**2025年春季学期《机器学习》实验报告**

**（第一次实验）**

班级： 230617 学号： 23373179 姓名： 周子强

一、实验过程中，是否对输入数据进行了归一化或标准化处理？试说明这两种方法的区别，并分析为什么线性回归模型可能对特征的尺度敏感。

**（1）标准化的选取**：对输入数据的X\_train和X\_test使用StandardScaler进行标准化。

**（2）归一化和标准化的区别**：

1.归一化（MinMaxScaler）是将数据缩放到一个固定范围（通常是[0,1]或[-1,1]），公式为。因为依赖最大值和最小值，所以对异常值敏感，适用于数据分布不固定或需要严格限定范围的场景如图像处理。

2.标准化（StandardScaler）是将数据转换为均值为0、标准差为1的分布，公式为。保留了异常值的影响但减弱其主导作用，适用于大多数机器学习模型。

**（3）为什么线性回归模型可能对特征的尺度敏感**：线性回归模型的目标是找到特征权重，使得预测值和真实值之间的误差最小化。

1.如果特征尺度差异大，梯度下降时不同方向的参数更新速度不一致，导致模型收敛速度变慢或无法收敛。

2.此外，正则化惩罚对不同尺度特征也存在差异。

3.计算解析解时，尺度差异大的特征句子可能导致协方差矩阵的条件数过高，使得矩阵求逆不稳定。

因此，使用标准化统一特征尺度，使模型训练更加高效可靠。

二、对于线性回归目标函数，推导给出参数的解析解形式，并思考对于实验所使用的数据集而言，采用标准方程组法求解参数相较于梯度下降法有何优势或劣势。

**（1）参数解析解的推导**：

线性回归模型的目标函数是最小化平方误差损失：

表示为矩阵形式：

其中是的设计矩阵（为特征维度），是的目标向量，是的待求参数向量。

1. 展开目标函数：
2. 对求导：利用矩阵微分规则，求导并令导数为零
3. 解得解析解：整理方程为  
   若可逆，则参数解为

**（2）标准方程组法和梯度下降法的优劣**：

1.标准方程组法：

优势：

1. 直接通过矩阵运算得到全局最优解，是一个精确解，无需迭代。不需要调参（学习率、迭代次数等超参数）。
2. 特征维度较小时（如本实验，时间、历史收盘价、波动），计算速度快。

劣势：

1. 如果接近奇异矩阵（特征之间存在强相关性），会求逆失败或结果不正确。
2. 如果特征维度极大，计算复杂度很高，效率急剧下降。

2.梯度下降法：

优势：

1. 每次迭代仅需部分数据（小批量梯度下降），内存占用低。
2. 无需矩阵求逆，适合存在共线性的场景。
3. 可逐步更新参数，适应数据动态变化（实时股票数据流）。

劣势：

1. 需要调参，参数选择不当会导致收敛慢或震荡。
2. 需要迭代足够次数才能逼近最优解，是一个近似解，对初始值敏感。
3. 对高维小样本数据效率低于解析解。

三、实验中使用的评估指标（如均方误差MSE、均方根误差RMSE、决定系数）分别反映了模型的哪些性能？如果某次实验的值为负，可能是什么原因导致的？

**（1）评估指标**：

1.均方误差：反映预测值与真实值的平均平方偏差，直接衡量模型的预测误差。

2.均方根误差与MSE类似，但量纲与因变量一致，数值更直观。

3.决定系数：反映模型对数据方差的解释能力。说明模型完美拟合数据，说明模型与直接预测均值效果相同，说明模型预测效果比直接用均值预测差。

**（2）值为负的可能原因**：

1.模型拟合极差：如果时间序列数据没有合理划分数据集，可能会过拟合/欠拟合。

2.数据预处理错误：测试集的信息（未来股价）泄漏到训练集会导致模型在测试集上表现差；如果没有对训练集和测试集统一使用训练集的均值和标准差，可能会破坏数据分布的一致性。

3.基准模型选择问题：定义混淆，计算时公式写错。

四、在实验中，如果原始数据中存在非线性关系（如特征与目标变量呈二次函数关系），直接使用线性回归会导致模型性能不佳，思考通过何种方式能够更好的拟合特征与目标变量之间的关系。

1.添加多项式特征：如果原始特征为，可以添加，等。经过标准化后即可参与线性回归拟合。

2.非线性特征变换：通过对数、指数、平方根等变换转换为线性关系。

3.引入交互项：通过特征间的乘积项捕捉交互作用。

4.使用非线性模型：决策树/随机森林、支持向量回归、神经网络等。

5.局部加权回归：对数据的不同区域分别拟合简单模型，适应局部非线性变化。

五、你对本次实验课程内容、课程形式、实践平台使用等方面有哪些意见及改进建议？

内容、形式都挺好的，实验平台本身不太行。并且希望这学期实验课能涉及尽可能多的机器学习模型，因为感觉这个真的很重要但没怎么系统地学过。