统计学习Lab1 SVM

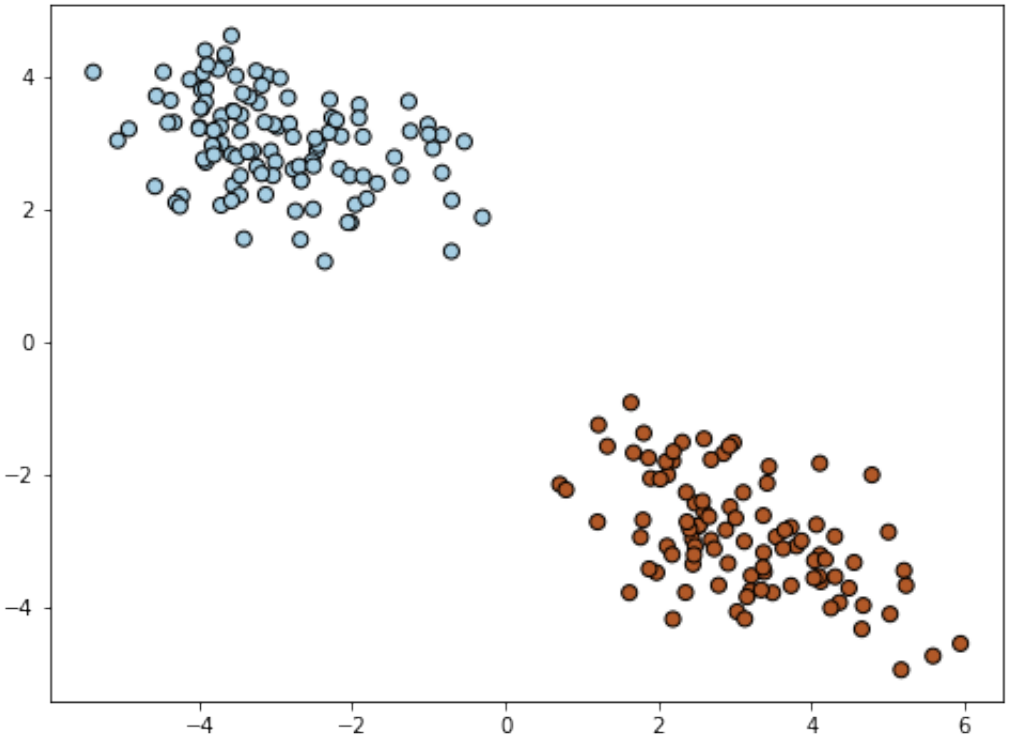
声明

1. 出现抄袭现象，抄袭双方均按零分计
2. 请严格按照deadline提交，超出每天扣除总分的20%
3. 更多问题咨询助教

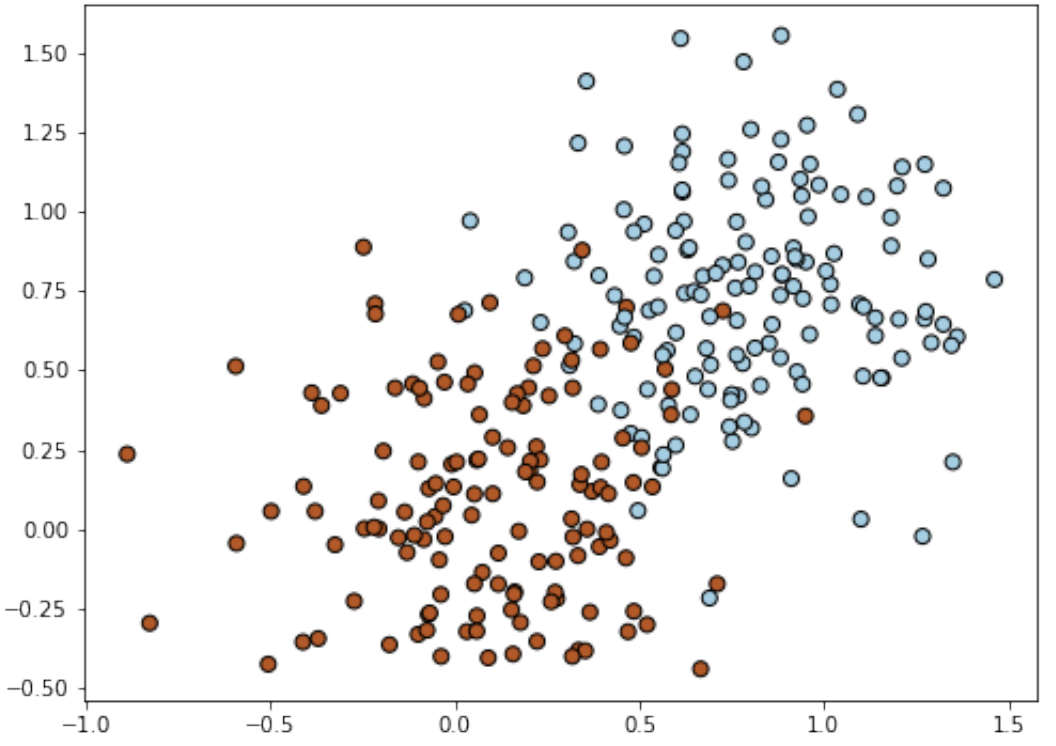
数据集说明

数据存放于`data`目录下，训练集与测试集为不同文件，测试集并不给出，大家在训练集上自行划分进行训练与验证。数据集包含三类：

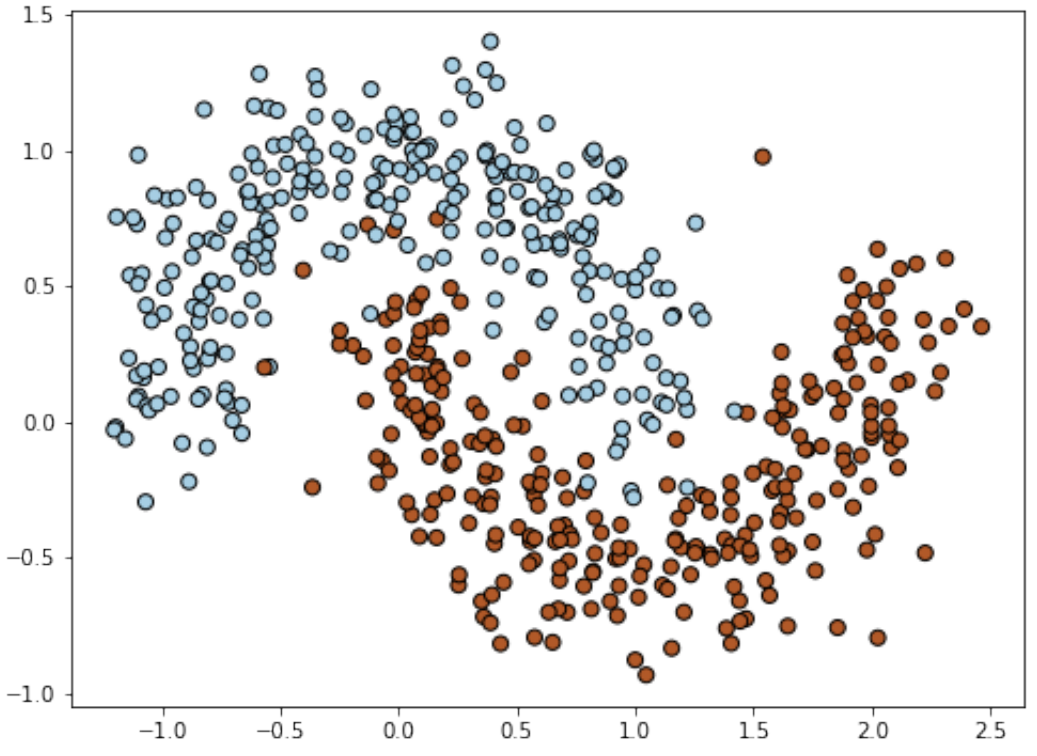
1. `train_linear.txt`：线性可分的数据



2. `train_linear_intersect.txt`：存在特异点的数据



3. `train_kernel.txt`：非线性数据



Deadline: 2024/11/30 23:59　　mzyang20@fudan.edu.cn

任务：

基于给定的不同数据集，实现SVM算法，对数据进行二分类。任务包括：

1. 实现基本的线性SVM，对线性可分的数据进行分类（3分）
2. 实现带有软间隔的线性SVM，处理带有特异点的数据（3分）
3. 使用核函数实现非线性SVM并实现SMO算法求解二次规划，对非线性数据进行分类（4分）
4. 编写实验文档，可以包括但不限于：主要代码结构、不同实验参数的比较、降低对偶问题计算复杂度的思路等（3分+2分（降低对偶问题计算复杂度的思路））

要求：

1. 使用 `python` 实现，建议使用 `numpy` 等数学运算库进行矩阵运算（否则运行速度可能很慢），任务1、2可以使用 `cvxopt` 等库进行二次规划的求解，任务3

要求使用自己实现的SMO算法进行二次规划求解。不能直接使用已经实现SVM算法的库（比如`sklearn`）。

2. 代码在给出的`svm.py`结构上进行实现，每个实验只需要实现`SVM`类的三个方法即可，读入数据集的路径可以自由修改，不需要修改其他部分的代码（比如读入数据的函数、计算分类准确率的函数）。

3. 用于参考的测试准确率（基于`sklearn`，大家不需要达到这个准确率，只作为参考，但如果在验证集上的准确率与参考差距过大可能是代码实现存在问题，实验的数据集是非常简单的）：

linear	linear intersect	kernel
100%	91.7%	97.0%

4. 文档要求工整、详实、美观，格式为pdf。