



浙江工业大学

人工智能原理及应用 实验报告

实验名称: 线性回归-房价预测

学号: 202205240220

姓名: 潘家航

专业班级: 自动化 2304

学院: 信息工程学院

指导教师: 付明磊

目录

一、实验目的.....	3
二、实验设备.....	3
三、实验内容.....	3
四、实验过程.....	3
五、实验结果分析.....	3
六、实验小结.....	4
七、其它.....	7

一、实验目的

1. 掌握线性回归模型的原理与实现流程；
2. 使用 PaddlePaddle 构建简单的回归模型对波士顿房价数据进行预测；
3. 理解假设函数、损失函数（MSE）与优化算法（SGD）的作用；
4. 学会评估回归模型的预测性能并可视化结果。

二、实验设备

1. 硬件设备： 游侠 G15。
2. 软件环境： 主流浏览器（如 Chrome、Edge）、代码编辑器 pycharm、Python 编程环境。
3. 开发平台： 百度 AI 开放平台（ai.baidu.com）的 EasyDL 产品。

三、实验内容

本实验使用 PaddlePaddle 的 dygraph 模式构建一个简单的线性回归模型（单层全连接，输出维度=1），对 UCI 波士顿房价数据集（13 个特征）进行训练与预测，目标为预测房价中位数。

四、实验过程

1. 数据准备：使用 `paddle.dataset.uci_housing.train()/test()` 加载数据，对数据进行 `shuffle` 与 `batch`；
2. 网络配置：定义 `Regressor` 类，使用一层 `Linear(input_dim=13, output_dim=1)`；
3. 训练流程：在 `dygraph.guard()` 下进行训练，使用 `SGD` 优化器，学习率 0.01，训练轮次 `EPOCH_NUM=5`；

4. 评估与预测：保存模型后加载进行测试，计算 `MSE` 作为评价指标，并对若干测试样本进行预测比较。

五、实验结果分析

训练过程中的记录损失值（节录，来自你的日志输出）：

- 记录点 1: Avg Cost = 818.36414
- 记录点 2: Avg Cost = 381.62445
- 记录点 3: Avg Cost = 556.19336
- 记录点 4: Avg Cost = 226.93808
- 记录点 5: Avg Cost = 179.14398
- 记录点 6: Avg Cost = 337.02646
- 记录点 7: Avg Cost = 117.68654
- 记录点 8: Avg Cost = 195.28535
- 记录点 9: Avg Cost = 9.21903
- 记录点 10: Avg Cost = 114.44053
- 记录点 11: Avg Cost = 122.28262
- 记录点 12: Avg Cost = 49.79018
- 记录点 13: Avg Cost = 81.98496
- 记录点 14: Avg Cost = 51.08876
- 记录点 15: Avg Cost = 7.97365

模型在测试集上的平均测试损失（来自你提供的日志）: Test, Cost: 12.31293

在一次具体的预测示例中，模型对 10 个样本的预测结果与真实值如下（用户提供）：

样本编号	真实值 (ground truth)	预测值 (predicted)
0	8.50	13.7515
1	5.00	14.0201
2	11.90	13.2920
3	27.90	14.9062
4	17.20	13.8965
5	27.50	14.7681
6	15.00	14.7362
7	17.20	14.2682
8	17.90	12.1298
9	16.30	13.9853

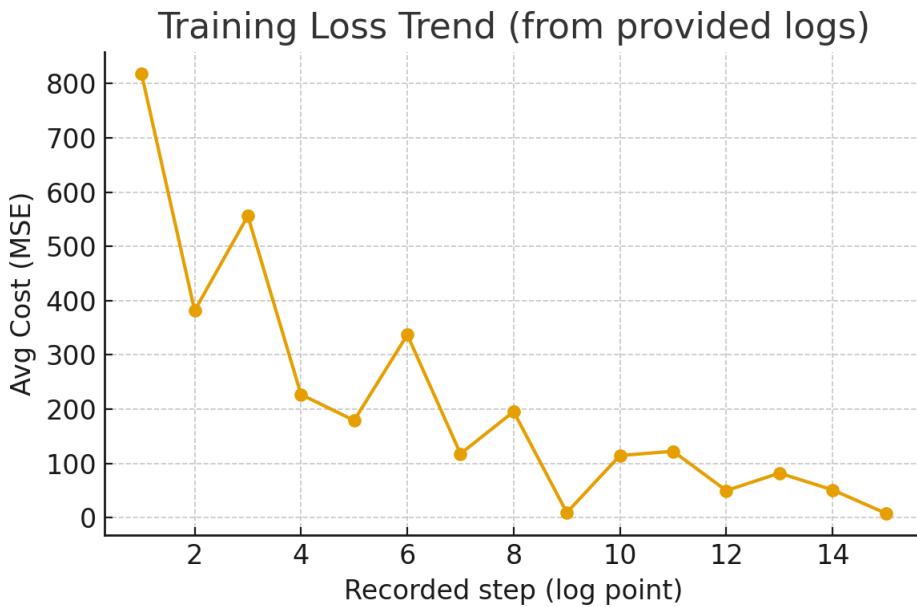


图 1：训练损失变化曲线（基于日志记录点）

本次预测的平均误差（用户提供）: 50.00491 (注意: 该值看起来像是平均平方误差或
累计平方误差, 应与 Test Cost 指标进行核对)。

从训练损失曲线和预测对比图可见:

- 训练过程中损失总体呈下降趋势, 但在中期存在波动, 最后几次记录点损失显著下降到个位数, 可能是由于某些 batch 的样本使得平均损失极低;
- 在预测结果中, 许多预测值集中在 12~15 范围, 但真实值分布更宽 (5 到 27.9),
说明模型可能欠拟合或数据/标签标准化导致尺度差

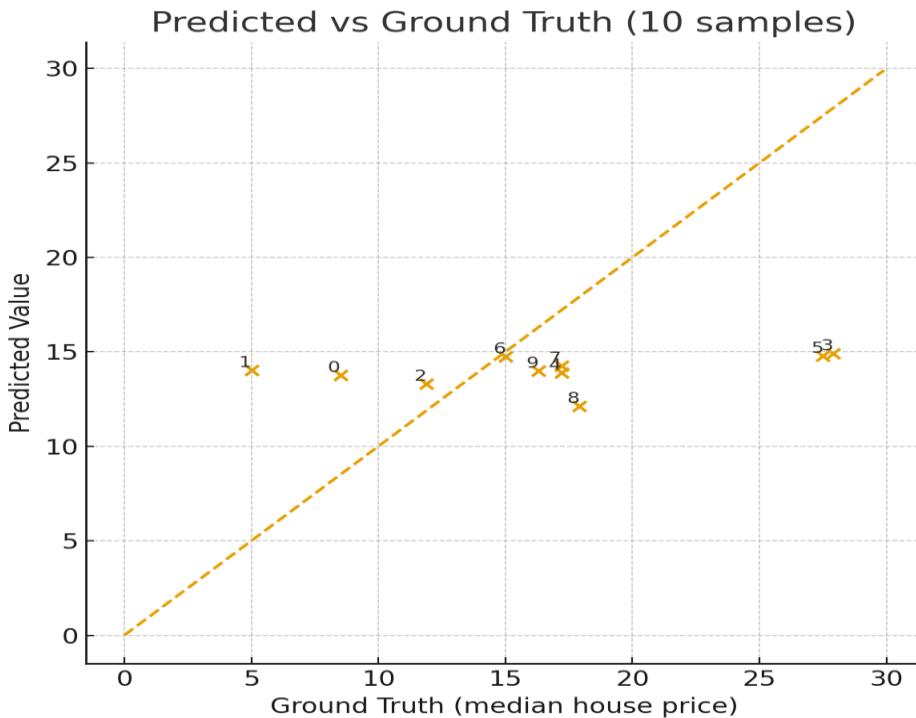


图 2：预测值与真实值对比（10 个样本）

- 建议核实数据预处理（标准化/归一化）与标签缩放是否一致，以及检查是否存在数据泄漏或训练/测试集分布差异。

六、实验小结

1. 成功实现了基于 PaddlePaddle 的线性回归模型训练流程，并能对测试样本进行预测；
2. 模型训练过程中损失总体下降，测试集上得到一定性能，但预测误差较大，说明需要进一步调参或改进模型；
3. 后续改进方向：调整学习率、增加训练轮次、尝试不同优化器（Adam）、加入正则化或更复杂的模型(多项式回归、神经网络)，以及更仔细的数据预处理(归一化/去离群点)。

七、其它

保存训练时每个 batch 的损失到文件，便于绘制更细致的损失曲线；记录随机种子以保证实验可复现；对预测结果计算更多指标（MAE, RMSE, R^2）以全面评估模型。