模式识别—实验三：Fisher 线性判别分析

实验报告

姓名　　文泓力　　学号　202111150036　　专业　人工智能

一、 实验任务

实验目的和要求

1、 理解 Fisher 线性判别分析的基本原理；
2、 熟练掌握降维、分类以及可视化的代码实现过程；
3、 要求写出实验对应的算法原理并写出对应代码的测试步骤；

实验内容

实验：Fisher 线性判别分析

1、问题描述：

已知 Iris 数据集包含了三种不同品种的鸢尾花（Iris），每种品种有 50 个样本，一共包含 150 个样本。每个样本都包括以下四个特征：花萼长度、花萼宽度、花瓣长度和花瓣宽度。基于这四个特征，数据集中的每个样本都被标记为以下三个鸢尾花品种之一：山鸢尾花（Iris Setosa）、变色鸢尾花（Iris Versicolor）和维吉尼亚鸢尾花（Iris Virginica）。

这三种品种的鸢尾花在这些特征上具有差异，要求基于这四个特征，使用 Fisher 线性判别分析的方法进行降维区分，并可视化结果。

2、实验内容：

（1） 使用 sklearn 库导入 Iris 数据集，对特征进行降维，并可视化降维后的结果。
（2） 使用感知器分类器进行分类，计算准确率（调用 Perceptron 类）。

二、 算法原理概述

（简单写一下实验用到的核心知识点或公式，不需要过于详细）

Fisher 线性判别分析（Fisher LDA），也称为 Fisher 判别分析，是一种用于特征提取和降维的监督学习方法，通常用于分类问题。它的主要目标是找到一个投影方向，使得不同类别的数据在这个方向上有最大的分离度，同时类内方差最小，从而提高分类的效果。以下是 Fisher LDA 的原理概述：

1.问题描述：

假设我们有一个具有两个或更多类别的数据集，每个类别包含多个特征的样本。Fisher LDA 的目标是将数据从原始的高维特征空间投影到一个低维的线性子空间，以便最大程度地分离不同类别的样本。

2.基本思想：

Fisher LDA 的核心思想是最大化类间离散度（between-class scatter）并最小化类内离散度（within-class scatter）。类间离散度衡量不同类别之间的分离程度，而类内离散度衡量同一类别内部数据点的分散程度。

3.步骤：

假设我们有两类数据，也就是说这里是针对二分类问题以及对应的标签。数据表示为 D=(x1,y)是 1，m=2,y1∈(0,1)。我们希望可以找到一条直线 y=w1x1 十 w2x2=wx 把两种样本分开。为了达到这个目的，我们首先要让同类的样本投影点尽可能小，即希望 wTΣw 小，同时我们还想让两个类的中心 u1,u2 投影后的距离尽量大，即 wTu1 一 wTu2 尽量大。我们采用以下步骤实现：

（1）对于每个类别，计算该类别的均值向量（平均值向量）。

（2）计算整体均值向量，即所有样本的平均值向量。

（3）计算类内离散度矩阵（within-class scatter matrix）和类间离散度矩阵（between-class scatter matrix），它们即一个类内的差异大小与类间差异大小的最直观体现。

$$S_b = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

$$S_w = \Sigma_1 + \Sigma_2 = \sum_{x \in X1}(x - \mu_1)(x - \mu_1)^T + \sum_{x \in X2}(x - \mu_2)(x - \mu_2)^T$$

（4）计算 Fisher 准则函数，这是类间离散度和类内离散度的比值，其又被称为“广义瑞利商“。

$$J = \frac{w^T S_b w}{w^T S_w w}$$

（5）最优化准则函数，等价于求解以下优化问题：

$$\min_w \ -w^T S_b w$$

$$s.t. w^T S_w w = 1$$

可以通过拉格朗日乘子法实现：

$$c(w) = w^T S_b w - \lambda(w^T S_w w - 1)$$

$$\Rightarrow \frac{dc}{dw} = 2S_b w - 2\lambda S_w w = 0$$

$$\Rightarrow S_b w = \lambda S_w w (2)$$

通过观察可知 Sbw 的方向恒为 μ1−μ2，即：

$$S_b w = \lambda(\mu_1 - \mu_2)$$

可以得出：

$$w = S_w^{-1}(\mu_1 - \mu_2)$$

但是我们进行分类的鸢尾花数据集是有三个类别的，对此计算不再适用，我们定义的“类间散度矩阵”

$$S_b = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

就不再适用，所以我们这里引入“全局散度矩阵”：

$$S_t = S_w + S_b = \sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

从上面两个式子推得:

$$S_b = S_t - S_w = \sum_{i=1}^{N} m_i(\mu_i - \mu)(\mu_i - \mu)^T$$

使用同样的优化方法可以求得 W 满足下式，即它是 SwSb 的特征向量：

$$S_w^{-1} S_b W = \lambda W$$

（6）找到使 Fisher 准则函数最大化的投影方向，通常是通过计算特征值和特征向量来实现，我们可以根据目标维数来选择需要的最大的几个特征向量组合成为降维矩阵，它们和样本的积即为降维样本。

4.投影：

一旦找到了最佳投影方向，可以将数据投影到这个方向上，从而将数据降维到一维或更低维的空间。在这个新的低维空间中，数据点的坐标就是它们在最佳投影方向上的投影值。

5.分类：

降维后的数据可以用于分类问题。一般来说，Fisher LDA 的目标是使不同类别之间的投影均值尽可能远离，以便更容易进行分类。

Fisher LDA 通常与其他分类算法（如最近邻、支持向量机等）结合使用，以实现更准确的分类。这是因为 Fisher LDA 仅关注降维，而不包括分类边界的构建。它在特征提取和数据可视化方面非常有用，尤其是需要将高维数据转化为更易理解和处理的低维表示时。

三、 测试步骤

（如运行哪个文件，可以得到什么样的结果，可以简单描述便于助教去测试）

代码结构如下所示：

```
| dataset.py
| main.py
| plot.py
| requirements.txt
| utils.py
| imgs
```

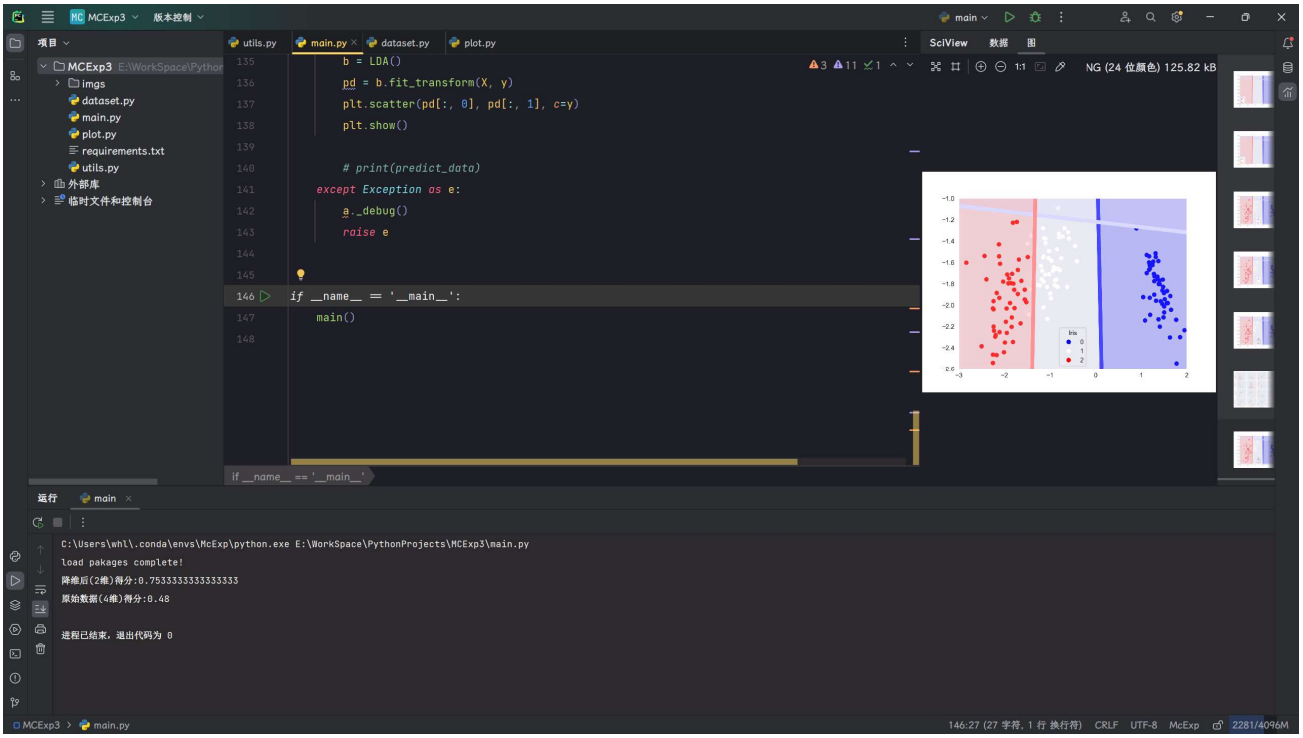
Dataset.py 处理数据集

Plot.py 定义画图函数，运行可以得到原始样本的分布图

Utils.py 导入工具包

Main.py 定义测试入口

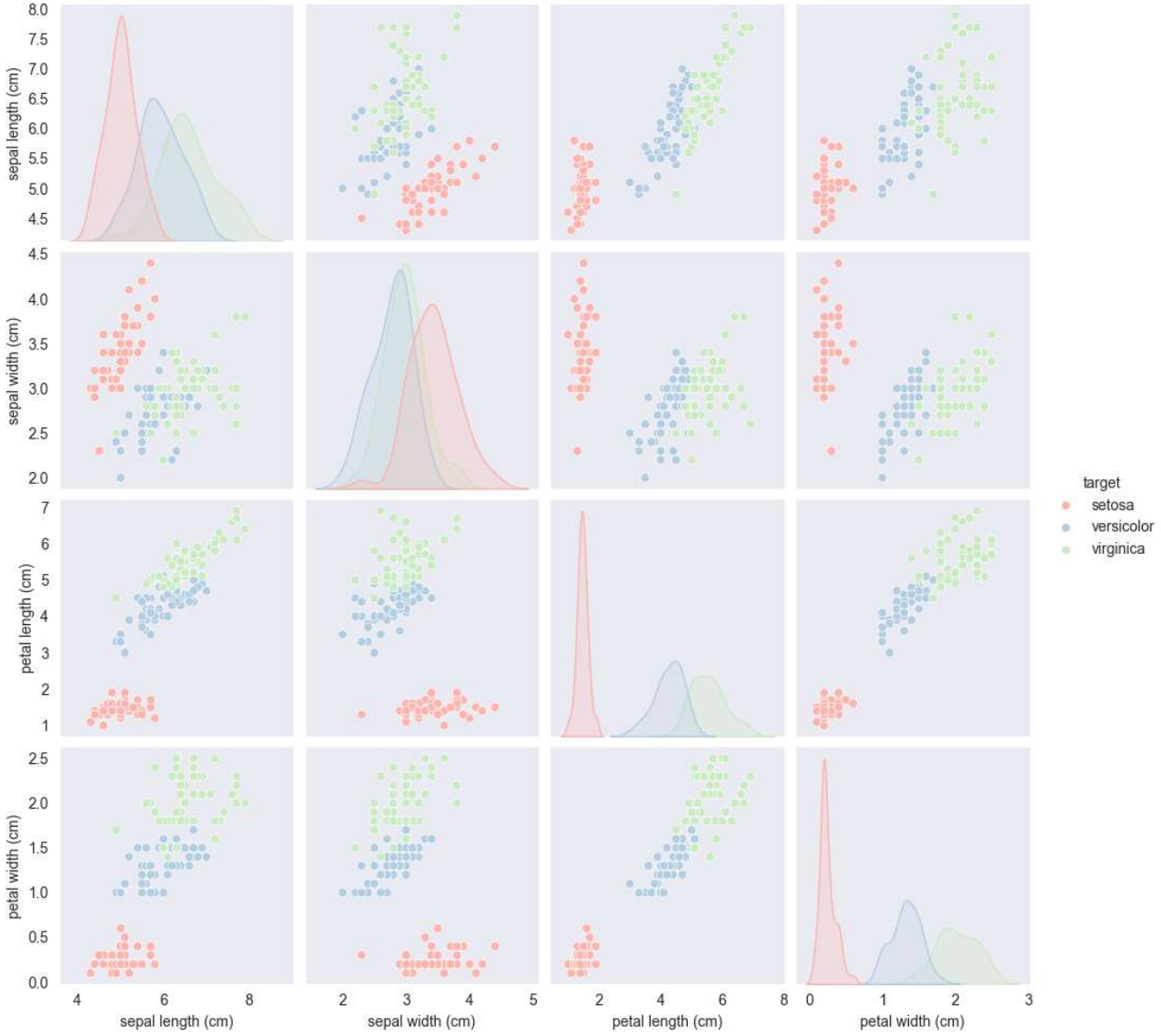
相关依赖在 requirements.txt 中定义，运行 main.py 进行测试，程序将读取数据集，进行对样本降维和分类，并打印降维前后分类的结果得分，然后绘制降维成 2 维后的样本分布和分类决策面。



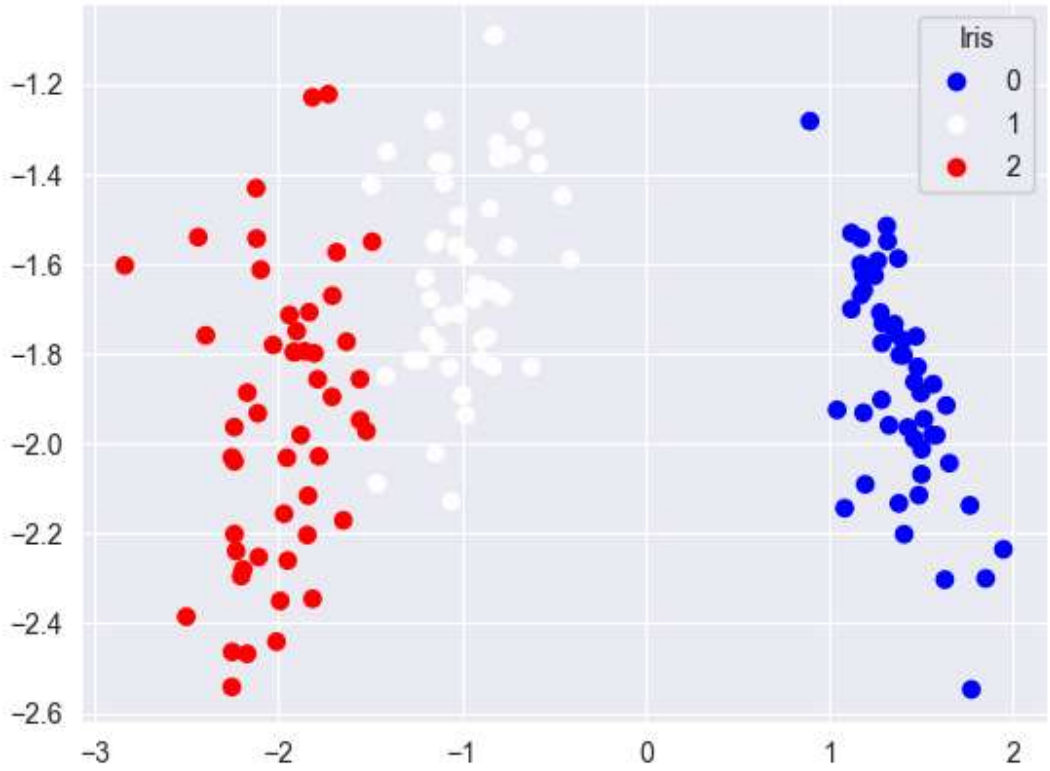
四、 实验结果与讨论

（可以根据得到的结果分析原因）

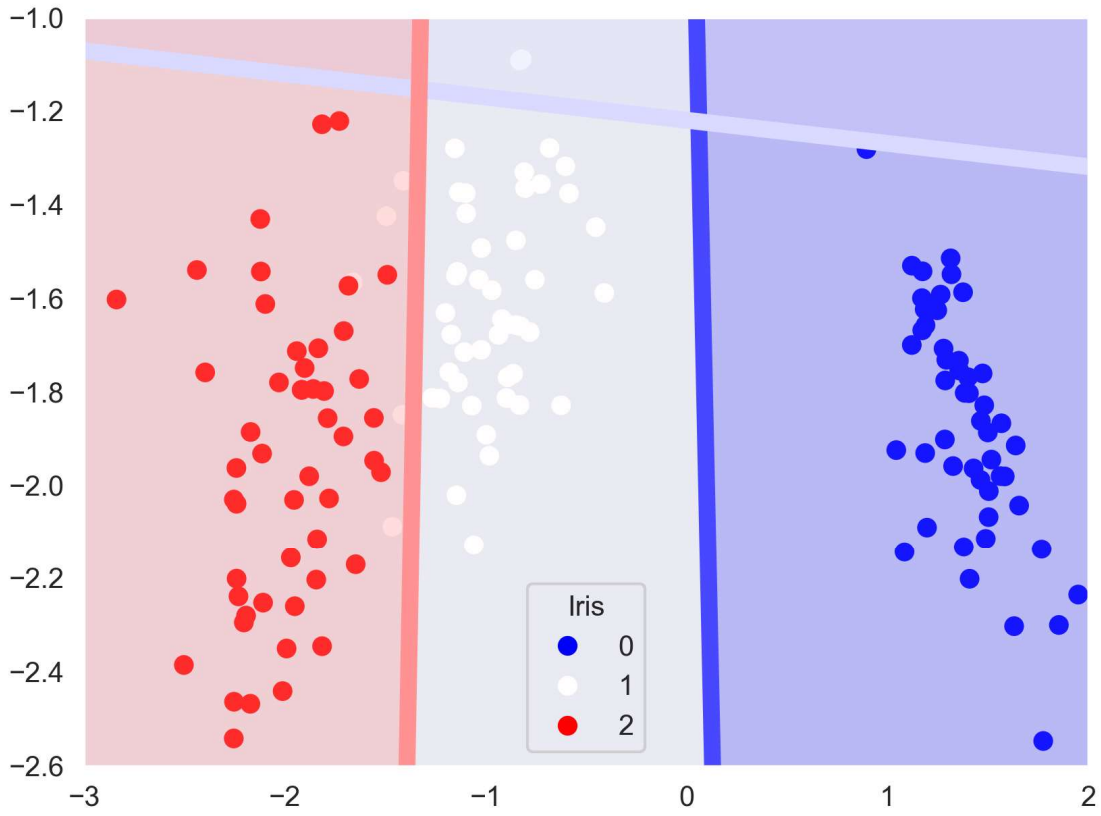
样本原始分布：


sepal length (cm)
sepal width (cm)
petal length (cm)
petal width (cm)
target
setosa
versicolor
virginica

降维后分布：



感知机分类后决策面：



分类正确率得分对比：

```
降维后(2维)得分:0.7533333333333333
原始数据(4维)得分:0.48
```

从实验结果分析可以得出，降维后的分类准确度明显上升，说明 Fisher 降维对感知机模型有着明显的提升，可以作为优化感知机模型的手段。

但是使用本来效果已经较好的分类器时，提升不明显，下图是采用 SGDClassifier()的分类结果：

```
降维后(2维)得分:0.9733333333333334
原始数据(4维)得分:0.94
```

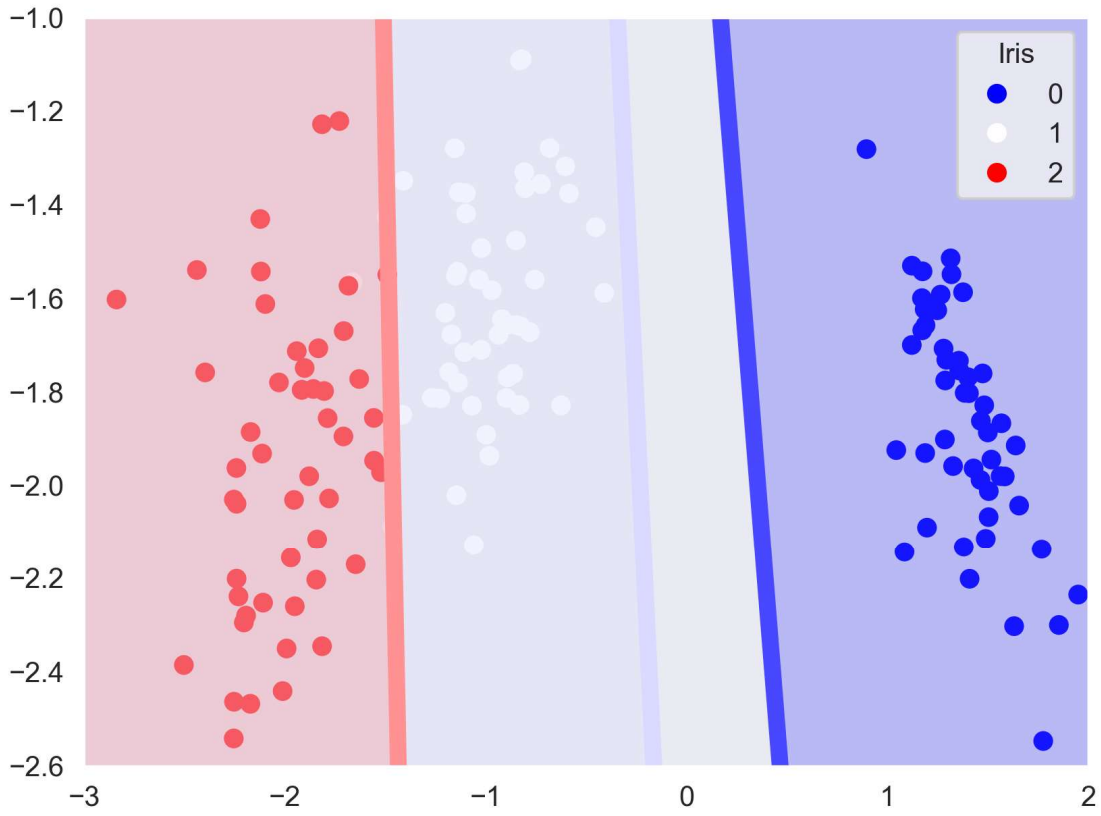
当使用非线性分类器例如核函数支持向量机时，几乎已经没有效果：

```
降维后(2维)得分:0.98
原始数据(4维)得分:0.9733333333333334
```

值得注意的是，使用线性核的支持向量机甚至在原始数据上取得了更好的效果：

```
降维后(2维)得分:0.98
原始数据(4维)得分:0.9933333333333333
```

如下图所示，其几乎完美分类：



五、 实验总结(总结本次实验课学习的内容、结论、反思)

本次实验我们使用了 Python 实现 Fisher 线性分析，在鸢尾花数据集上使用了该算法进行了降维实验，并且将降维后的数据输入感知机中进行分类，得到结果与未降维的数据进行对比，发现使用 Fisher 降维后的数据可以明显提升分类结果的准确性。

实验中我发现鸢尾花数据集为三类分类，不符合我们上课讲授的公式算法，通过查找资料，完善了多分类的 Fisher 线性降维的算法。

实验结果说明了降维可以帮助某些分类器取得更好的效果，有以下原因。 Fisher 降维通常会选择在原始特征空间中具有最大方差的投影方向，这有助于减少特征之间的冗余信息。减少冗余信息可以提高模型的泛化能力，减少过拟合的可能性。总之，Fisher 降维有助于改善分类问题的性能，因为它能够更好地利用数据的结构信息，提高类别分离度，并减少冗余信息，从而增强了分类器的性能和泛化能力。然而，它也有一些局限性，例如对数据的线性假设和需要满足高斯分布等假设。我们可以从实验结果中分析得出，对于信息获取能力较强的分类器例如支持向量机，降维反而会因为损失一些信息带来负作用。

通过本次实验我进一步加深了对 Fisher 算法的理解，强化了实践应用能力，充分意识到每种算法的优势、适用区间以及它们的局限性。