

**模式识别与机器学习实验报告**

实验 二

|  |  |
| --- | --- |
| 题 目 | 逻辑回归 |
| 学 院 | 未来技术学院 |
| 专 业 | 人工智能 |
| 学 号 | 2023112419 |
| 学 生 | 陈铠 |
| 任 课 教 师 | 刘扬 |

哈尔滨工业大学计算机科学与技术学院

2025年秋季

1. **实验内容**

（一）逻辑回归简要介绍

逻辑回归（Logistic Regression）是以二分类为主的判别式、监督学习的分类模型。

1逻辑回归的核心任务

逻辑回归的核心任务是针对输入的特征向量学习映射关系：

通过学习数据特征，建立拟合函数来**估计后验概率**，再基于概率阈值判别最终类别。

在本实验中，主要研究逻辑回归的**线性二分类**问题。而事实上，也可以将逻辑回归算法拓展更广泛的问题上，例如通过对特征进行非线性映射，可以学习到非线性决策边界；通过设计One-vs-Rest或Softmax回归，可以实现多分类问题。

2逻辑回归的数学模型

逻辑回归选用的拟合函数为

其中为sigmoid函数。。

对于待分类样本，有

3逻辑回归的优化方法

3.1符号定义

表1 本实验使用符号

|  |  |  |
| --- | --- | --- |
| 名称 | 符号 | 描述 |
| 特征矩阵 |  | 每行为第个样本，含个特征（常置第一个特征为1） |
| 输出向量 |  | 每行为第个样本的预测值（0或1） |
| 参数向量 |  | 模型待优化变量 |
| 预测向量 |  | 每行 |

记训练集为且样本独立同分布，使用极大似然估计MLE以及交叉熵损失，可以得到

于是损失函数为

其中为正则化项

模型的训练目标是

3.2梯度下降法求解

梯度下降核心是求解损失函数梯度向量，再沿负梯度方向更新参数

单个参数的偏导数为

于是

更新公式为

其中为正则项。

3.3牛顿法求解

牛顿法使用二阶导数矩阵Hessian矩阵加速收敛

其中是对角矩阵，对角元素，是单位矩阵

更新公式为

其中，为正则项。

（二）实验研究内容

1研究样本数量、样本分布条件的影响

调整样本数量为100，200，类朴素贝叶斯分布与不满足类朴素贝叶斯分布。

2研究两种求解方法和正则化的影响，

分别为：**梯度下降+无正则化、梯度下降+正则化、牛顿方法+无正则化、牛顿方法+正则化**。

3研究模型在真实数据上的分类表现

使用UCI数据集（<https://archive.ics.uci.edu/dataset/267/banknote+authentication）>。

共1372个样本，特征维度为4，类别为2

1. **实验环境**

操作系统：Windows 11

实验平台：pycharm

解释器版本：Python 3.12

工具包：

表2 本实验使用工具包

|  |  |
| --- | --- |
| 工具包名称 | 版本 |
| pip | 25.2 |
| matplotlib | 3.10.6 |
| numpy | 2.3.3 |
| scikit-learn | 1.7.2 |

1. **实验结果及分析**

（一）样本数量、样本分布条件的影响

在该项中选用**梯度下降+无正则化**的求解方法

|  |  |  |
| --- | --- | --- |
|  |  | |
|  |  | |
| 图1-4样本数量、样本分布条件的影响训练结果 | |

表3 样本数量、样本分布条件的影响

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 样本量/类朴素贝叶斯分布 | 训练轮次 | 训练集acc | 验证集acc | 训练集loss |
| 100/是 | 200 | 0.994 | 1.000 | 0.0181 |
| 200/是 | 200 | 0.991 | 1.000 | 0.0187 |
| 100/否 | 200 | 0.981 | 1.000 | 0.0664 |
| 200/否 | 200 | 0.978 | 0.975 | 0.0751 |

样本分布的影响：当样本符合类朴素贝叶斯分布时，训练集acc≥0.99、loss≤0.0187、验证集acc=1.0，说明逻辑回归在数据满足线性可分或近似线性分布（如类朴素贝叶斯分布）时，模型拟合效率高、泛化稳定性强，且参数少。反之，逻辑回归性能略有下降，证明逻辑回归对数据分布敏感，依赖“特征与标签线性相关”假设

样本量的影响：类朴素贝叶斯分布下，逻辑回归在较小样本量的情况下均能保持极高准确率和泛化率，但当样本量较大时其训练时间、性能消耗增加，效果可能下降。

（二）两种求解方法和正则化的影响

在该项中选择样本数量为200，分布为类朴素贝叶斯分布。

|  |  |
| --- | --- |
|  |  |
|  |  |
| 图5-8 求解方法和正则化的影响训练结果 | |

表4 求解方法和正则化的影响

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 求解方法/正则化 | 训练轮次 | 训练集acc | 验证集acc | 训练集loss |
| 梯度下降/否 | 200 | 0.991 | 1.000 | 0.0187 |
| 梯度下降/是 | 200 | 0.991 | 1.000 | 0.0235 |
| 牛顿法/否 | 14 | 0.991 | 1.000 | 0.0132 |
| 牛顿法/是 | 10 | 0.991 | 1.000 | 0.0215 |

求解方法的影响：牛顿法显著降低了逻辑回归的训练轮次，说明其在解决该类问题上的高效性，但考虑到其求解方法需要计算Hessian矩阵，当样本特征映射至高维后，求解复杂度将极大增加，这是其局限性。

正则化的影响：逻辑回归加入正则化后acc不变、loss略升，说明其过拟合控制能力有限，正则化仅通过惩罚参数抑制复杂度，无法从模型结构上避免过拟合。

（三）模型在真实数据上的分类表现

表5 真实数据集上的表现

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 求解方法/正则化 | 训练轮次 | 训练集acc | 验证集acc | 训练集loss |
| 梯度下降/否 | 200 | 0.985 | 0.971 | 0.0418 |
| 梯度下降/是 | 200 | 0.985 | 0.971 | 0.0461 |
| 牛顿法/否 | 17 | 0.991 | 0.985 | 0.0183 |
| 牛顿法/是 | 10 | 0.993 | 0.985 | 0.0355 |

可以发现，模型在真实数据集上仍有较好的表现，且与自建数据集上的表现相近，验证了模型的正确性。

1. **结论**

样本符合类朴素贝叶斯分布时，模型训练精度高（acc≥0.991）、损失低（loss≤0.0187）、泛化能力强（验证集acc≥1.0）；而非该分布下，模型性能显著下降，且样本量增大会进一步加剧拟合难度与泛化能力下滑。​

牛顿法求解效率与性能均优于梯度下降，无论在类朴素贝叶斯分布数据还是真实数据中，牛顿法的收敛轮次仅为梯度下降的1/10-1/20，且能实现更高的训练精度与更低的训练损失，是逻辑回归模型更高效的求解方法。​

正则化在实验条件下作用有限：在样本符合类朴素贝叶斯分布（无过拟合）或真实数据（轻微过拟合风险）中，正则化仅轻微提高训练损失，对验证集泛化精度无影响；仅在牛顿法+真实数据的组合中，正则化小幅提升训练精度，但整体未改变模型核心性能。

1. **参考文献**

[1]（美）SHELDON AXLER著；杜现坤，刘大艳，[马晶](https://zhida.zhihu.com/search?content_id=183206416&content_type=Article&match_order=1&q=%E9%A9%AC%E6%99%B6&zhida_source=entity)译. [线性代数应该这样学](https://zhida.zhihu.com/search?content_id=183206416&content_type=Article&match_order=1&q=%E7%BA%BF%E6%80%A7%E4%BB%A3%E6%95%B0%E5%BA%94%E8%AF%A5%E8%BF%99%E6%A0%B7%E5%AD%A6&zhida_source=entity) 第3版[M]. 北京：人民邮电出版社, 2016.10.

[2][周志华](https://zhida.zhihu.com/search?content_id=183206416&content_type=Article&match_order=1&q=%E5%91%A8%E5%BF%97%E5%8D%8E&zhida_source=entity)著. 机器学习[M]. 北京：清华大学出版社, 2016.01.

[3]谢文睿，[秦州](https://zhida.zhihu.com/search?content_id=183206416&content_type=Article&match_order=1&q=%E7%A7%A6%E5%B7%9E&zhida_source=entity)编著. 机器学习公式详解[M]. 北京：[人民邮电出版社](https://zhida.zhihu.com/search?content_id=183206416&content_type=Article&match_order=2&q=%E4%BA%BA%E6%B0%91%E9%82%AE%E7%94%B5%E5%87%BA%E7%89%88%E7%A4%BE&zhida_source=entity), 2021.03.

[4][李航](https://zhida.zhihu.com/search?content_id=183206416&content_type=Article&match_order=1&q=%E6%9D%8E%E8%88%AA&zhida_source=entity)著. 统计学习方法 第2版[M]. 北京：清华大学出版社, 2019.05.