《机器学习导论》作业一

LSC 和 FLD 分类界线可视化

聂嘉一　niejy20@lzu.edu.cn　220240945241

Lanzhou University

Date: 2024 年 11 月 30 日

1 算法简介

1.1 最小二乘分类器

最小二乘分类器（LSC）是一种基于线性回归的简单分类算法，适用于二分类问题。

它的核心思想是通过最小化平方误差来找到一个线性决策边界，将不同类别的数据点分开。用伪代码表示如下：

Algorithm 1: 最小二乘分类器

Input: 训练数据集 $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$，其中 $x_i \in \mathbb{R}^d$，$y_i \in \{0, 1\}$

Output: 权重向量 $w \in \mathbb{R}^d$ 和偏置项 $b \in \mathbb{R}$

foreach $i \in \{1, 2, \ldots, n\}$ **do**

  $\tilde{x}_i = [1, x_i^T]^T$;

$\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n]^T$;

$y = [y_1, y_2, \ldots, y_n]^T$;

$\tilde{w} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T y$;

$w = \tilde{w}_{1:d}$;

$b = \tilde{w}_0$;

return w, b ;

分类规则: 对于新的数据点 x：

$$\hat{y} = w^T x + b$$

如果 $\hat{y} > 0$，则预测为类别 1；否则预测为类别 0。

1.2 Fisher 判别分析

Fisher 判别分析（FLD），也称为线性判别分析（LDA, Linear Discriminant Analysis），是一种用于分类和降维的监督学习方法。

它的核心思想是找到一个投影方向，使得不同类别的数据在该方向上的类间散布最大化，同时类内散布最小化。用伪代码表示如下：

Algorithm 2: Fisher 判别分析 (FLD)

Input: 训练数据集 $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$，其中 $x_i \in \mathbb{R}^d$，$y_i \in \{0, 1\}$

Output: 最优投影方向 $w \in \mathbb{R}^d$ 和偏置项 $b \in \mathbb{R}$

$\mu_1 = \frac{1}{n_1} \sum_{i:y_i=1} x_i$;
$\mu_0 = \frac{1}{n_0} \sum_{i:y_i=0} x_i$;
$S_1 = \sum_{i:y_i=1} (x_i - \mu_1)(x_i - \mu_1)^T$;
$S_0 = \sum_{i:y_i=0} (x_i - \mu_0)(x_i - \mu_0)^T$;
$S_W = S_1 + S_0$;
$S_B = (\mu_1 - \mu_0)(\mu_1 - \mu_0)^T$;
$w = S_W^{-1}(\mu_1 - \mu_0)$;
$b = -\frac{1}{2} w^T (\mu_1 + \mu_0)$;
return w, b ;

分类规则: 对于新的数据点 x：

$$y = w^T x + b$$

如果 $y > 0$，则预测为类别 1；否则预测为类别 0。

2 数据处理

2.1 数据集

本次作业需要对二维数据进行分类，我使用了 Iris 这个典型数据集，它包含了 3 种鸢尾花（Iris Setosa、Iris Versicolor、Iris Virginica）的 4 种特征数据：

- 萼片长度 (Sepal length)
- 落片宽度 (Sepal width)
- 花瓣长度 (Petal length)
- 花瓣宽度 (Petal width)

在使用 matplotlib 绘图前，先筛选出 2 种鸢尾花的 2 种特征作为输入。**filter.py** 文件实现了上述处理过程。

2.2 可视化过程

运行 main.py 时，程序会自动执行以下步骤：

1. 调用 filter_data 函数筛选数据并保存到 iris_filter.data
2. 读取筛选后的数据
3. 提取特征和标签，划分训练集和测试集
4. 训练 LSC 和 FLD 模型
5. 评估模型性能、输出准确率
6. 绘制数据点和两条决策边界

结果如下图1所示：

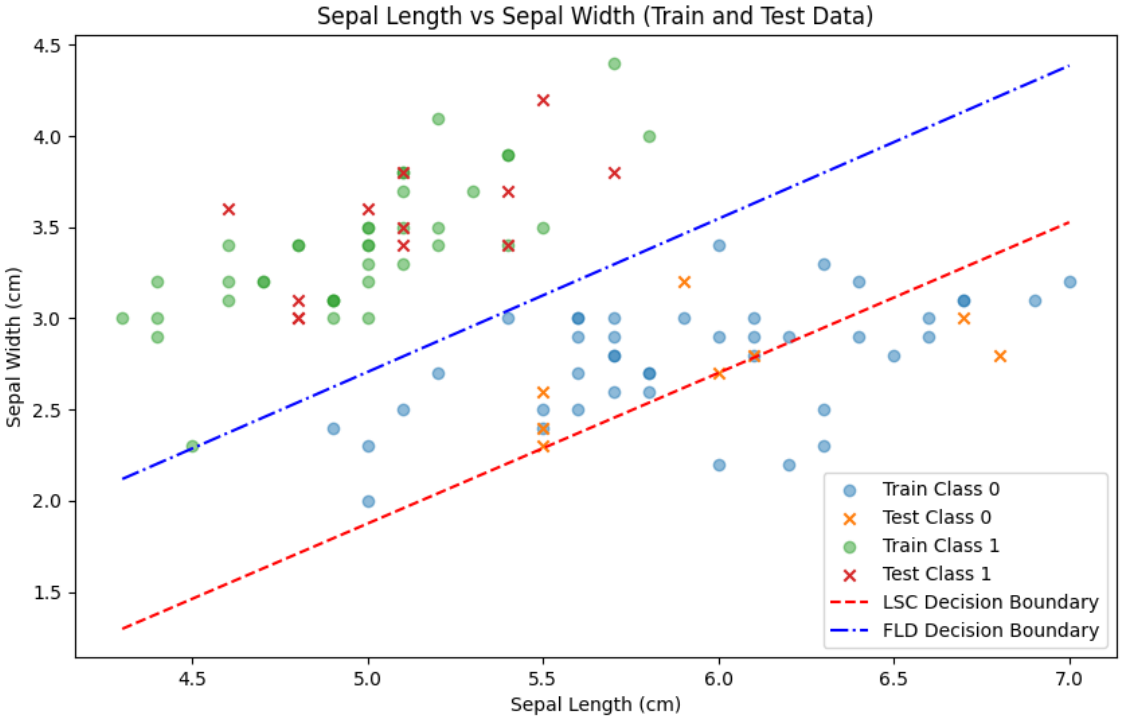

图 1: 可视化结果

3 结论与展望

根据测试结果，对于 Iris 数据集的分类，FLD 模型表现更好。

本次作业有两处待改进的地方：

1. 数据集体量小：Iris 数据集共有 150 条数据且三种类别数目相等，由于我只使用了其中两种类别，因此训练集与测试集共计 100 条数据，数据集过小可能导致了 FLD 的过拟合（LSC 和 FLD 模型准确率分别为 70% 与 100%）。因此可考虑使用体量更大的数据集。

2. 数据选择待优化：Iris 数据集包含了鸢尾花的 4 种特征，只选择其中 2 种特征进行类别判断，对于模型性能可能会有影响。另外，还可以遍历所有的类别与特征组合，使用 matplotlib 子图进行可视化，研究选择不同的特征是否会影响模型性能。

4 致谢

本文 LaTeX 模板为ElegantPaper。
本文部分知识来自 机器学习导论。