**深 圳 大 学 实 验 报 告**

|  |
| --- |
| **课程名称 机器学习**  **项目名称 实验四：SVM算法**  **学 院 计算机与软件学院**  **专 业 软件工程（腾班）**  **指导教师 赖志辉**  **报 告 人 黄亮铭 学号 2022155028**  **实验时间 2024年4月8日至2024年4月21日**  **实验报告提交时间 2024年4月21日** |

**教务处制**

# 实验目的与要求

1. 熟悉SVM算法，要求理解SVM算法的基本原理。
2. 实现简单的SVM算法。
3. 使用实际数据集测试所写的SVM算法。
4. 分析SVM算法的分类效果。

# 二、实验内容与方法

1. 分别给出经典的/ 软间隔/核-SVM的优化问题并推导其求解优化过程，实现经典的SVM算法进行图像识别；在二维平面对二类问题给出support vector的一个示例。
2. 用PCA、LDA算法提取前 10,20,30,...,160维的图像特征，然后再用不同SVM进行识别，并比较识别率。

# 三、实验步骤与过程

1. 分别给出经典的/ 软间隔/核-SVM的优化问题并推导其求解优化过程，实现经典的SVM算法进行图像识别；在二维平面对二类问题给出support vector的一个示例。
   1. **SVM算法简介**

支持向量机是一种经典的机器学习算法，常用于分类和回归分析。它在许多实际问题中都表现出色，特别是在高维空间中的数据分类任务上。

* + - **基本原理：**SVM的核心思想是找到一个最优的超平面，将不同类别的数据点分开。对于二分类问题，这个超平面可以被看作是一个将特征空间划分为两个部分的决策边界。最优的超平面是指能够最大化类别间间隔（margin）的超平面，即使得不同类别的数据点离超平面的距离最大化。
    - **关键概念：**

**①** 支持向量（Support Vectors）： 离超平面最近的数据点被称为支持向量。这些支持向量决定了超平面的位置和方向。

② 间隔（Margin）： 类别间间隔是指在超平面和支持向量之间的距离，SVM试图最大化这个间隔，以提高分类器的泛化能力。

③ 核技巧（Kernel Trick）： SVM可以通过引入核函数将数据映射到高维空间中，从而使得在原始特征空间中线性不可分的数据在新的空间中变得线性可分。

* + - **优点：**1）SVM对于高维空间中的数据分类效果好，适用于特征维度较高的情况。

2）通过核技巧，SVM能够处理非线性可分的数据。

3）SVM的训练过程中通常只涉及到一部分训练样本，因此对内存消耗较小。

* + - **小结：**

支持向量机是一种强大且灵活的机器学习算法，适用于各种分类和回归任务。它的优雅数学基础以及在高维空间中的优良性能使得它成为了许多实际问题中的首选算法之一。

* 1. **经典SVM优化问题及推导优化过程**

样本点与超平面的距离为：

若分类正确，则有：

两异类支持向量到超平面的距离之和为：

我们的目标是找到最大间隔以分类，即：

问题等价于如下形式：

令

对L求偏导：

令，得：

将结果回代L：

原问题可以转化为：

* 1. **软间隔SVM优化问题及推导优化过程**

软间隔即允许某些样本不满足要求，为此我们引入松弛变量

目标函数为：

令

对L求偏导：

令 ，得：

将结果回代L：

原问题可以转化为：

* 1. **核-SVM优化问题及推导优化过程**

目标函数为：

令

对L求偏导：

令，得：

将结果回代L：

由核方法可知：

原问题可以转化为：

* 1. **经典SVM算法的实现**

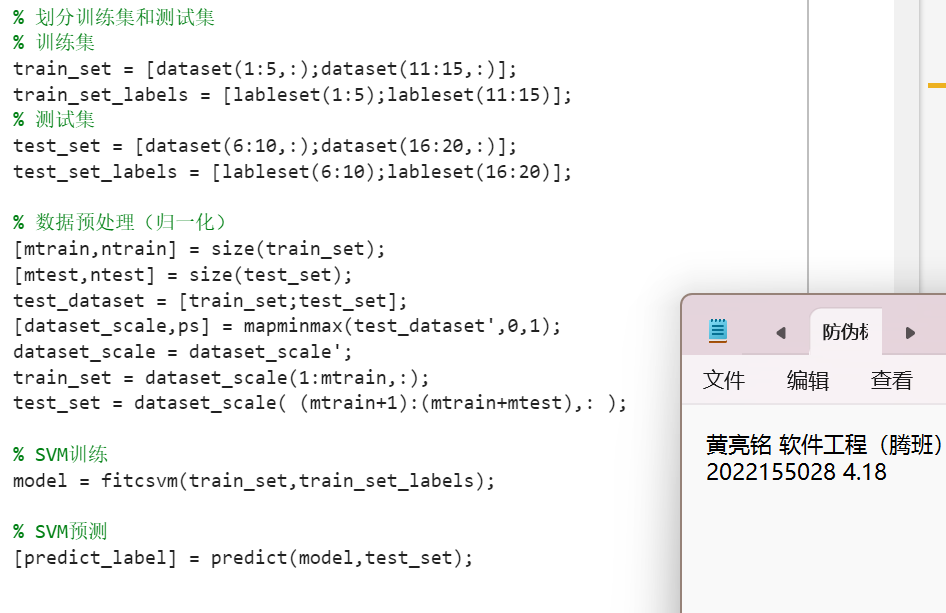


图1：经典SVM

* 1. **在二维平面对二类问题给出support vector的一个示例**

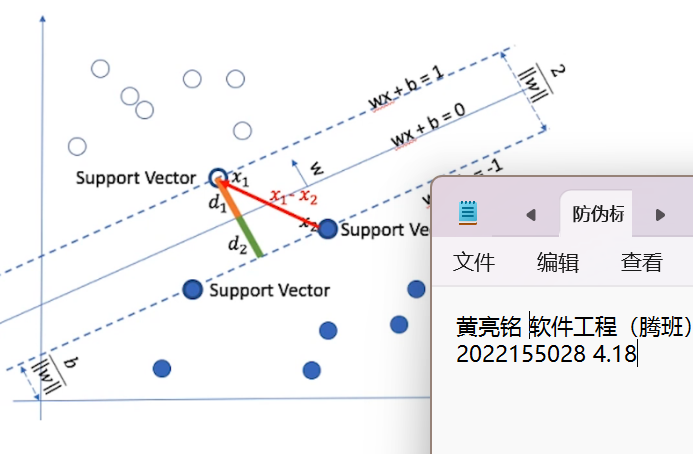


图2：支持向量示例

1. 用PCA、LDA算法提取前 10,20,30,...,160维的图像特征，然后再用不同SVM进行识别，并比较识别率。
   1. **SVM核心代码**

这里使用经典SVM进行人脸识别。对于c个类别的情况，我训练了个SVM，每个SVM用于判断任意个样本是属于c种类别中的特定的两种类别。测试时，使用个SVM做次测试，遵循多数服从少数的原则，样本被分到哪个类别的次数最多，则认为样本属于哪个类别。实现代码如下图：

****

图3：SVM核心代码

* 1. **使用PCA提取图像特征**

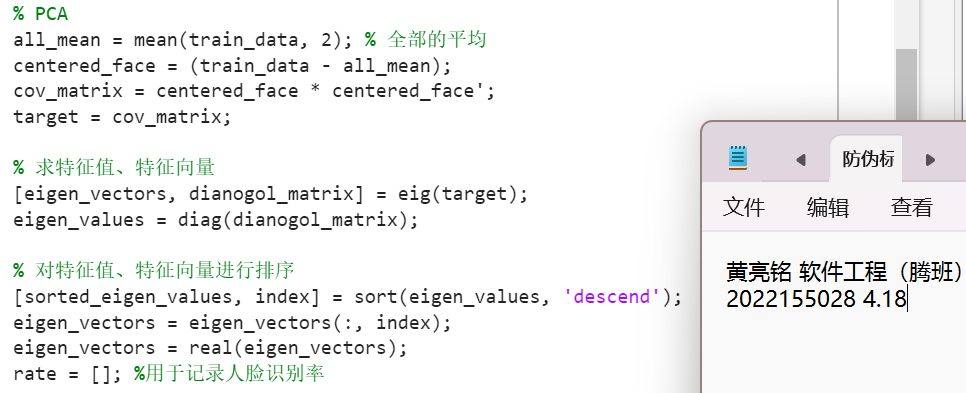
****

图4：PCA提取特征

* 1. **使用PCA提取图像特征的SVM识别率**

由图5可以得到信息：维度与识别率的关系为对数关系。维度较低时，识别率较低，这是因为从样本中获得的信息过少，导致SVM识别不准确；当维度达到60时，识别率稳定在97%。

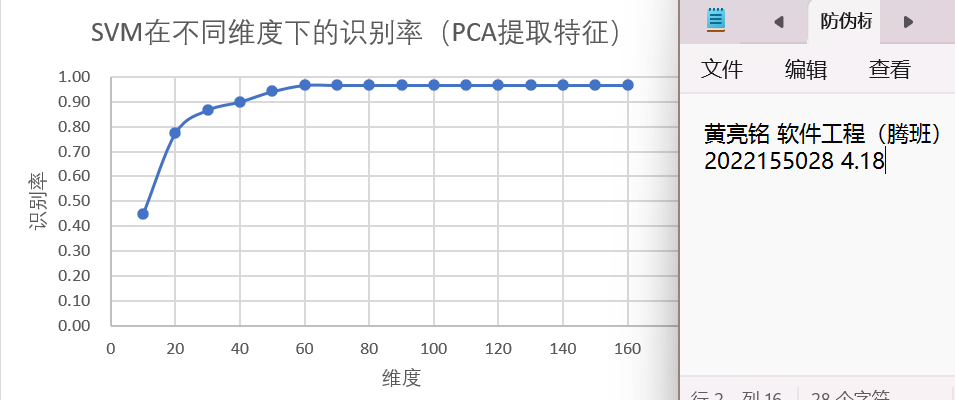
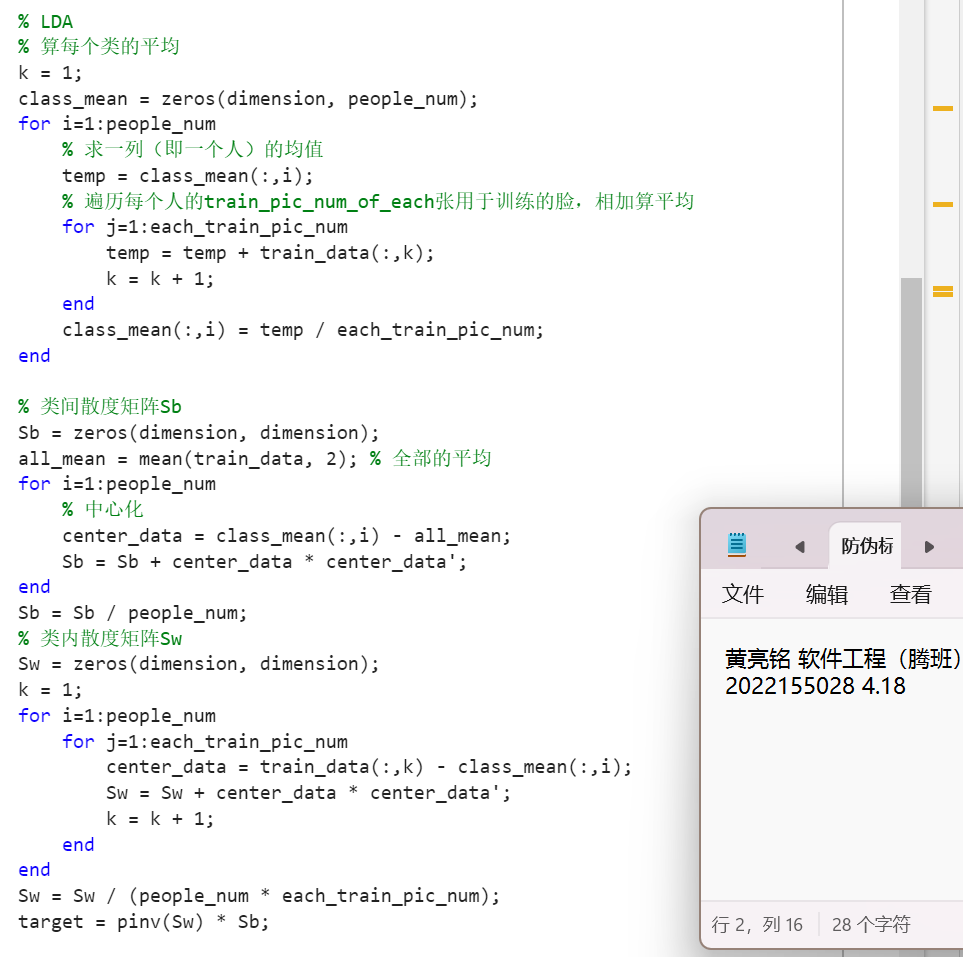


图5：PCA提取图像特征的SVM识别率

* 1. **使用LDA提取图像特征**

****

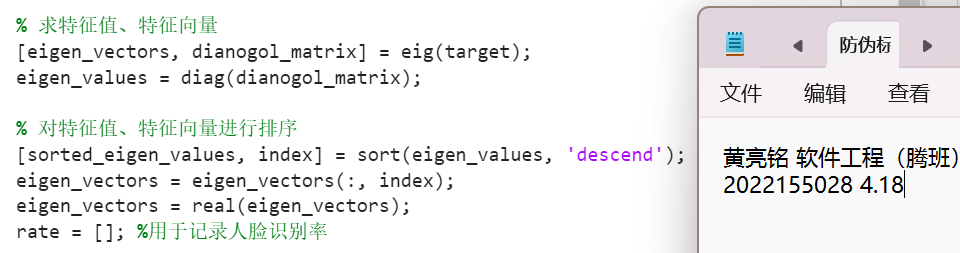


图6：LDA提取特征

* 1. **使用LDA提取图像特征的SVM识别率**

由图5可以得到信息：维度与识别率的关系为对数关系。维度较低时，识别率较低，这是因为从样本中获得的信息过少，导致SVM识别不准确；当维度达到60时，识别率稳定在98%。

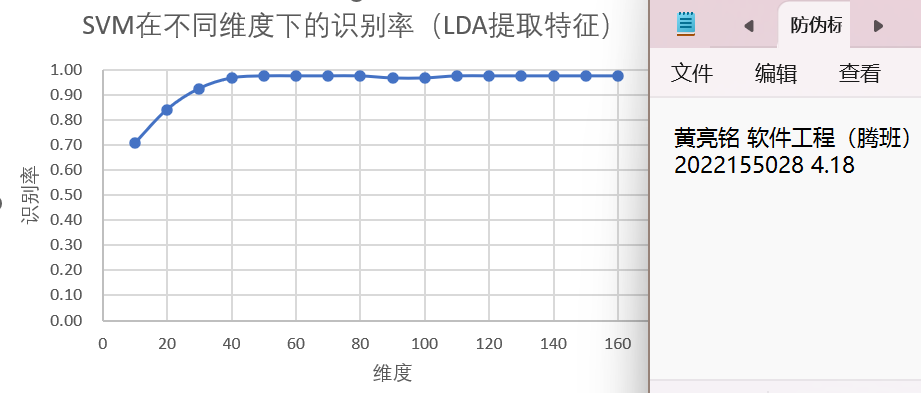


图7：LDA提取图像特征的SVM识别率

* 1. **小结**

由图（图8）我们可以得到信息：① 维度和PCA提取特征的SVM识别率、LDA提取特征的SVM识别率在图像上均呈现对数关系；② 在维度较低时，LDA提取特征的SVM识别率优于PCA提取特征的SVM识别率；③ 在维度较高时，LDA提取特征的SVM识别率也略高于PCA提取特征的SVM识别率。

综上，如果只看识别率，LDA提取特征的方法优于PCA提取特征的方法，但是LDA在处理高维数据的计算代价非常大。事实上，执行程序的所用时间也体现了这一点。因此，选择PCA提取特征，还是选择LDA提取特征，需要根据具体情况确定。

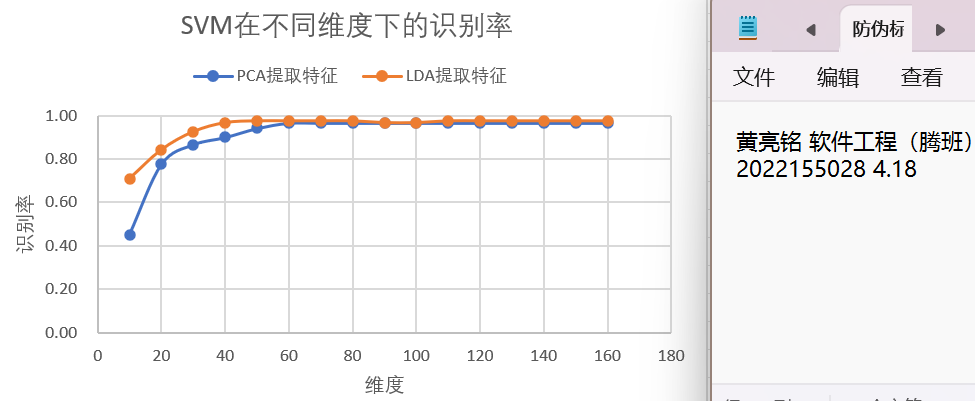
****

图8：不同提取特征方法的SVM识别率对比

# 四、实验结论或体会

1. 经典的支持向量机算法在处理线性可分问题时表现良好，但在面对线性不可分问题时效果有限。
2. 软间隔支持向量机通过引入松弛变量允许部分数据点位于间隔内部，提高了模型的鲁棒性，适用于处理一些噪声较多的情况。
3. 核-SVM通过核技巧将数据映射到高维空间中，从而处理非线性可分问题，但需要注意选择合适的核函数以及调节相关参数。
4. PCA和LDA作为常用的降维算法，能够有效地减少特征维度，提高计算效率，但在处理非线性数据时可能存在局限性。
5. 随着特征维度的增加，图像识别的性能通常会有所提高，因为更多的信息被保留下来，但也存在维度灾难的问题，可能导致过拟合。

|  |
| --- |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  年 月 日 |
| 备注： |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。