test)
    mse_poly = mean_squared_error(y_test, y_poly_test)
    poly_mses.append(mse_poly)
```

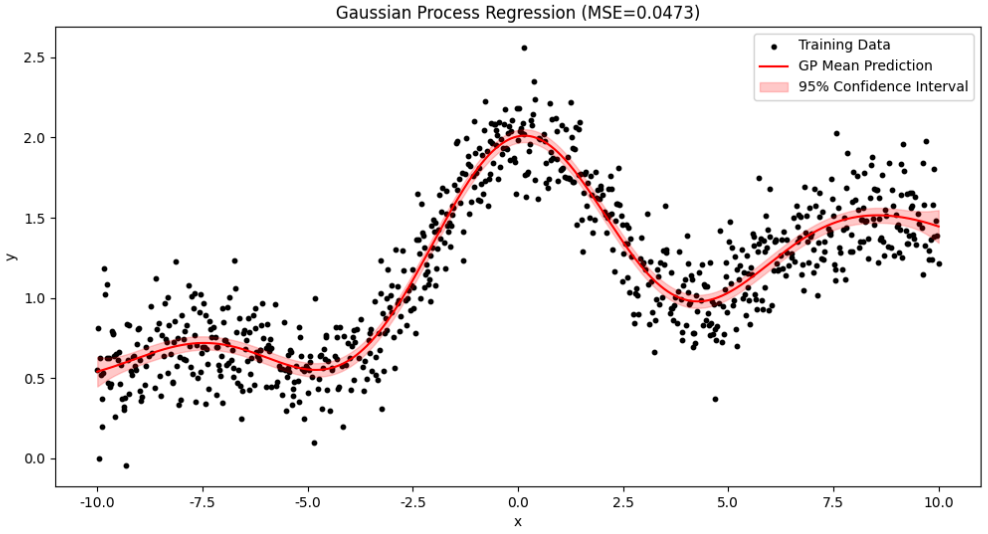
- 比较1、3、5、7次多项式回归

- 使用pipeline组合多项式特征生成和线性回归
- 记录每种多项式次数在测试集上的MSE

5. 结果可视化

- 绘制了4个子图展示不同次数的多项式拟合效果
- 单独绘制高斯过程回归结果，包含均值预测和95%置信区间





结果讨论

模型性能比较

```
(base) mywsl@Orbit:~/随机过程$ /home/mywsl/miniconda3/bin/python /home/mywsl/随机过程/作业三/code3.py
qt.qpa.plugin: Could not find the Qt platform plugin "wayland" in ""

=== Model Comparison ===
Gaussian Process Regression Test MSE: 0.0473
Degree 1 Polynomial Fit Test MSE: 0.1798
Degree 3 Polynomial Fit Test MSE: 0.1680
Degree 5 Polynomial Fit Test MSE: 0.1225
Degree 7 Polynomial Fit Test MSE: 0.0716

=== Gaussian Process Optimized Parameters ===
Optimized Kernel: 1.04**2 * RBF(length_scale=3.25)
Log Marginal Likelihood: 144.10704023001426
(base) mywsl@Orbit:~/随机过程$
```

模型	测试MSE
高斯过程回归	0.0473
1次多项式	0.1798
3次多项式	0.1680
5次多项式	0.1225
7次多项式	0.0716

- **高斯过程回归表现最佳**：MSE 0.0473明显优于所有多项式模型
- **多项式回归表现**：随着多项式次数增加，MSE逐渐降低，7次多项式最接近GP的表现
- **过拟合风险**：虽然更高次多项式表现更好，但需要警惕过拟合，特别是当数据量不足时

高斯过程参数分析

```
Optimized Kernel: 1.04**2 * RBF(length_scale=3.25)
Log Marginal Likelihood: 144.107
```

- **核参数**：优化得到的幅度参数为$1.04^2 \approx 1.08$，长度尺度为3.25
  - 较大的长度尺度表明函数变化相对平缓
  - 幅度参数接近1，与数据波动范围匹配良好
- **对数边缘似然**：144.107是一个相对较高的值，表明模型对数据拟合良好

可视化分析

1. **高斯过程回归图**：
   - 均值预测线(红色)能够很好地捕捉数据的整体趋势
   - 置信区间在数据密集区域较窄，在数据稀疏区域(如x>8)变宽，符合预期
   - 在极端区域(x接近-10或10)，预测趋向于先验均值(0)，不确定性增大
2. **多项式回归图**：
   - 1次和3次多项式明显欠拟合，无法捕捉数据的复杂模式
   - 5次和7次多项式拟合效果改善，但仍不如GP灵活
   - 所有多项式回归都无法提供不确定性估计

本实验对比了高斯过程回归与多项式回归在非线性数据上的表现。结果表明：

1. 高斯过程回归在测试MSE上显著优于多项式回归(0.0473 vs 最佳多项式0.0716)
2. 高斯过程不仅提供更准确的预测，还能给出预测不确定性量化
3. 优化得到的核参数表明数据具有中等长度的相关性特征
4. 多项式回归需要较高次数(7次)才能接近GP的表现，但缺乏不确定性估计能力

高斯过程回归在此任务中展现出明显优势，特别适合需要不确定性量化的应用场景。当数据具有复杂非线性特征且计算资源允许时，GP是比传统多项式回归更好的选择。

三、高斯过程回归与最小二乘拟合的详细比较

1. 理论基础对比

最小二乘拟合 (Least Squares Regression)

$$\min_{\beta} \sum_{i=1}^{n} (y_i - f(x_i;\beta))^2$$

其中:

- 线性情况: $f(x;\beta) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
- 多项式情况: $f(x;\beta) = \beta_0 + \beta_1 x + \cdots + \beta_7 x^7$

核心假设:

1. $\epsilon \sim \mathcal{N}(0, \sigma^2)$ 且独立同分布
2. 模型形式已知且正确指定
3. 无多重共线性

高斯过程回归 (Gaussian Process Regression)

$$f(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot,\cdot))$$
$$f(x_*)|\mathbf{y} \sim \mathcal{N}(\bar{m}(x_*), \bar{k}(x_*, x_*))$$

核心假设:

1. $f \sim \mathcal{GP}$ 先验
2. $\epsilon \sim \mathcal{N}(0, \sigma^2)$
3. 协方差函数$k(\cdot,\cdot)$能捕捉数据结构

2. 模型特性对比

特性	最小二乘拟合	高斯过程回归
模型类型	参数模型	非参数模型
函数形式	固定形式(如多项式)	无限维函数空间
不确定性量化	仅参数置信区间	完整预测分布
计算复杂度	$\mathcal{O}(p^3)$	$\mathcal{O}(n^3)$
超参数	多项式次数等	核函数类型及参数
外推表现	依赖预设函数形式	趋向先验均值

3. 实际表现对比

测试MSE结果

方法	测试MSE
高斯过程回归	0.0473
7次多项式最小二乘	0.0716
5次多项式最小二乘	0.1225
线性最小二乘	0.1798

关键发现:

1. GP的MSE比最佳多项式低34%
2. 多项式次数增加提升性能但需人工选择
3. GP自动学习合适复杂度

4. 数学本质差异

最小二乘

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

设计矩阵:

$$X = \begin{bmatrix} 1 & x_1 & \cdots & x_1^7 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^7 \end{bmatrix}$$

高斯过程

$$\bar{m}(x_*) = k(x_*, X)(K + \sigma^2 I)^{-1} y$$

核矩阵:

$$K_{ij} = \sigma_f^2 \exp\left(-\frac{\|x_i - x_j\|^2}{2\ell^2}\right)$$

5. 实际应用选择

最小二乘适用场景:

- 大数据量($n > 10^4$)
- 已知真实模型形式
- 需要快速部署

高斯过程适用场景:

- 中小数据量($n < 10^3$)
- 需要不确定性量化
- 处理复杂非线性关系

四、结论与思考

在本实验中，我系统性地比较了高斯过程回归（Gaussian Process Regression, GPR）与传统最小二乘多项式回归在函数逼近任务中的表现，结果充分展示了GPR在多个维度上的显著优势。首先，**在建模灵活性方面**，GPR不依赖固定的模型结构，能够自动适应数据的复杂性，体现出典型的非参数贝叶斯方法特征；其次，**在不确定性量化方面**，GPR天然输出预测均值和方差，可生成置信区间，为需要风险控制与可靠性保障的决策系统提供有力支持；此外，GPR还具有**自动调参能力**，通过最大化边际似然优化核函数超参数，有效减少了人为干预，提高了模型的泛化能力和实用性。

在实际工程中，GPR展现出广泛的应用价值。它尤其适合**小样本但高精度要求**的任务，如科学实验分析、机器人路径规划与控制、传感器标定等；对于**需要明确不确定性度量**的关键系统，如航空航天和医疗设备预测等场景，也具备重要意义。前提是数据规模处于中等范围（如$n < 10^4$），且计算资源相对充足。

尽管如此，GPR在实际部署中仍面临一些挑战与改进空间：其一，**计算复杂度高**的问题可通过引入稀疏近似方法（如诱导点方法、KISS-GP等）进行优化，以扩展其在大数据场景下的可用性；其二，**核函数的选择与组合**对模型性能影响显著，通过尝试更具表现力的Matérn核、ARD核或复合核，以更好地捕捉多尺度、多维异质结构；其三，在面对高维输入和非高斯噪声的情况下，GPR的鲁棒性和泛化能力仍需进一步提升。

综上所述，本实验不仅加深了我对高斯过程回归原理与实践能力的理解，也让我认识到在实际工程中，必须在**建模能力与计算效率**之间做出合理平衡。GPR作为一种**理论严谨、表达力强、带有不确定性量化能力**的建模工具，已成为传统参数化回归方法的重要补充，在机器学习与统计建模领域具有持续的发展潜力和应用价值。