机器学习第三次作业
——支持向量机（SVM）

2052902 韩意

一、 问题描述

试使用 LIBSVM，在西瓜数据集 3.0α 上分别用不同 γ 参数的 2 次多项式核 $K(x_i, x_j) = (\gamma \cdot x_i \cdot x_j)^2$ 训练一个 SVM，比较他们支持向量的差别，最后推荐一个合适的 γ 参数并说明理由。

表 4.5 西瓜数据集 3.0α

编号	密度	含糖率	好瓜
1	0.697	0.460	是
2	0.774	0.376	是
3	0.634	0.264	是
4	0.608	0.318	是
5	0.556	0.215	是
6	0.403	0.237	是
7	0.481	0.149	是
8	0.437	0.211	是
9	0.666	0.091	否
10	0.243	0.267	否
11	0.245	0.057	否
12	0.343	0.099	否
13	0.639	0.161	否
14	0.657	0.198	否
15	0.360	0.370	否
16	0.593	0.042	否
17	0.719	0.103	否

二、 算法实现

本题使用 Python 实现，用到的库包括 NumPy、Matplotlib、LIBSVM、tkinter，其中 LIBSVM 用于训练支持向量机模型，tkinter 用于展示交互 GUI，以便实时调整 γ 参数训练不同的 SVM。

LIBSVM 库在 Python 上可以直接通过以下指令安装：

```
pip install -U libsvm-official
```

随后在 Python 中按照如下方式调用：

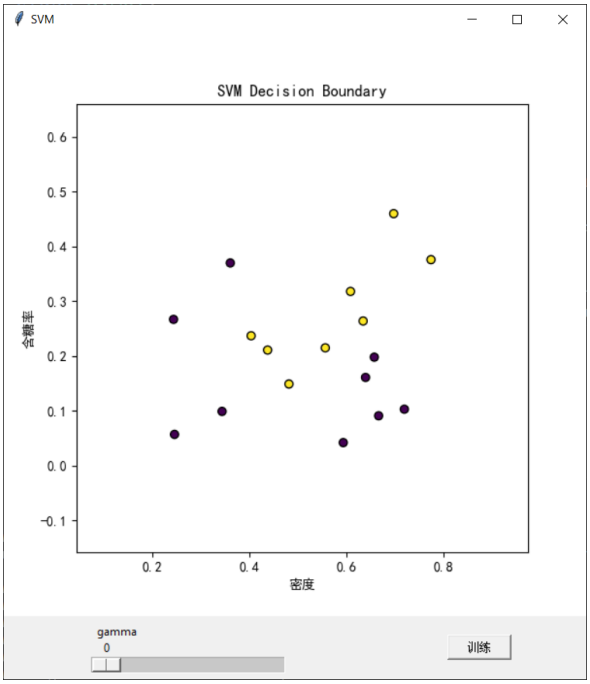
```
from libsvm.svmutil import *
```

根据题意，要求选用不同γ参数的 2 次多项式核，对应 svm_train 的参数为：

```
-t 1 -d 2 -g <gamma>
```

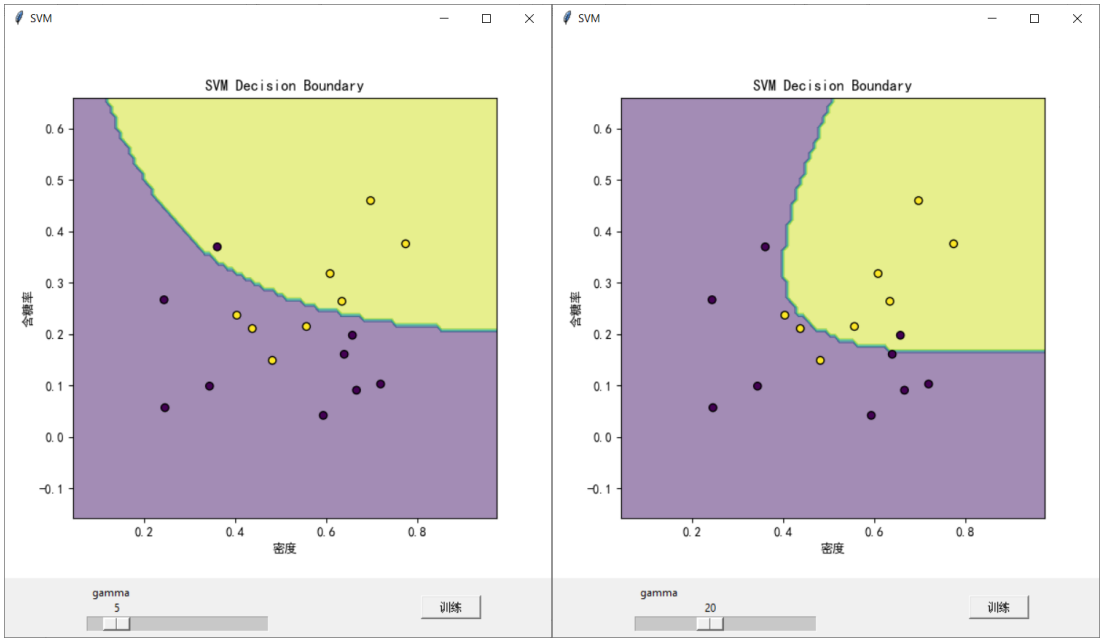
这里`-t 1`表示多项式核，`-d 2`表示次数为 2，`-g <gamma>`即表示多项式核的 γ 参数。

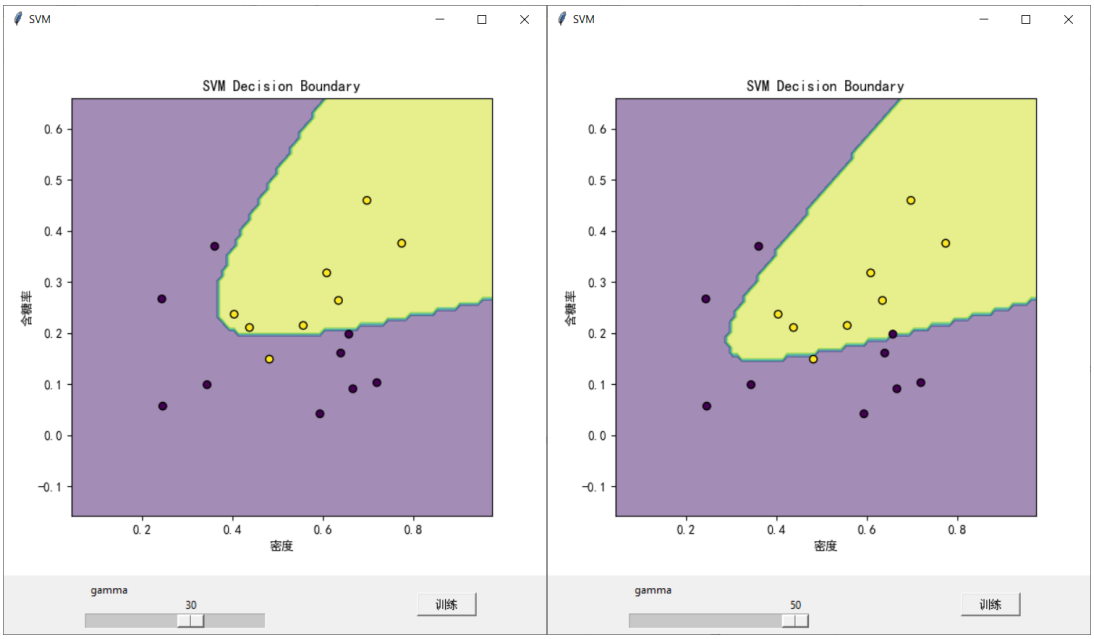
西瓜数据集散点图如下：



这里，黄点表示好瓜，紫点表示坏瓜。左下角滑动条可以拖动，得到不同的γ值，接下来点击训练即可得到不同的 SVM 训练结果。

γ 参数对模型的复杂度有很大影响。γ 值越大，模型越复杂，可能会导致过拟合；γ 值越小，模型越简单，可能会导致欠拟合。因此，选择合适的 γ 值对模型的性能至关重要。下面列出几组不同 γ 值的 SVM 训练结果，黄色区域代表好瓜，紫色区域代表坏瓜：





由此可见，随着 γ 的增大，模型复杂程度增加，非线性程度增加，对于训练集的划分越来越好，同时过拟合程度也越来越大。综合考虑过拟合与欠拟合的影响，最终选择 $\gamma = 35$ 作为针对该数据集的 γ 参数。