

人工智能实验报告

实验名称： 逻辑回归

|  |  |  |  |
| --- | --- | --- | --- |
| 学生学号 | 202110120401 | 202110120420 | 202110120408 |
| 学生姓名 | 刘广润 | 赵威 | 喻慧超 |
| 分工 | 逻辑回归、硬阈值线性分类实现,测试算法性能，撰写报告 | 算法速度测试，绘制表并进行对比，撰写报告 | 优化算法、撰写报告 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **实验评价结果** | | | | |
| **占比** | 10% | 40% | 50% | 总分 |
| **评分标准** | 报告的规范性。报告中的术语、格式、图表、数据、公式、标注及参考文献是否符合规范要求。 | 报告的严谨性。结构是否严谨，论述的层次是否清晰，逻辑是否合理，语言是否准确。 | 实验的充分性。实验是否完备，数据是否合理，是否有创新性成果或独立见解。 |
| 评分 |  |  |  |  |

报告提交时间： 2023 年 12 月 23 日

# 试验任务

本报告对硬阈值线性分类和Logistic线性分类两种方法在乳腺癌数据集上的应用进行了详细的比较和分析。通过对损失曲线、准确度、混淆矩阵的观察，以及参数分析，我们旨在揭示这两种方法在乳腺癌诊断上的有效性和特点。

# 实验平台

Windows操作系统，Python编译环境，numpy, sklearn,pytorch,seaborn,matplotlib

等软件包。

# 实现与损失曲线的收敛情况

在这一部分，我们首先介绍了硬阈值线性分类和Logistic线性分类的基本原理与实现，然后展示并比较了两种方法在训练集上的损失收敛情况。通过绘制损失曲线图，我们可以观察到模型在训练过程中的学习效率和稳定性。

## 方法介绍

### 硬阈值线性分类

硬阈值线性分类是一种简单的线性分类方法，其核心思想是基于线性方程的输出对数据进行分类。这种方法通过构造一个线性方程来描述特征和目标变量之间的关系，其中w表示权重，b表示偏置项。对于给定的输入特征x，线性方程产生一个实数值输出。

硬阈值线性分类中，我们设置一个阈值。如果线性方程的输出大于或等于这个阈值，数据被分类为一个类别；如果小于这个阈值，数据被分类为另一个类别。这种方法适用于二分类问题，但它的主要局限性在于对数据的线性可分性的假设。

### Logistic回归

Logistic回归是一种广泛用于二分类问题的线性分类模型。与硬阈值线性分类类似，Logistic回归模型使用一个线性方程z = wx + b来预测目标变量，但它进一步将线性输出通过一个Sigmoid函数转换为概率值。

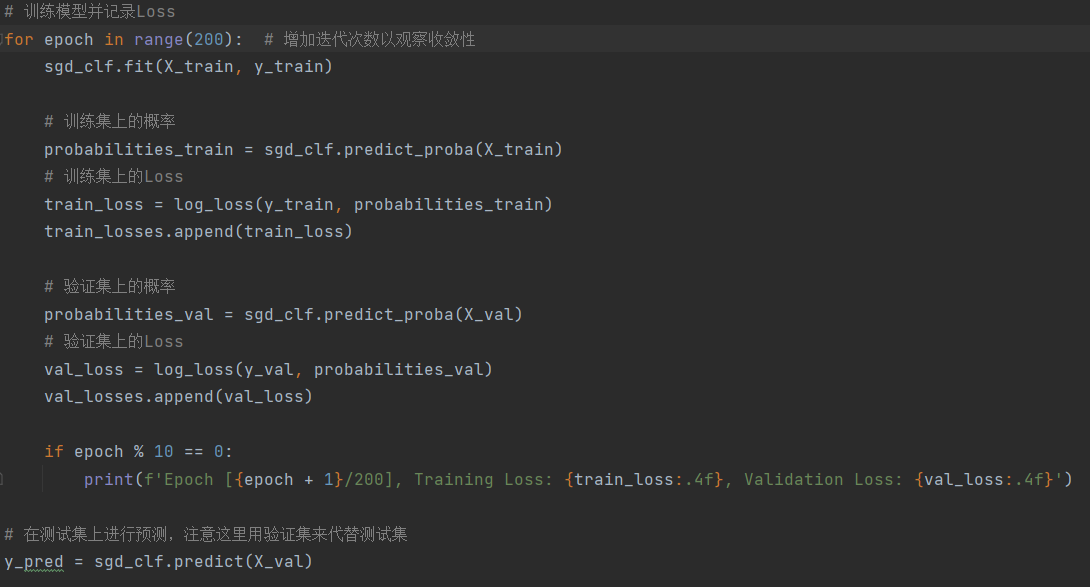
Sigmoid函数的输出范围在 0 到 1 之间，表示属于阳性类别的概率。如果Sigmoid函数的输出大于0.5，数据被预测为阳性类别，否则为阴性类别。

Logistic回归的优势在于它不仅能够进行分类，还能提供概率解释，这对于理解预测的不确定性非常有帮助。此外，它通过Sigmoid函数解决了硬阈值线性分类可能过于严格的二元分类限制。

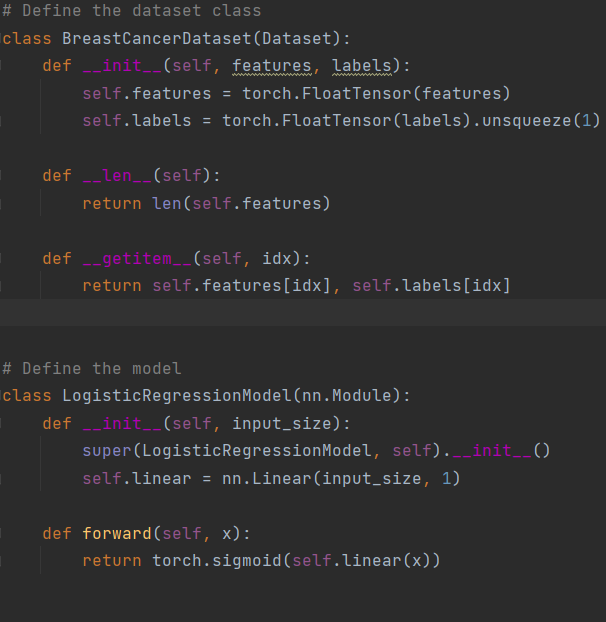
## 方法实现

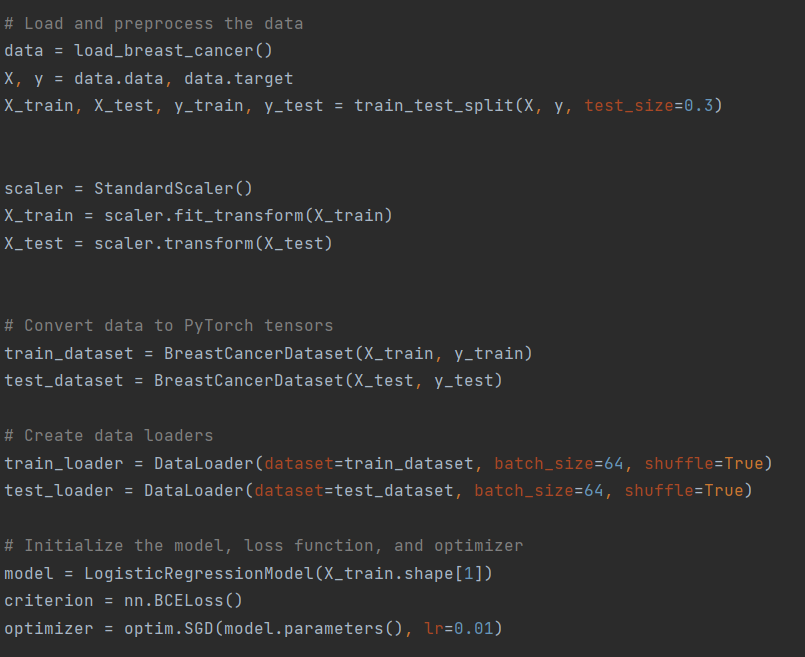
### 硬阈值线性分类

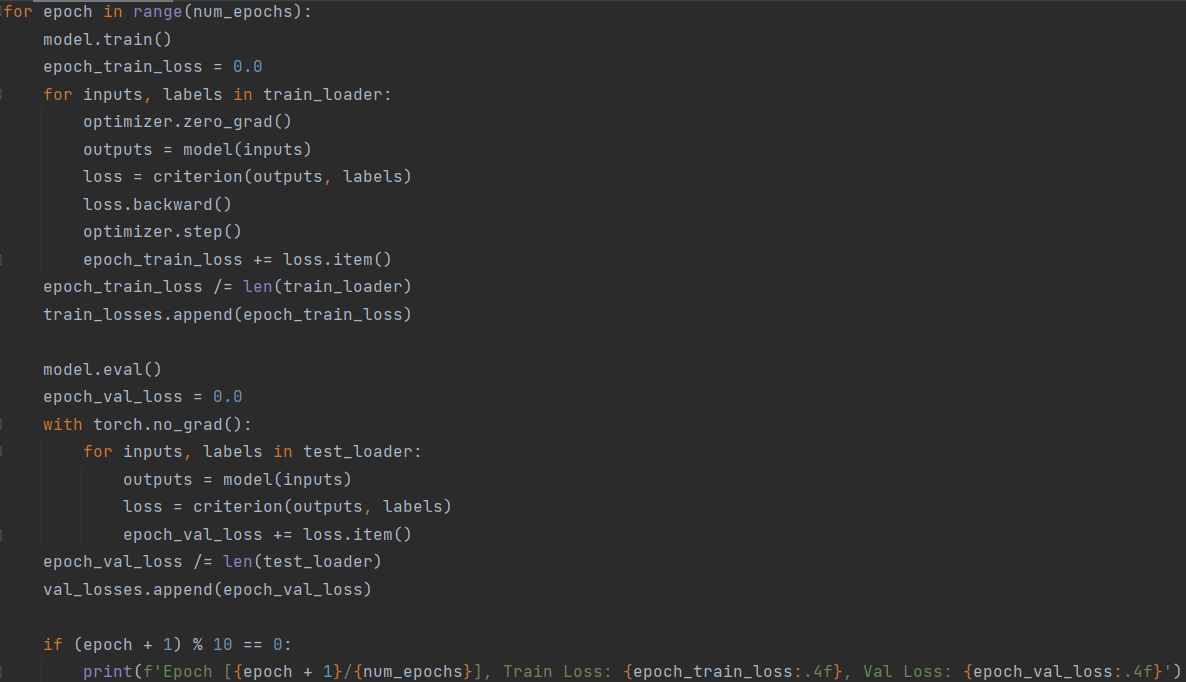




### Logistic回归

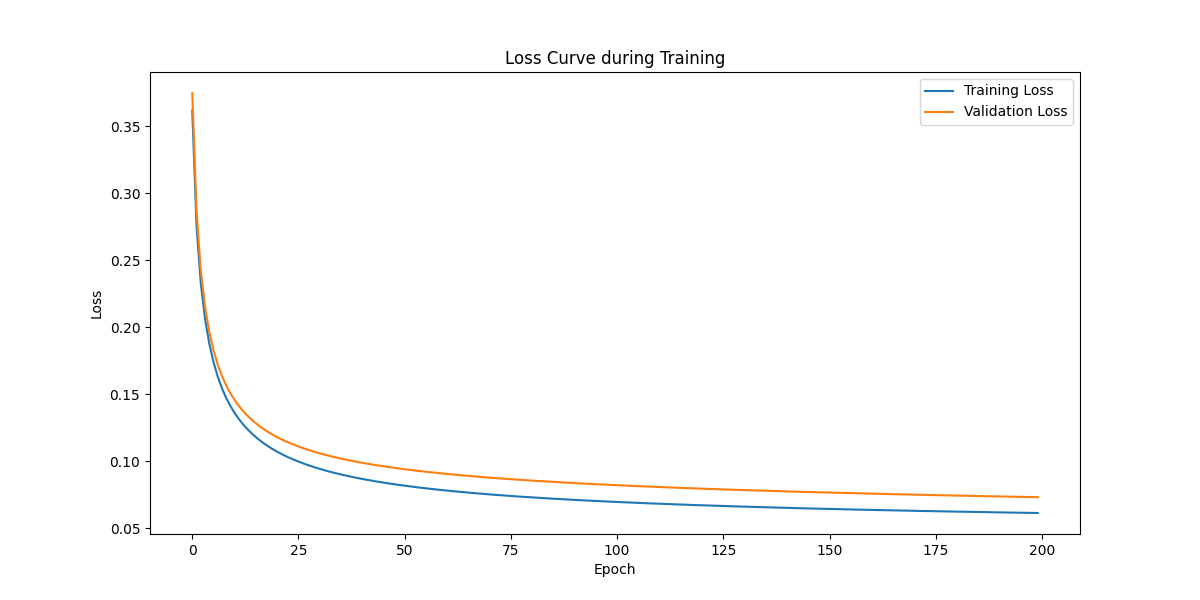




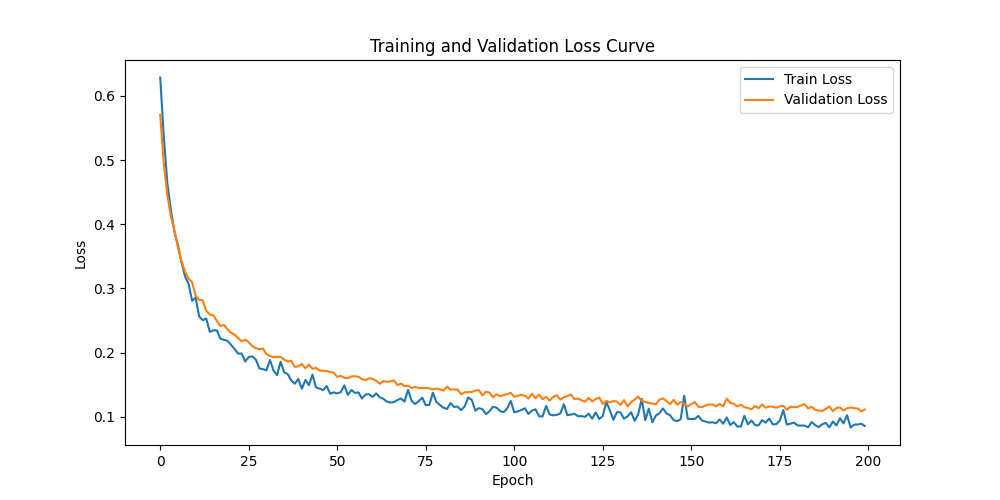


## 损失曲线对比

### 硬阈值线性分类

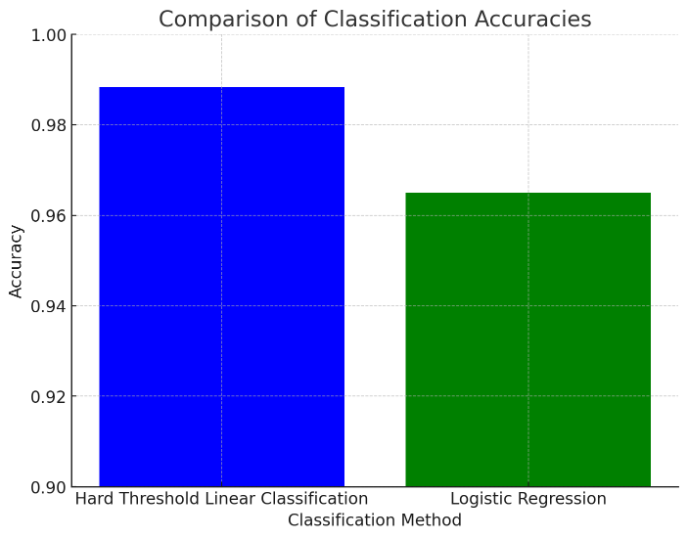


### Logistic回归



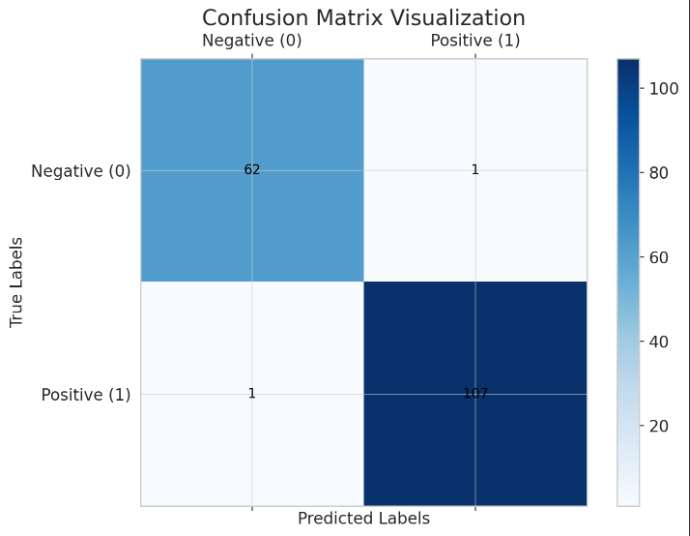
# 性能比较：准确度与混淆矩阵

## 准确度比较(epoch=200)



## 混淆矩阵分析

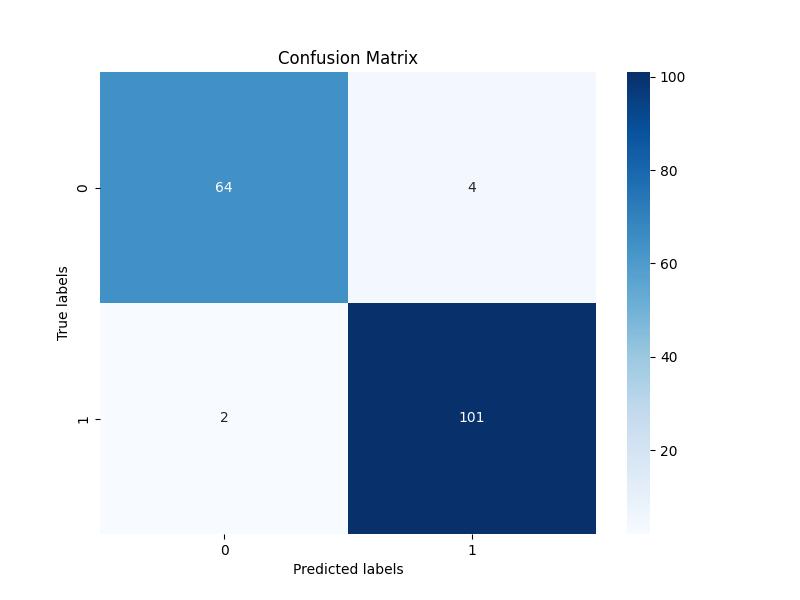
### 硬阈值线性分类



* **真阴性（True Negatives, TN）：**左上角的单元格显示了62，这表示模型正确地将62个实际为阴性的样本预测为阴性。
* **假阳性（False Positives, FP）：**右上角的单元格显示了1，这表示有1个实际为阴性的样本被模型错误地预测为阳性。
* **假阴性（False Negatives, FN）：**左下角的单元格也显示了1，这表示有1个实际为阳性的样本被模型错误地预测为阴性。
* **真阳性（True Positives, TP）：**右下角的单元格显示了107，这表示模型正确地将107个实际为阳性的样本预测为阳性。

混淆矩阵的主对角线显示了正确预测的数量，而非对角线显示了错误预测的数量。由此可见绝大多数的样本都被正确分类，模型的总体预测准确性很高。

### Logistic回归

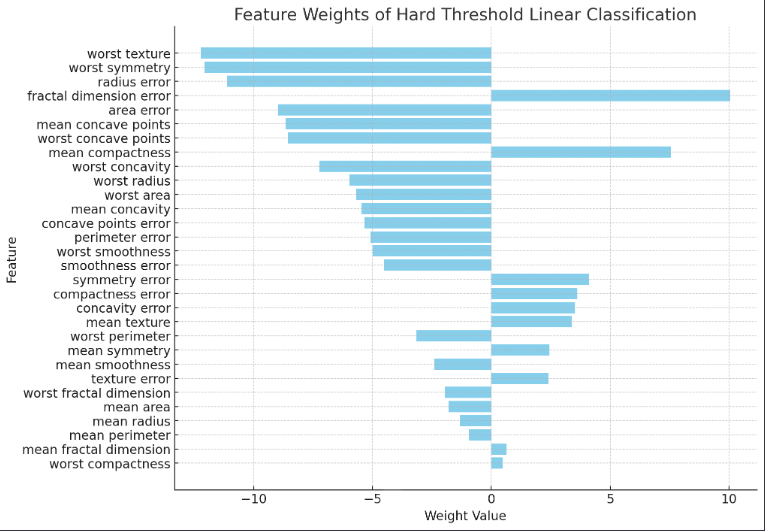


* **真阴性（True Negatives, TN）:** 左上角的单元格显示了64，这意味着模型正确地将64个实际为阴性的样本预测为阴性。
* **假阳性（False Positives, FP）:** 右上角的单元格显示了4，这意味着有4个实际为阴性的样本被模型错误地预测为阳性。
* **假阴性（False Negatives, FN）:** 左下角的单元格显示了2，这意味着有2个实际为阳性的样本被模型错误地预测为阴性。
* **真阳性（True Positives, TP）:** 右下角的单元格显示了101，这意味着模型正确地将101个实际为阳性的样本预测为阳性。

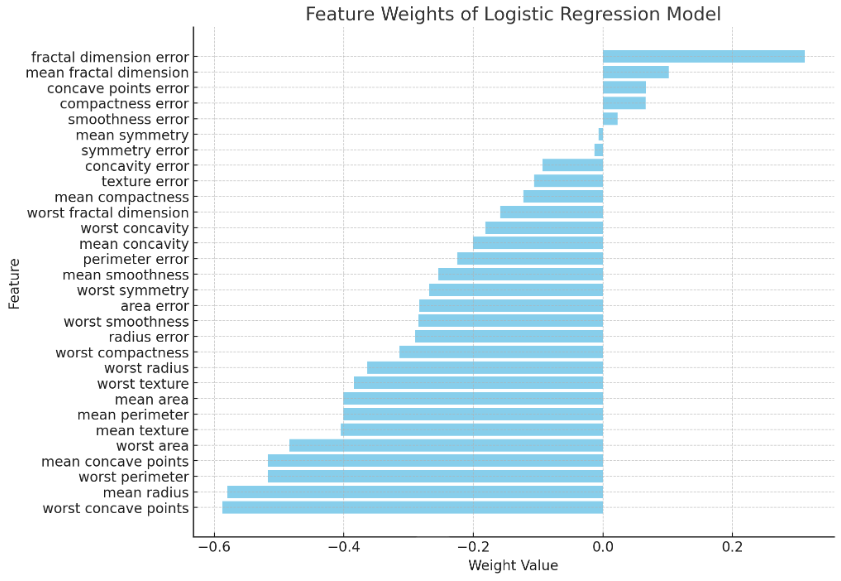
混淆矩阵的主对角线显示了正确预测的数量，而非对角线显示了错误预测的数量。由此可见绝大多数的样本都被正确分类，模型的总体预测准确性很高。

# 参数分析

## 硬阈值线性分类的参数分析



## Logistic线性分类的参数分析



# Logistic分类方法的特点和优势

## 特点

* **概率输出：**Logistic回归提供了概率估计，即每个类别的概率，而不仅仅是硬分类结果。这使得结果解释更为丰富，可以了解模型对其预测的确定性程度。
* **线性决策边界：**虽然它可以输出非线性概率，但Logistic回归本身是一个线性分类器，决策边界是特征空间中的线性划分。
* **低方差，高偏差：**作为线性模型，Logistic回归在处理复杂或非线性可分数据时可能显示出高偏差，但它的低方差特性使其对训练数据的小波动不那么敏感。
* **正则化能力：**Logistic回归可以通过L1或L2正则化来处理特征间的多重共线性，这有助于防止过拟合，并可以进行特征选择。

## 优势

* **解释性强：**Logistic回归模型简单，易于理解和解释。模型的每个特征都有一个权重系数，可以直观地表达该特征对预测结果的影响。
* **计算效率高：**相对于更复杂的模型，Logistic回归在训练和预测时都非常高效。
* **广泛应用：**由于其稳健和高效，它被广泛应用于各种领域，尤其是医学和社会科学领域。
* **良好的概率估计：**提供的概率输出对于需要权衡不同类型错误（如假阳性和假阴性）的决策过程非常有价值。
* **适应性强：**通过调整阈值，可以调整分类的灵敏度和特异性，这在不平衡的数据集或特定应用（如疾病筛查）中非常有用。

# 代码

https://github.com/SonicAge1/Logistic