实验2　对数几率回归

1. 实验目的

（1）掌握对数几率回归算法。

（2）理解分类相关指标。

2. 数据集介绍

数据集	样本数	维度	类数	数据类型
MINIST	3000	784	10	手写体数字
Yale	165	1024	15	人脸图像
lung	203	3312	5	生物数据

数据集中的labels都已经展平为一维数据。

3. 实验内容

1. **数据预处理**：采用随机方法划分训练与测试集（7:3），并进行数据的标准化。
2. **模型架构**：基于一对多策略的多分类逻辑回归，参数矩阵维度为（特征数+1）×类别数。
3. **优化方法**：手动实现批量梯度下降（**学习率为0.1**，**迭代300个epochs**），同时加入L2正则化（λ=0.1）。
4. **评估指标**：计算准确率、精确率、召回率、F1值，输出混淆矩阵与分类报告等结果。

4. 实验结果

4.1 算法原理

4.1.1 一元对数几率回归算法

对数几率回归是一种用于解决分类问题的统计模型，尤其适用于**二分类**任务。它的核心思想是通过线性组合输入特征，并利用对数几率和Sigmoid函数将线性输出映射到 [0,1] 区间，从而表示样本属于某一类的概率。

模型定义：

给定输入特征向量$\mathbf{x}$，模型计算其线性组合：

$$z = \mathbf{w}^T \mathbf{x} + b$$

其中$\mathbf{w}$是权重向量，b是偏置项。然后通过Sigmoid函数：

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

将z转换为概率值$P(y = 1 \mid \mathbf{x})$，表示样本属于正类的概率。

将类别转换为几率：

几率是指事件发生的概率与不发生的概率之比，即$\frac{P}{1-P}$，而**对数几率**是对几率取自然对数，得到线性表达式：

$$\ln\left(\frac{P(y=1 \mid \mathbf{x})}{1-P(y=1 \mid \mathbf{x})}\right)=\mathbf{w}^T\mathbf{x}+b$$

这表明模型实际上是在用线性函数拟合对数几率。

参数估计：

模型参数$\mathbf{w}$和b通过最大似然估计来学习。假设样本独立，似然函数是所有样本预测概率的乘积。通常我们取对数似然的负值，得到**交叉熵损失函数：**

$$\ell(\mathbf{w},b)=-\sum_{i=1}^{n}\left[y_i\ln P(y_i \mid \mathbf{x}_i)+(1-y_i)\ln(1-P(y_i \mid \mathbf{x}_i))\right]$$

优化方法：

由于该损失函数是凸函数，可以使用梯度下降或牛顿法等优化算法进行求解。在本实验所用的梯度下降中，参数的更新规则为：

$$\mathbf{w}\leftarrow\mathbf{w}-\alpha\frac{\partial\ell}{\partial\mathbf{w}},\quad b\leftarrow b-\alpha\frac{\partial\ell}{\partial b}$$

其中α是学习率，同时**Sigmoid函数的导数和梯度的计算公式分别为**：

$$\sigma'(z)=\sigma(z)(1-\sigma(z))$$

$$\frac{\partial\ell}{\partial\mathbf{w}}=\sum_{i=1}^{n}(\sigma(\mathbf{w}^T\mathbf{x}_i+b)-y_i)\mathbf{x}_i$$

模型输出：

模型的输出$P(y=1 \mid \mathbf{x})$可以直接解释为样本属于正类的概率。在实际分类时，**通常会设定一个阈值**（如0.5），若概率大于等于阈值则预测为正类，否则预测为负类。

4.1.2 从二分类问题到多分类问题的推广

当目标变量Y有K个类别（$K>2$）时，对每个类别k训练一个二分类器，判断样本是否属于该类：

$$P(Y=k|\mathbf{x})=\frac{e^{\mathbf{w}_k^T\mathbf{x}}}{1+e^{\mathbf{w}_k^T\mathbf{x}}}$$

预测时选择**概率最大**的类别。

4.2 建模结果

4.2.1 MINIST 数据集

表1 MINIST 数据集预测结果分类报告

Class	Precision	Recall	F1-Score	Support
1	0.92	0.95	0.93	60
2	0.85	0.93	0.89	60
3	0.82	0.77	0.79	60
4	0.90	0.88	0.89	60
5	0.85	0.87	0.86	60
6	0.91	0.82	0.86	60
7	0.87	0.92	0.89	60
8	0.88	0.88	0.88	60
9	0.84	0.87	0.85	60
10	0.88	0.83	0.85	60
Accuracy			0.87	600
Macro Avg	0.87	0.87	0.87	600
Weighted Avg	0.87	0.87	0.87	600

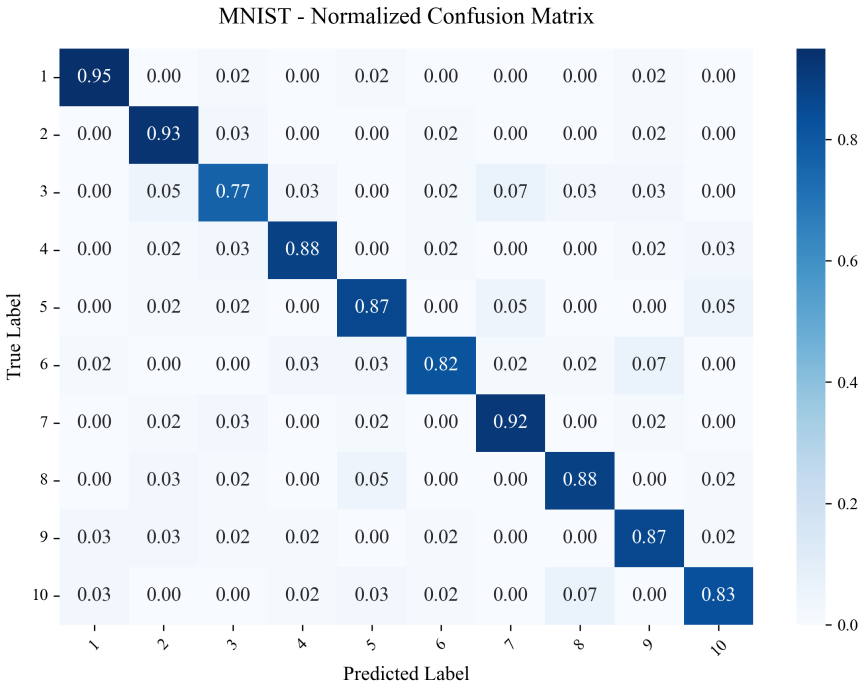

图1　MINIST 数据集预测结果的混淆矩阵

4.2.2 Yale 数据集

表2 Yale 数据集预测结果分类报告

Class	Precision	Recall	F1-Score	Support
1	0.40	1.00	0.57	2
2	1.00	0.50	0.67	2
3	1.00	1.00	1.00	2
4	0.60	1.00	0.75	3
5	1.00	1.00	1.00	2
6	0.50	0.50	0.50	2
7	1.00	1.00	1.00	3
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	2
10	0.67	1.00	0.80	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	2
13	1.00	0.67	0.80	3
14	0.00	0.00	0.00	2
15	0.00	0.00	0.00	2
Accuracy			0.79	33
Macro avg	0.74	0.78	0.74	33
Weighted avg	0.76	0.79	0.75	33

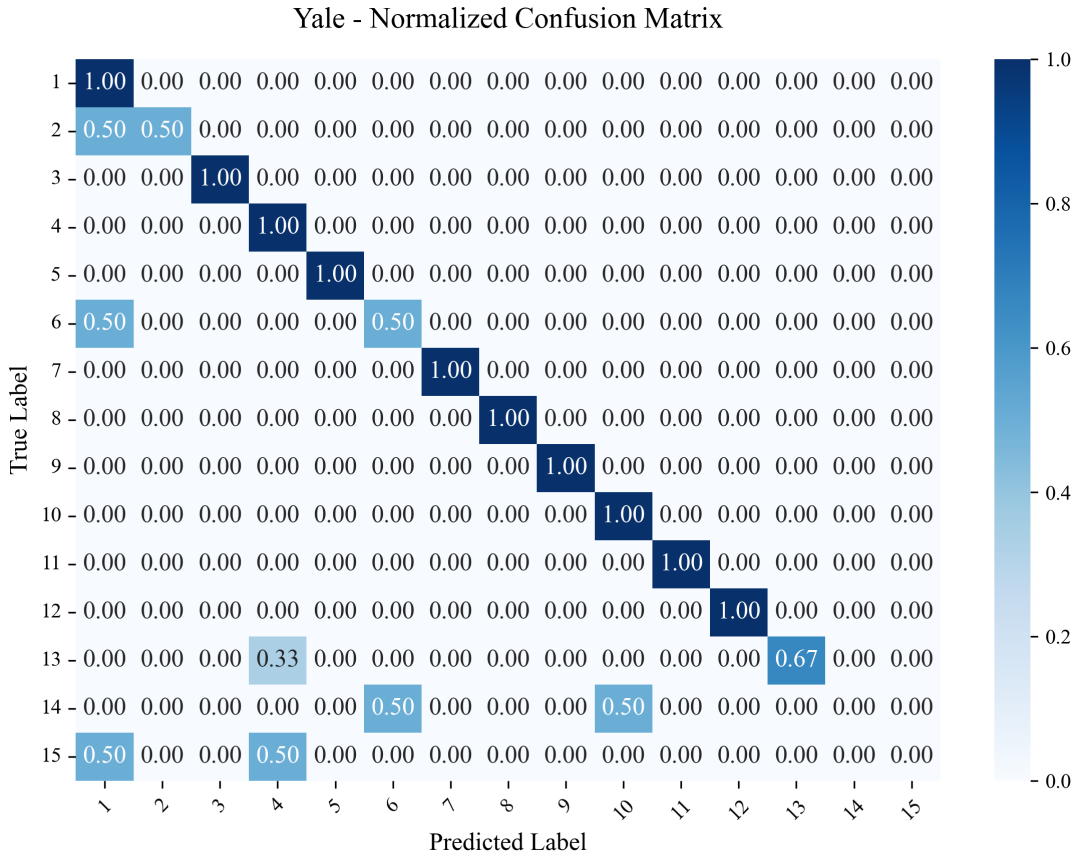

图2　Yale 数据集预测结果的混淆矩阵

4.2.3 lung 数据集

表3 lung 数据集预测结果分类报告

Class	Precision	Recall	F1-Score	Support
1	1.00	0.82	0.90	28
2	0.67	1.00	0.80	4
3	0.60	0.75	0.67	4
4	0.80	1.00	0.89	4
5	0.50	1.00	0.67	1
Accuracy			0.85	41
Macro avg	0.71	0.91	0.78	41
Weighted avg	0.90	0.85	0.86	41

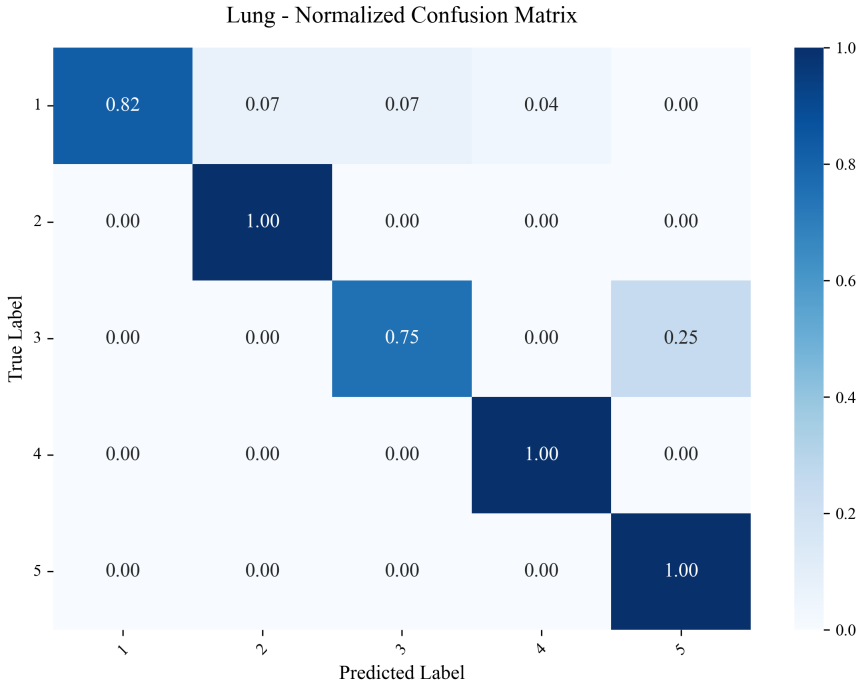

图3　lung 数据集预测结果的混淆矩阵

4.2.4 三数据集Accuracy结果对比

图4展示了模型在MNIST、Yale和Lung三个数据集上的分类准确率对比，结果显示模型在MNIST上表现最佳，在结构更复杂、类别更多的人脸数据集Yale上表现较差，而在医学图像数据集Lung上表现接近MNIST，说明对数几率回归模型对结构清晰的数据具有较强的识别能力，但在处理类间差异较小、样本不平衡的复杂任务时仍存在一定的挑战。

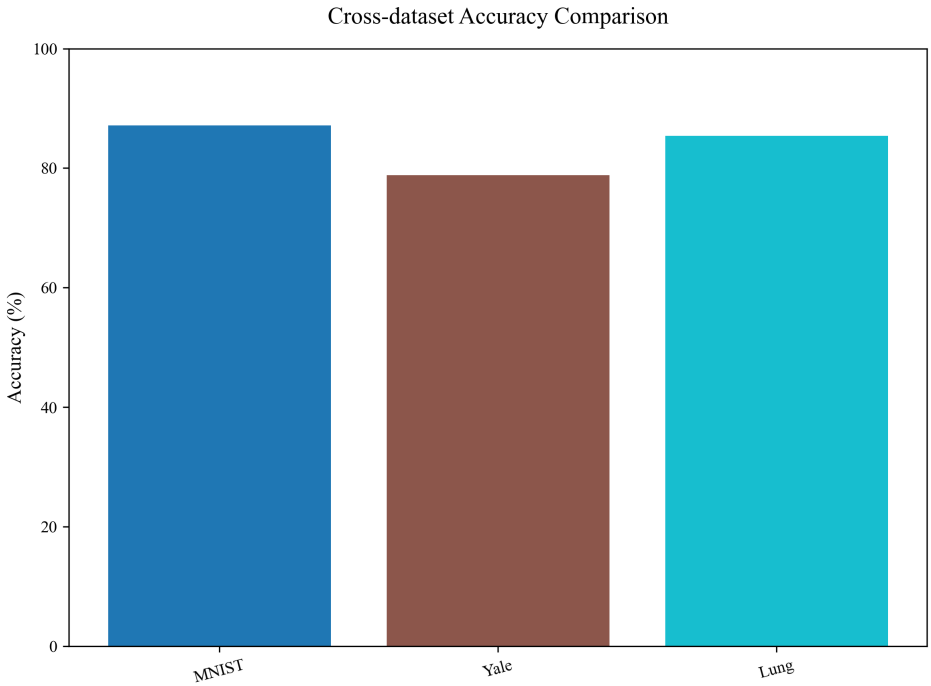

图4　各个数据集预测结果的Acc对比

5 实验感悟

通过本次基于对数几率回归的分类实验，我体会到传统统计学习方法在特定场景下的独特价值。在MNIST多分类任务中，尽管模型结构简单，依然取得了87%的整体准确率。实验中手动实现梯度下降的过程，让我直观理解了学习率与迭代次数对收敛性的影响：过大的学习率会导致震荡，而过小的学习率则需要更多迭代才能收敛。