**《机器学习》课程小作业（三）**

**报 告**

|  |  |
| --- | --- |
| **学院** | 深圳国际研究生院 |
| **班级** | 深数据硕212班 |
| **姓名** | 陈文硕 |
| **学号** | 2021214480 |
| **日期** | 2021年10月25日 |

# 预处理

## 标准化

将训练集与测试集分别存放至project/train与project/test文件夹下

将数据进行标准化至区间

## 主成分分析

在最初的实验中发现，原训练集的样本矩阵不可逆，说明部分数据线性相关。为了让为非奇异矩阵，可以使用PCA的方法进行降维。

我们选取scikit-learn中的PCA工具包（由于作业中并没有考察相关知识，故直接调包），设置主成分个数为‘mle’，即程序会自动根据主成分的方差选取主成分的个数。最终发现，降至96维是合理的。将PCA的降维与之前手动选取特征（选取前37维特征）对比，发现前者在实验中约有2%的准确率提升。由于scikit-learn的PCA库中会自动对数据集进行标准化，故对于标准化，无需再单独编写函数。

## 归一化

实验最初遇到如下提示

ConvergenceWarning: Solver terminated early (max\_iter=1000). Consider pre-processing your data with StandardScaler or MinMaxScaler.

warnings.warn('Solver terminated early (max\_iter=%i).'

经过归一化处理后，收敛成功。

# 实验六：SVM

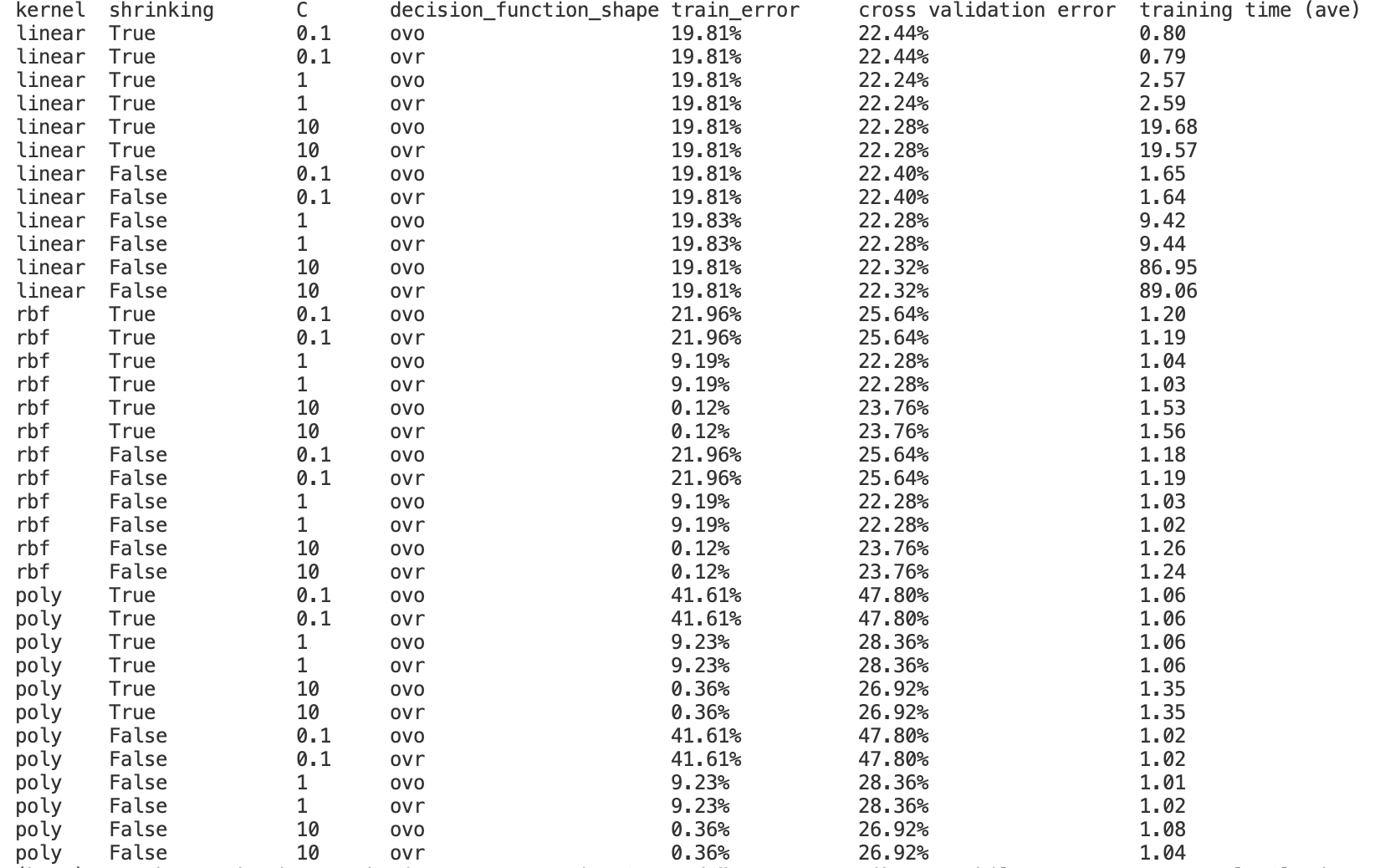
## sklearn中的参数

参数：

1. C L2正则化的参数C，必须为正数；  
   C越大，相当于惩罚松弛变量，希望松弛变量接近0，即对误分类的惩罚增大，趋向于对训练集全分对的情况，这样会出现训练集测试时准确率很高，但泛化能力弱。  
   C值小，对误分类的惩罚减小，容错能力增强，泛化能力较强。
2. kernel 可选用{‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’}, default=’rbf’；
3. degree 多项式的维度，对其他的核函数无效；
4. gamma ‘rbf’, ‘poly’ 和‘sigmoid’的参数，分为scale和auto两种选择
5. coef 核函数的常数项，默认为0，仅对poly和sigmoid有用
6. shrinking 预测哪些变量对应支持向量，加快训练速度，对结果没有影响。默认为True。
7. max\_iter 最大迭代次数，默认为无穷大
8. decision\_function\_shape ovo ovr默认为ovr

## 用linear, gaussian, polyomial和三种参数进行训练。计算训练误差和验证误差。

设置交叉验证的n\_splits=5，test\_size=0.1，统一参数对每轮交叉验证的结果取平均。



## 在TestSet-1上做测试

根据实验6.2中交叉验证的结果，我们选取参数C=1, kernel=‘linear’。 训练模型后，在TestSet-1上的正确率为79.3%。（结果保存在submissions.csv中）

## 实验小结

主要依据实验6-2的观察结果有以下结论：

1. 更复杂的模型取得的效果不一定比简单模型好，如我们采用linear核函数的效果就比高斯、多项式核函数效果好；
2. shrinking操作能智能选取局部向量进行分类，大大加快训练操作；
3. 在本次实验中，并没有观察到惩罚系数对测试准确率的影响。
4. 本实验还对比了decision\_function\_shape的参数。观察结果之前很惊讶ovo和ovr的训练时长是一样的。是由于本问题是二分类问题，所以没有差别。通常ovo在多分类问题中会比ovr慢。