

1. 线性回归模型的梯度下降法

步骤：

- (1) 加载数据：使用 pandas 加载波士顿房价数据集。
- (2) 预处理数据：将数据分为特征和目标变量，然后划分为训练集（80%）和测试集(20%)。
- (3) 归一化数据：对特征进行归一化处理。
- (4) 定义线性回归模型：实现一个线性回归模型，包括梯度下降算法(BGD 和 SGD)。
- (5) 训练模型：在原始数据+BGD，原始数据+SGD，归一化数据+BGD 和归一化数据+SGD 四种组合上训练模型。
- (6) 评估模型：计算每个模型在测试集上的均方误差（MSE）和决定系数（R2）。
- (7) 结果对比：以图表形式展示原始数据和归一化数据的预测结果。

方法：

归一化：使用 sklearn 中的 StandardScaler 方法进行特征缩放。

梯度下降：实现批量梯度下降（BGD）和随机梯度下降（SGD）。

BGD：在每次迭代中使用所有样本更新模型参数。

SGD：在每次迭代中只使用一个样本更新模型参数。

结果对比：

(1) 模型拟合图像

原始数据+BGD

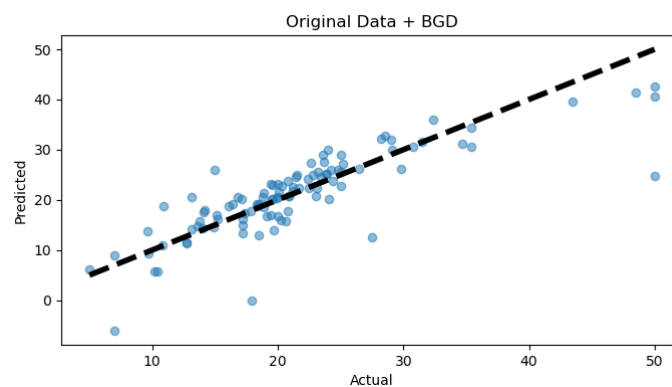


Figure 1 原始数据+BDG-1

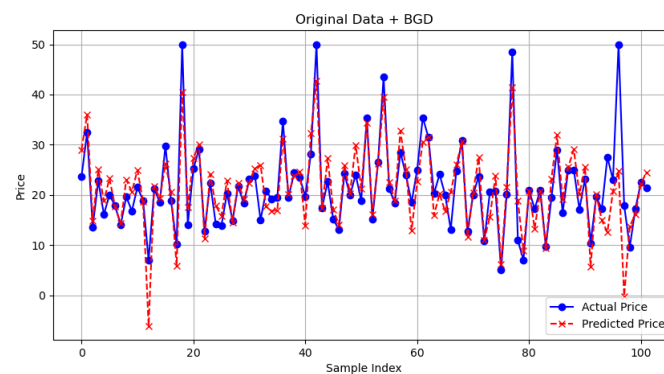


Figure 2 原始数据+BDG-2

原始数据+SGD

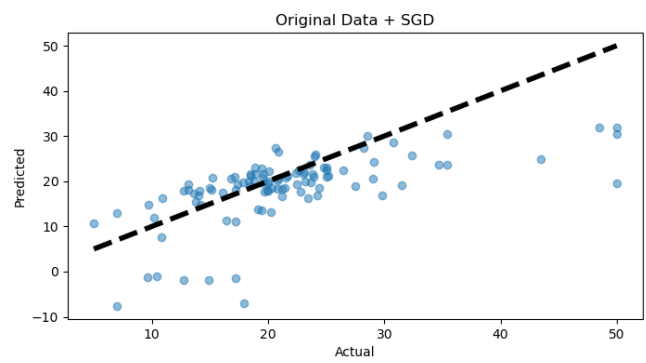


Figure 3 原始数据+SDG-1

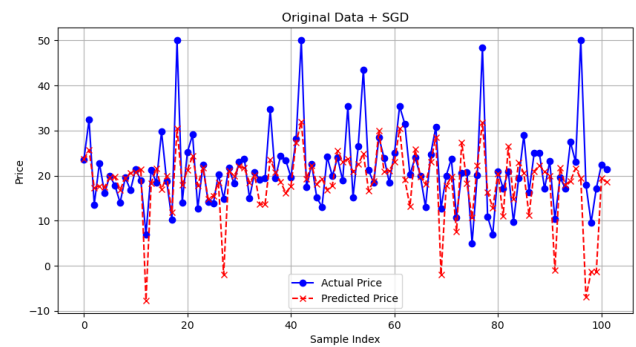


Figure 4 原始数据+SDG-2

归一化数据+BGD

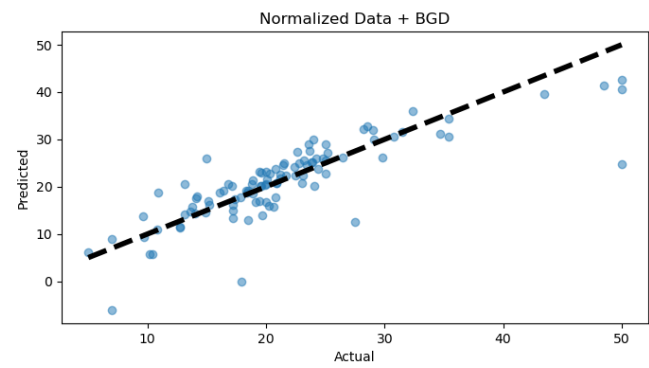


Figure 5 归一化数据+BDG-1

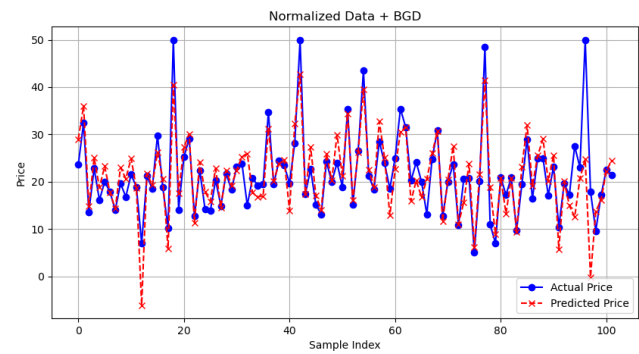


Figure 6 归一化数据+BDG-2

归一化数据+SGD

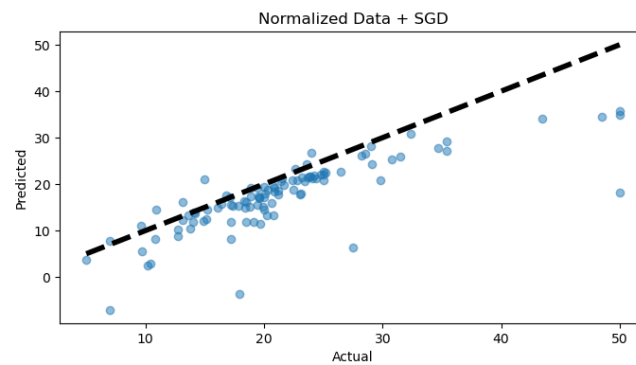


Figure 7 归一化数据+SGD-1

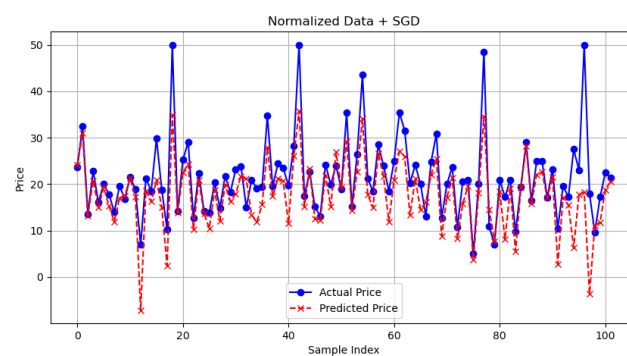


Figure 8 归一化数据+SGD-2

(2) 效果评估

均方误差 (MSE):

MSE for Original Data + BGD: 24.291119474973247

MSE for Normalized Data + BGD: 24.291119474973517

MSE for Original Data + SGD: 59.20273988728192

MSE for Normalized Data + SGD: 41.1536386373198

决定系数 (R2):

R2 for Original Data + BGD: 0.6687594935356357

R2 for Normalized Data + BGD: 0.668759493535632

R2 for Original Data + SGD: 0.1926948626412417

R2 for Normalized Data + SGD: 0.4388174608781691

结论:

- (1) 归一化数据对于 SGD 方法有显著的积极影响, 降低了 MSE 并提高了 R2 得分。
- (2) BGD 方法在原始数据和归一化数据上都表现更好, 尤其是在原始数据上, 其性能明显优于 SGD 方法。
- (3) SGD 方法在归一化数据上的性能提升表明, 归一化可能有助于 SGD 方法更快地收敛到更好的解。

2. 岭回归模型和 LASSO 回归模型

步骤：

- (1) 数据准备：将数据分为特征和目标变量，划分数据为训练集和测试集，对训练集和测试集的特征进行归一化处理。
- (2) 模型训练：使用归一化后的数据训练线性回归、岭回归和 LASSO 回归模型，岭回归和 LASSO 回归模型使用 SGD 优化器进行训练。
- (3) 模型评估：使用测试集评估每个模型的性能，计算每个模型的性能 MSE 和 R2 得分。
- (4) 系数分析：对比分析线性回归、岭回归和 LASSO 回归模型的系数。

方法：

岭回归：添加 L2 正则化项。

LASSO 回归：添加 L1 正则化项。

结果对比：

- (1) 模型拟合图像

岭回归+SGD

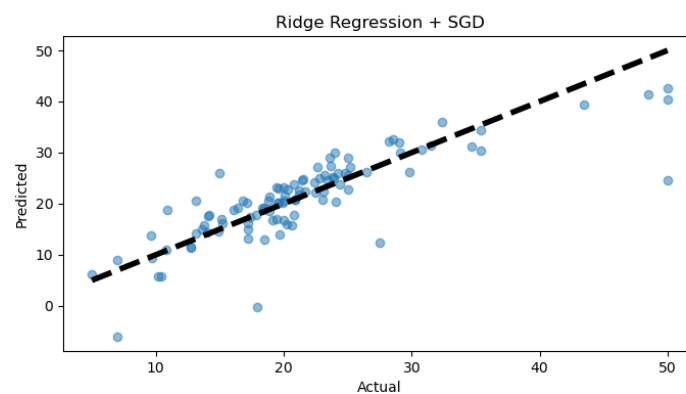


Figure 9 岭回归-1

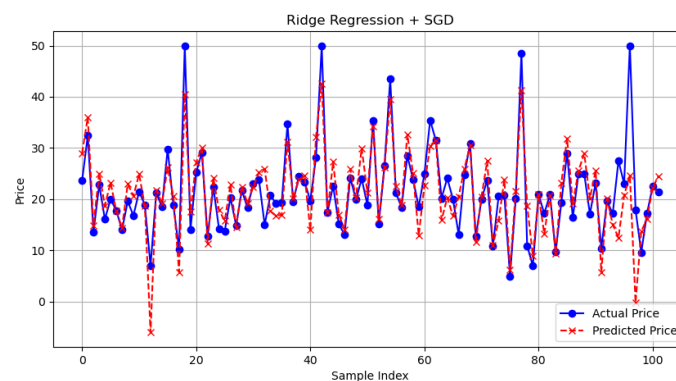


Figure 10 岭回归-2

LASSO 回归+SGD

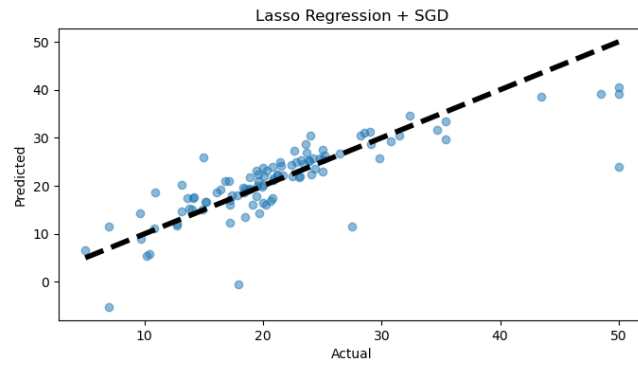


Figure 11 LASSO 回归-1

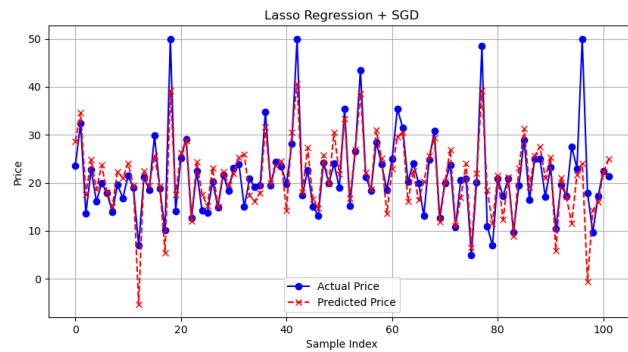


Figure 12 LASSO 回归-2

(2) 效果评估

均方误差 (MSE):

MSE for Ridge Regression + SGD: 24.31290

MSE for Lasso Regression + SGD: 25.65674

决定系数 (R2):

R2 for Ridge Regression + SGD: 0.66846

R2 for Lasso Regression + SGD: 0.65014

以上六种模型的 MSE 和 R2 比较:

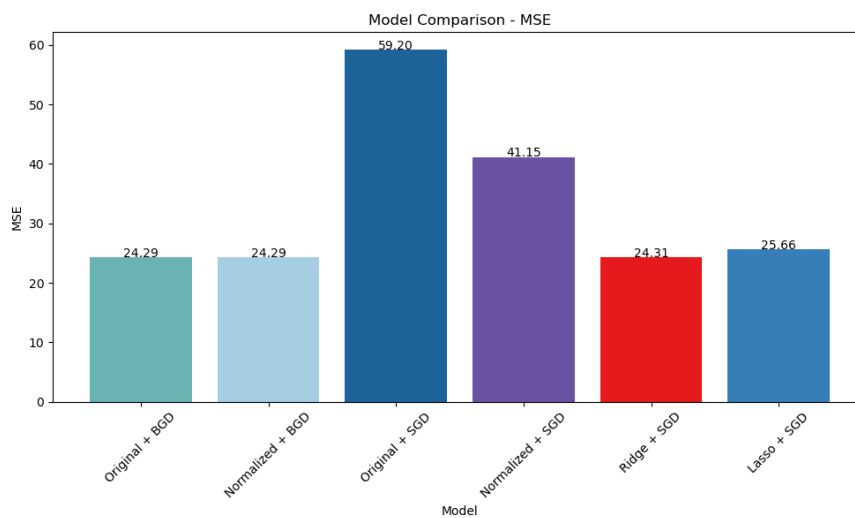


Figure 13 MSE

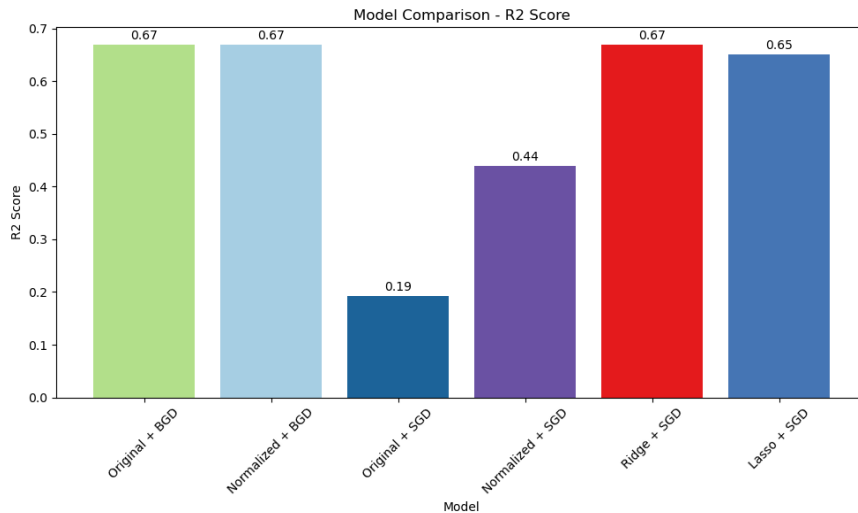


Figure 14 R2

(3) 系数比较

模型/系数						
Original BGD	-1.13055924e-01	3.01104641e-02	4.03807204e-02	2.78443820e+00	-1.72026334e+01	4.43883520e+00
-6.29636221e-03	-1.44786537e+00	2.62429736e-01	-1.06467863e-02	-9.15456240e-01	1.23513347e-02	-5.08571424e-01
Normalized BGD	-1.00213533	0.69626862	0.27806485	0.7187384	-2.0223194	3.14523956
-0.17604788	-3.0819076	2.25140666	-1.76701378	-2.03775151	1.12956831	-3.61165842
Original SGD	-0.10150258	0.10734008	-0.0822197	0.00976258	0.00324712	0.12130291
0.08581505	0.0276876	0.00446581	-0.00586322	0.07229635	0.04942595	-0.34381471
Normalized SGD	-0.65873366	0.14756146	-0.43237766	0.95316884	-0.55087258	3.34749208
-0.10449068	-1.23092844	0.1840841	-0.41477694	-1.63330497	0.93492954	-2.90877966
Ridge SGD	-0.99218679	0.6777488	0.2522143	0.72248078	-1.99083465	3.15157218
-0.17726162	-3.04502895	2.17324941	-1.69555879	-2.02783351	1.127197	-3.59897667
Lasso SGD	-0.71836455	0.25962714	-0.	0.69822096	-1.56814243	3.27150693
-0.	-2.28444944	0.67193802	-0.3566537	-1.89333519	1.03136581	-3.60941047

结论：

MSE：

- (1) 线性回归模型（无论是 BGD 还是 SGD）在 MSE 上表现相似，表明归一化对 BGD 方法的 MSE 影响不大。
- (2) 岭回归和 LASSO 回归的 MSE 低于原始数据的 SGD 方法，表明正则化有助于提高预测精度。
- (3) 岭回归的 MSE 与归一化数据的 BGD 相似，而 LASSO 回归的 MSE 略高。

R2：

- (1) 线性回归模型（无论是 BGD 还是 SGD）在 R2 得分上表现相似，表明归一化对 BGD 方法的 R^2 得分影响不大。
- (2) 岭回归和 LASSO 回归的 R2 得分高于原始数据的 SGD 方法，表明正则化有助于提高模型的解释能力。
- (3) 岭回归的 R^2 得分略高于归一化数据的 BGD，而 LASSO 回归的 R2 得分略低。

系数：

- (1) 线性回归的原始数据和归一化数据的系数相似，表明归一化对系数的影响不大。
- (2) 岭回归的系数与线性回归的系数相似，但可能略小，这表明 L2 正则化有助于减少过拟合。
- (3) LASSO 回归的系数中有一个为零，表明 LASSO 回归通过将不重要的特征的系数设置为零来进行特征选择。

附：代码地址 <https://github.com/JLU-KDZ/MachineLearning3.git>