2025年春季学期《机器学习》实验报告

班级: 230617 学号: 23371007 姓名: 刘嘉明

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. 本次实验对输入数据进行了标准化处理。这两种方式有一定的区别，前者的缩放范围固定在[0,1],而且容易受到异常最大值,最小值的影响,后者则是把数据调整为均值为0,标准差为1的统计数据,不容易受极端值影响
2. 如果特征A的范围在[0,1],而特征B的范围在[0,1000],这样在加权的时候,导致权重的变化或缓或快,不利于找到稳定的结果

二、对于线性回归目标函数$J(\mathbf{w}) = \sum_{i = 1}^{N} (\mathbf{w}^{T}\mathbf{x}*{i} - y{i})^{2}$，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。

矩阵形式： $$ J(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^\mathrm{T}(\mathbf{X}\mathbf{w} - \mathbf{y}) $$ 其中：

- $\mathbf{X} \in \mathbb{R}^{N \times D}$ 为设计矩阵（每行为一个样本）
- $\mathbf{y} \in \mathbb{R}^N$ 为目标向量
- $\mathbf{w} \in \mathbb{R}^D$ 为参数向量

求导过程： $$ \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 2\mathbf{X}^\mathrm{T}(\mathbf{X}\mathbf{w} - \mathbf{y}) = 0 $$

解析解： $$ \mathbf{w} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{y} $$

优势

1. 可以直接得到解析解
2. 更加精确

劣势

1. 内存消耗大：需计算$(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}$,时间复杂度高
2. 数值不稳定：若$\mathbf{X}^\mathrm{T}\mathbf{X}$接近奇异矩阵时求逆困难

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R^2的值为负，可能是什么原因导致的？

1. MSE: $\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ 反映预测值与真实值的平均平方误差,体现模型的误差程度
2. RMSE: $\sqrt{\text{MSE}}$也反映了预测值与真实值的平均平方误差,体现模型的误差程度,但是与MSE相比,它的单位与股价一致
3. 决定系数R^2: $1 - \frac{\text{MSE}}{\text{Var}(y)}$反映出构建的模型解释目标变量变异性的比例,越接近于1,说明效果越好. 如果R^2为负数,可能代表模型欠拟合,也就是模型的构建不合理,同时,也有可能是模型过拟

> 合,体现在训练集R^2接近于1,测试集R^2为负数,这是因为模型过于符合训练集,以至于完全把训练集的特征生搬硬套的自己的模型中,不具有普适性,不具有灵活性,因此面对新的数据时,表现很不好

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。

使用线性回归的优点在于简便,但是不精确,因此可以考虑使用多项式回归

```
poly = PolynomialFeatures(degree=2, include_bias=False)
X_train_poly = poly.fit_transform(X_train_scaled)
X_test_poly = poly.transform(X_test_scaled)
```

这里采用二次函数的回归形式,我将线性回归改为多项式回归后,训练集和的R^2更加接近于1,训练集R^2由原来的0.9996360560429334改进为0.9996404412620434

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?

我认为本次实验的内容很丰富,充实,而且构建股价模型这一任务具有现实意义,能够很好的调动学生的积极性. 之后希望能提升实验平台的稳定性.