**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**  **指导教师 赖志辉**  **报 告 人 唐健龙 学号 2018132100**  **实验时间 2020年5月4日至2020年5月18日**  **实验报告提交时间 2020年5月18日** |

**教务处制**

# 一、实验目的与要求

**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SVM进行识别，并比较识别率。**

# 二、实验内容与方法

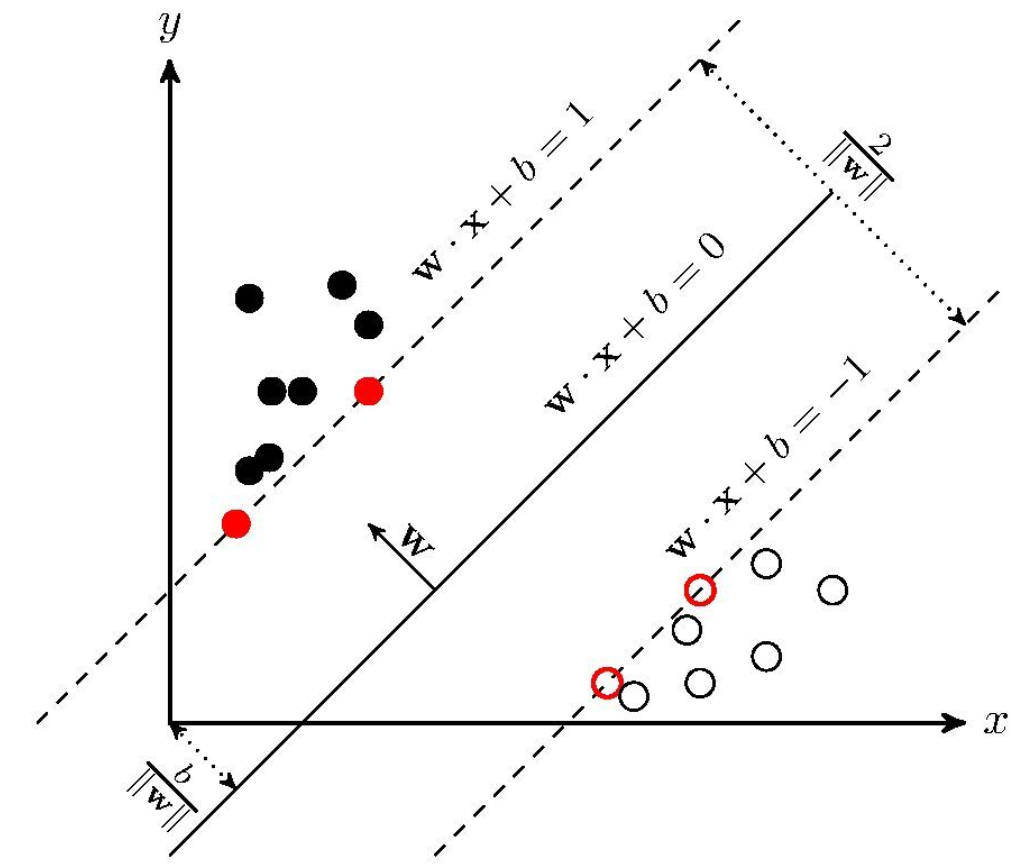
**硬间隔SVM：**

**原理：**

在二分类问题中，假设两类数据集合是线性可分的，我们可以设想存在一个超平面，能够将两类数据完全分开，且这个平面距离某一类的与之最小距离的点的距离等于另一类的与之最小距离的点的距离，这样能够保证最优化容错率的同时做到区分两类数据

**推导：**

首先假设存在数据集 ，以及数据集中的每个数据的标签的值对应 中的某个值，每个数据的标签都是1或 -1，代表不同类别，观察二维平面的实例（下图）



首先假设我们需要的超平面方程为

其中

假设当 ，，当 ，，于是有

假设超平面到两类中任意一类数据点的最小距离为 ，利用点到超平面的距离公式，我们可以得到一个不等式

进行变量替换

不等式化为

为了构造一个超平面，使得容错率最优化的同时能够完全正确地分割不同类数据，我们需要使之满足以下条件

根据上面的变量替换，我们发现 的模长恰好是 ，所以问题可以化为

显然，还可以进一步转换为更容易求的等效问题

为了将不等式约束扔掉，我们首先引入一个函数

其中的 ，且 ，现在尝试分析该函数的最大值

任取一组 ，假设

那么，令 ，就能使得

若上述假设不成立，就会有

显然，这恰好满足我们的要求，并且此时 具有最大值 ，并且在取得最大值的时候，一定有

于是，我们知道，只有满足完全正确分类数据的超平面，才会使得该函数最大值不为无穷大，我们又有

这也就是说，以下问题

恰好将问题的解限定在了可行域，但这一类问题并不好求解，首先列出该问题的对偶问题，如下所示

这是一个弱对偶问题，假设原问题的解的值为 ，对偶问题的解为 ，

因为上面的不等式最右边的表达式对 无限制，显然就会有

于是得到结论

但是我们希望 ，如此便可间接求得 的值，我们希望知道在什么情况下，才会使得 ，实际上，要使得等式取等号，需要满足两个条件

首先，从优化表达式能够知道，优化模型满足第一个条件，其次由于

这使得优化模型也满足了第二个条件，于是就有 ，也就是说，能够通过求解对偶问题以间接求解原问题

回到对偶问题

首先处理下述问题

令 对 求偏导，并使之为0，得到

代入 得到

再令 对 求偏导，并使之为0，得到

继续代入 得到

于是对偶问题化为

或者转化为求最小值

求解 可以采用坐标上升法，但是在数据量很大的时候会很慢，这个时候就需要用到SMO算法，在李航老师的《统计学习方法》一书中有证明过程和求解方法，实验过程正是使用SMO算法进行 的求解

**软间隔SVM：**

**原理：**

在大多数情况下，硬间隔并不能适应分类任务。因此引入一个容错范围，使得在允许少许错误的情况下，仍然能够令超平面有较好的分类效果

**推导：**

将硬间隔的不等式

修改为

若有少许样本点在超平面的两个边界之外，即满足 且能正常分类，那么令 ；若在两个边界之内，即虽然不满足 但是仍能正常分类，就令 ；若被误分类，则令 ，如此便能保证所有点满足不等式约束，但是相应地，对于不能正确分类的点，需要给予一定惩罚，将优化表达式约束改为

类似地，构造一个拉格朗日函数

其中 令该函数对 求导，得到

再对 求导，得到

最后对 求导，得到

因为

合并得到

可以发现，软间隔与硬间隔的差别只是多了 ，于是可以使用硬间隔的求解思路进行求解，其最终的优化模型为

**核SVM：**

**原理：**

实际情况下，还有一些样本是软间隔也无法解决的，如部分数据可能是环形的，无论如何都不能通过线性超平面进行分割，此时引入核函数将数据隐式映射入高维空间，就有可能在高维空间进行线性超平面的分割了

**推导：**

假设存在映射

且存在核函数

由于核函数内部进行内积运算并不在高维进行，因此计算速度相比将数据映射入高维再做内积运算要高效很多，并且有些情况下，映射空间可能会是无穷维空间，这时候核函数就能发挥其作用了，常用的核函数有

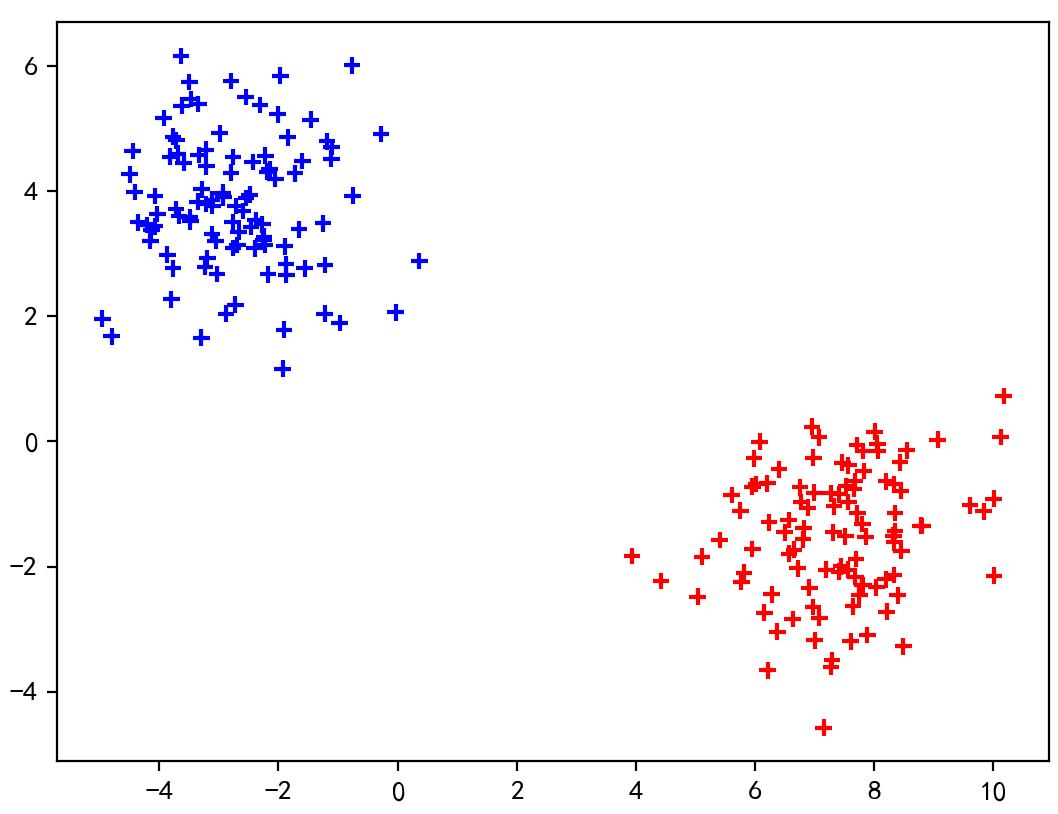
1. 线性核
2. 多项式核
3. 拉普拉斯核
4. 高斯核

容易发现，在硬间隔和软间隔中，所有内积运算表达式都对求导结果无影响，将软间隔的优化表达式中的内积运算用核函数代替后，得到

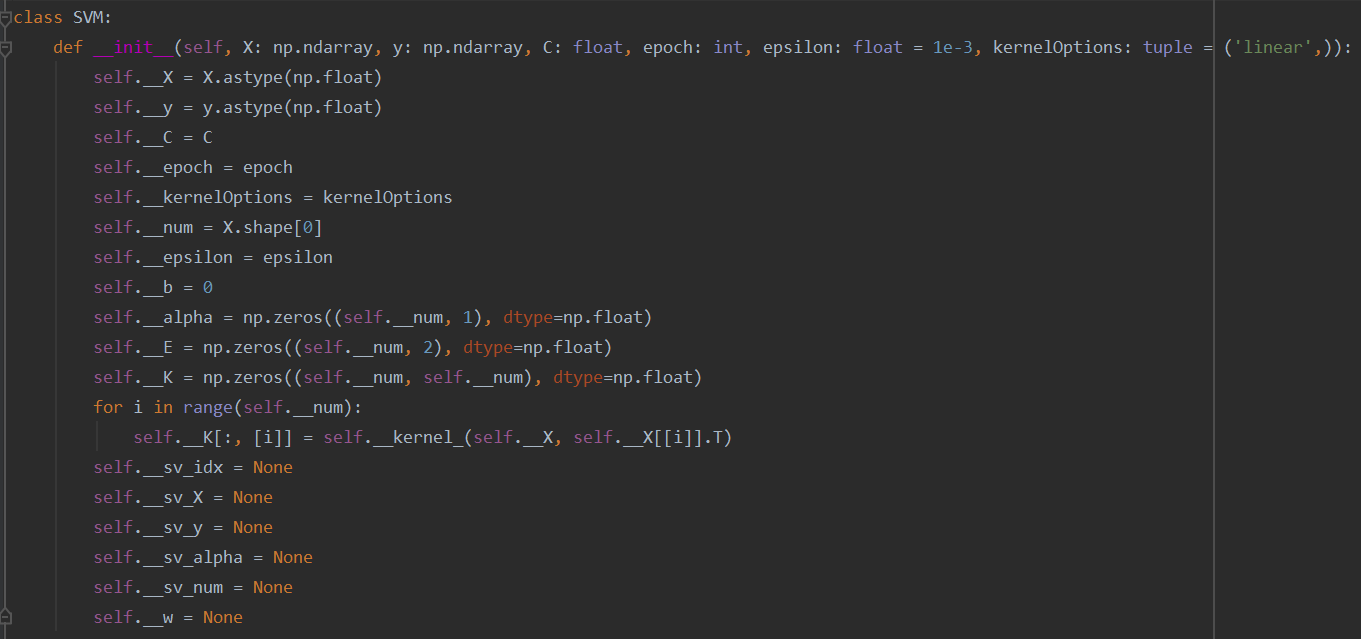
这就是同时包含软间隔和核函数的优化模型

# 三、实验步骤与过程

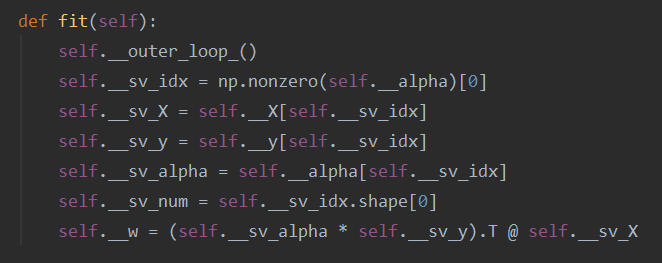
1. 为了在二维显示SVM的效果，首先引入sklearn.datasets生成两个类别的数据，如图所示



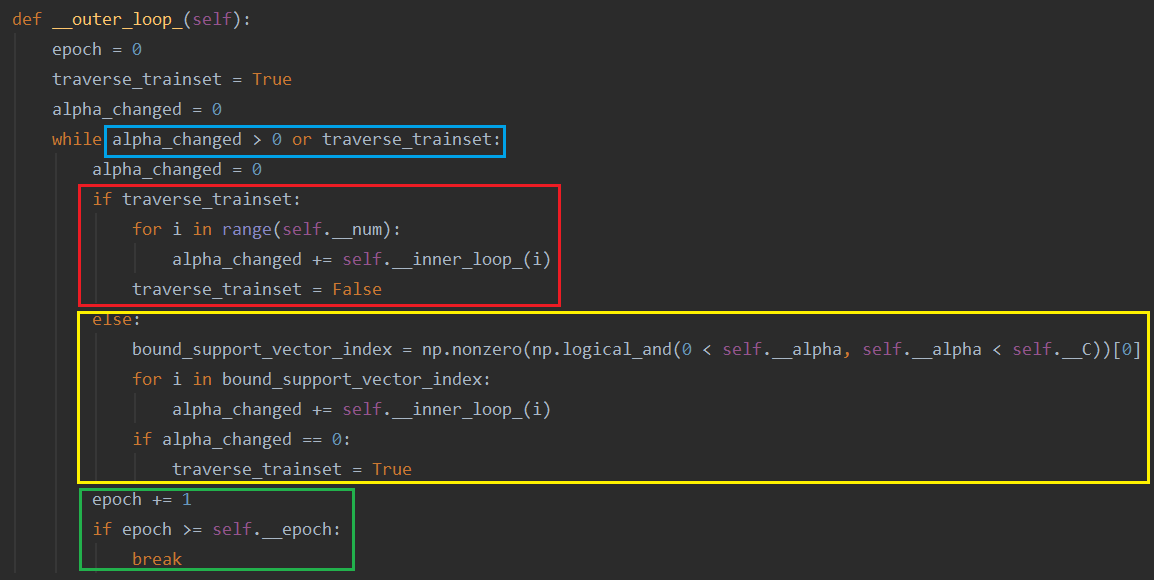
2. 首先构造一个类，能够存储建立模型的数据，以及求解中用到的参数，如图所示，在类的构造中，还进行了初始化核函数矩阵



3. 初始化类后，需要进行模型的训练和提取支持向量，如图所示，首先进行外循环，进行所有的迭代工作，随后找到所有支持向量



4. 在外循环中，有两种情况退出迭代，分别是达到最大迭代次数（绿框）和不进行全样本遍历且 不再变化（蓝框）。在循环内部，要么遍历所有样本进行内循环（红框），要么找到所有对应于 的边界点进行内循环（黄框），如图所示



5. 在内循环中，找到所有满足 且 ，或 且 的 ，随后寻找 使得 在迭代后有最大变化，在《统计学习方法》一书中，给出了 的迭代式

其中

最后再将 剪切到给定范围 内，如下所示，《统计学习方法》指出，若 ，范围为

若 ，则范围为

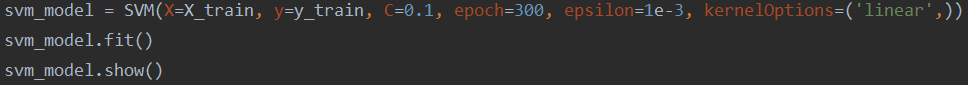
《统计学习方法》指出，在剪切 后，再通过

迭代 ，最后再通过

更新 ，最后成功迭代并返回1，否则返回0

内循环代码过长，不放出截图

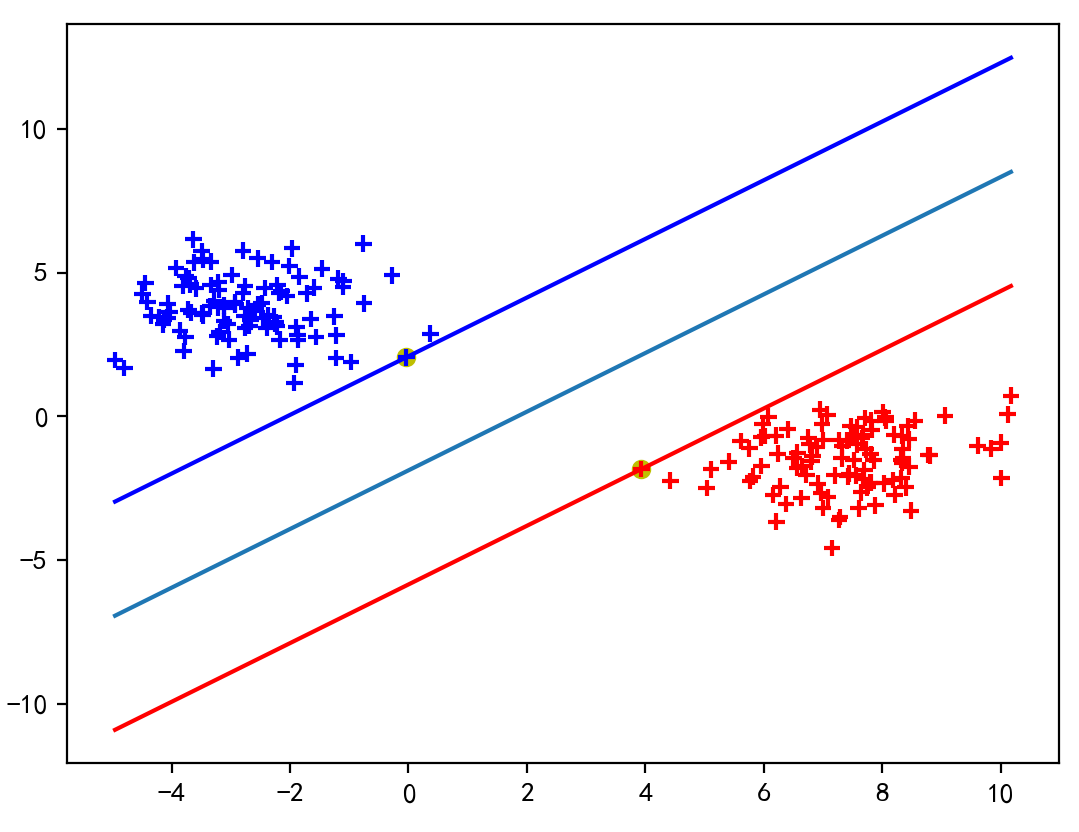
6. 至此，模型训练完毕，接下来传入数据并进行训练



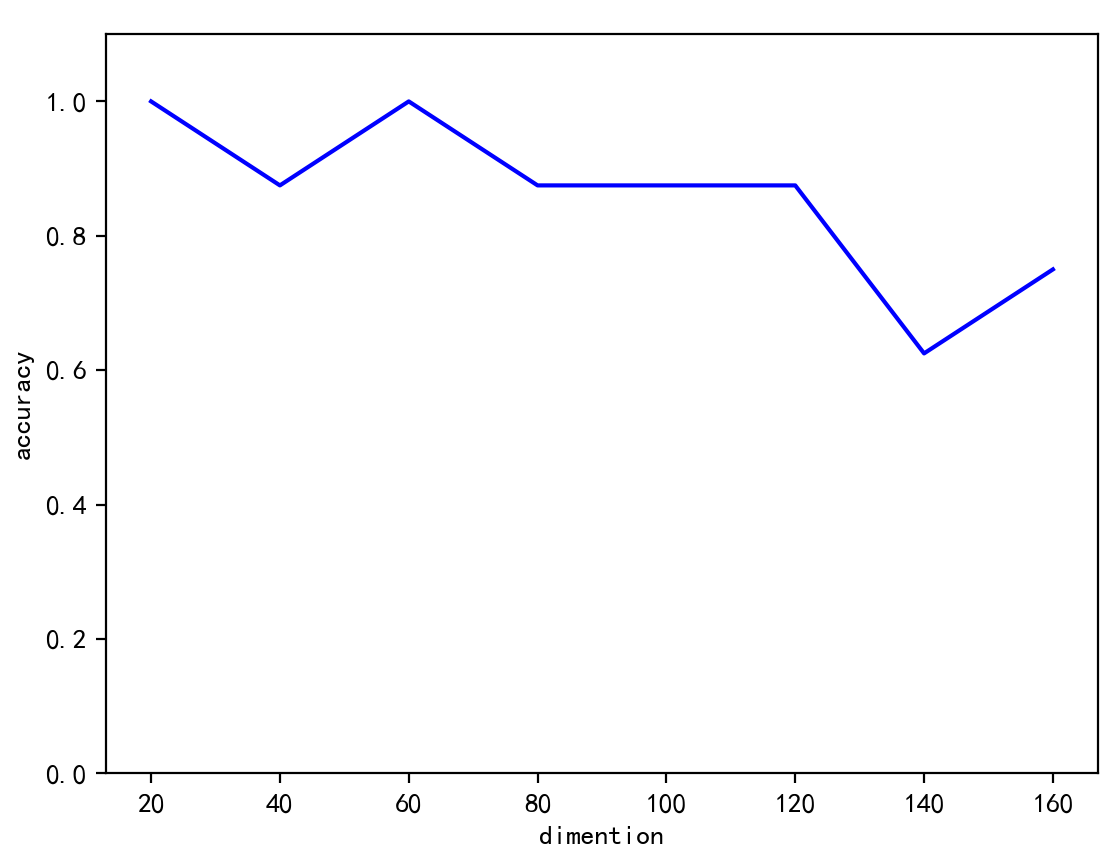
7. 训练完毕后，展示效果，在二维例子中，有

三个平面分别是超平面和边界平面，将等式变形，得到三条直线

最后将支持向量和超平面画出来，如图所示

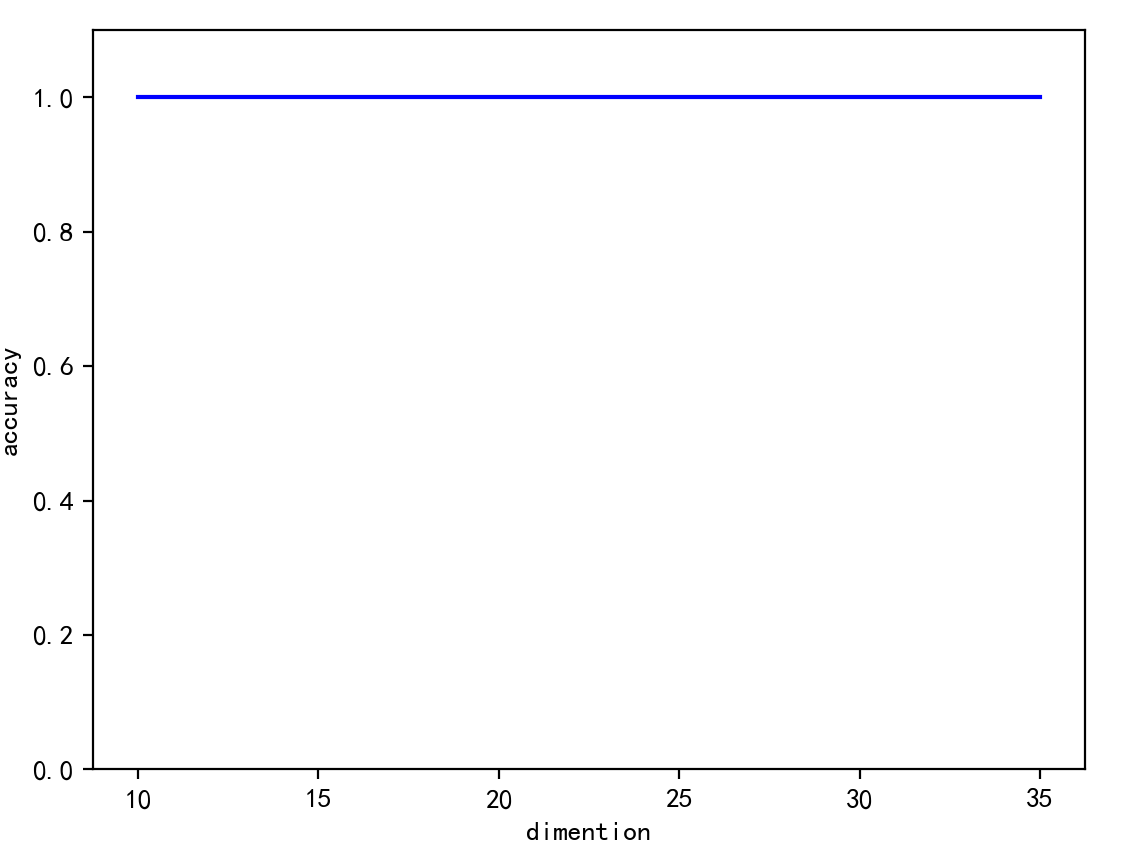


8. 接下来测试人脸识别，首先加载 ORL4646.mat文件，并随机选择某个人的4张照片，并在剩余的人中也随机选择4张，接下来每次循环都使用一次PCA降维，在维度分别降维到20,40,60,80,100,120,140,160的情况下训练SVM模型，软间隔参数设为 ，随后验证预测标签是否正确，结果如图所示



可以发现，总体趋势是，在较高维情况下，SVM的分类能力就较弱

接下来验证经过LDA处理的情况，同样随机选择某个人的4张照片，并在剩余的人中也随机选择4张，接下来每次循环都使用一次LDA降维，在维度分别降维到10, 15, 20, 25, 30, 35的情况下训练SVM模型，软间隔参数设为 ，随后验证预测标签是否正确，结果如图所示



实验结果不出所料，LDA所得结果比PCA所得结果要更好，一是因为维度更低了，二是因为LDA能够更好地将不同类别的数据分离开，利用这一特性，配合SVM能够起到很好的效果

# 四、实验结论或体会

1. SVM作为曾经机器学习领域炙手可热的算法，实际表现也非常优秀

2. SVM有多种变形与发展，包括软间隔SVM，核SVM等

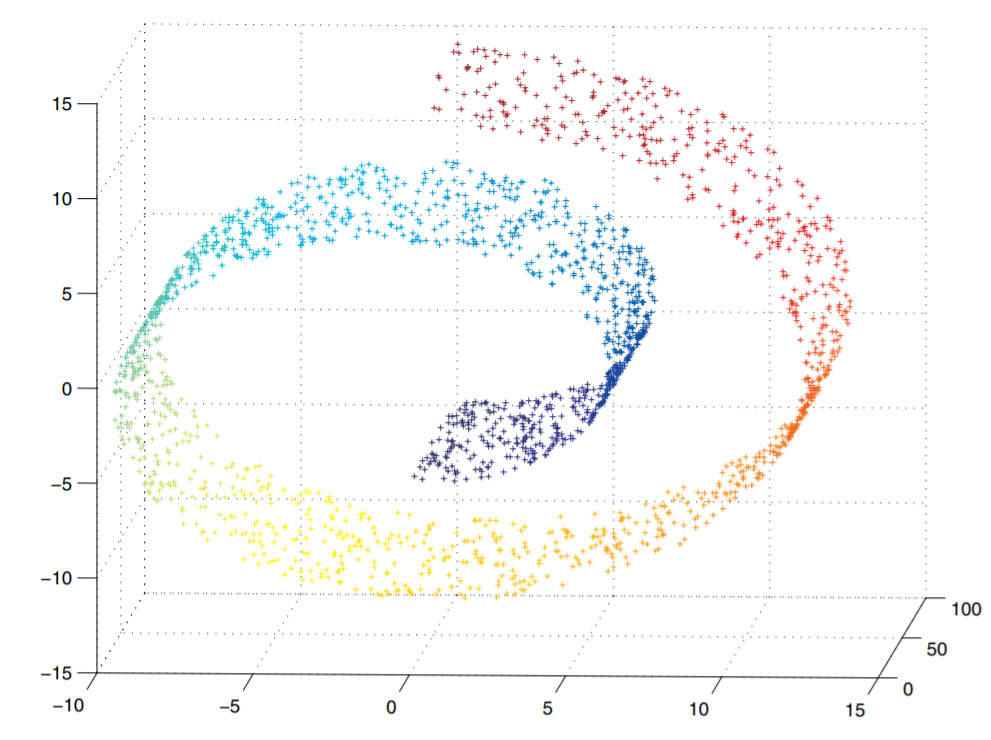
3. SVM借助其他算法更能够凸显优势

4. 与此同时，SVM的推导与计算也更加复杂

# 五、思考题

**个人想法：**

在一些复杂的非平面但局部类似平面的一些总体连续型数据集中，如瑞士卷形状的数据集



一些常用的核函数可能也无法进行较好的映射使得其线性可分，但是如果配合流形学习，如拉普拉斯特征映射（Laplacian Eigenmaps），就能将其降维至低维连续流形，从而就能相当容易对其进行超平面分割

当然，也有可能在进行流形学习算法后，仍然无法直接线性分割的情况，这种时候可以请出LDA算法和配合核函数，理论上应该能起到很好的效果

|  |
| --- |
| 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。