作业 3　支持向量机

1752931 胡斌

- **作业内容**

本作业使用 PYTHON 语言和 libsvm 软件包完成，实现了根据给定数据集，训练出一个支持向量机模型，以进行分类任务。

根据支持向量机的理论知识，普通的支持向量机只能较好地解决线性可分性较强的问题。为了解决更多的问题，可以使用核方法，引入核函数，将数据映射到更高维的空间，然后在这一更高维的空间中进行分类器的训练。

本作业使用线性核与高斯核两种核函数，分别根据西瓜数据集 3.0α（教材 89 页表 4.5）训练出支持向量机模型，并对得到的支持向量和支持向量机的结果进行分析。

➢ 作业说明

本作业的程序逻辑相对清晰，在代码中有充分注释，能够比较容易地看懂。下面列出比较关键的点或在完成过程中遇到的一些问题进行说明。

- 数据集的格式

为了使编写的程序有更好的移植性，本作业首先将使用的数据集写入到一个 excel 表格中，然后通过 python 程序从 excel 表格中读取相关的数据。这样的方法没有把数据集写死在程序中，因此当训练使用的数据集改变时，只需使用新的 excel 表格即可，有较强的移植性。

为此，需要先编写一段将 excel 表格中的数据转换为 libsvm 规定的数据格式。本作业中的代码实现了这一功能，其实现的效果如图所示。图 1 是 excel 中的训练集数据，图 2 是转换后的满足 libsvm 要求的数据集格式。

	A	B	C	D	E
1	编号	密度	含糖率	好瓜	
2	1	0.697	0.46	1	
3	2	0.774	0.376	1	
4	3	0.634	0.264	1	
5	4	0.608	0.318	1	
6	5	0.556	0.215	1	
7	6	0.403	0.237	1	
8	7	0.481	0.149	1	
9	8	0.437	0.211	1	
10	9	0.666	0.091	-1	
11	10	0.243	0.267	-1	
12	11	0.245	0.057	-1	
13	12	0.343	0.099	-1	
14	13	0.639	0.161	-1	
15	14	0.657	0.198	-1	
16	15	0.36	0.37	-1	
17	16	0.593	0.042	-1	
18	17	0.719	0.103	-1	
19					

图 1 excel 中的训练集数据

```
the category of each sample:
 [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, -1.0, -1.0, -1.0, -1.0, -1.0, -1.0, -1.0, -1.0, -1.0]
the values of attributes of each sample:
 [{1: 0.697, 2: 0.46}, {1: 0.774, 2: 0.376}, {1: 0.634, 2: 0.264}, {1: 0.608, 2: 0.318}, {1: 0.556, 2: 0.215},
```

图 2 转换后的训练集数据（部分数据）

- **效果分析**

本作业使用了线性核和高斯核两种核函数进行支持向量机的训练。一般来说，线性核能够较好的解决线性可分问题，但对线性不可分问题的效果较差。为了解决线性不可分问题，需要使用更复杂的核函数，将数据映射到更高维的空间，再使用支持向量机进行线性分类。高斯核就是一种常用的核函数。

首先将本作业使用的实验数据进行可视化，如图 3 所示。可以看到，该问题的线性可分性较差，所以使用线性核的效果可能较差，而使用高斯核的效果应该会相对更好。

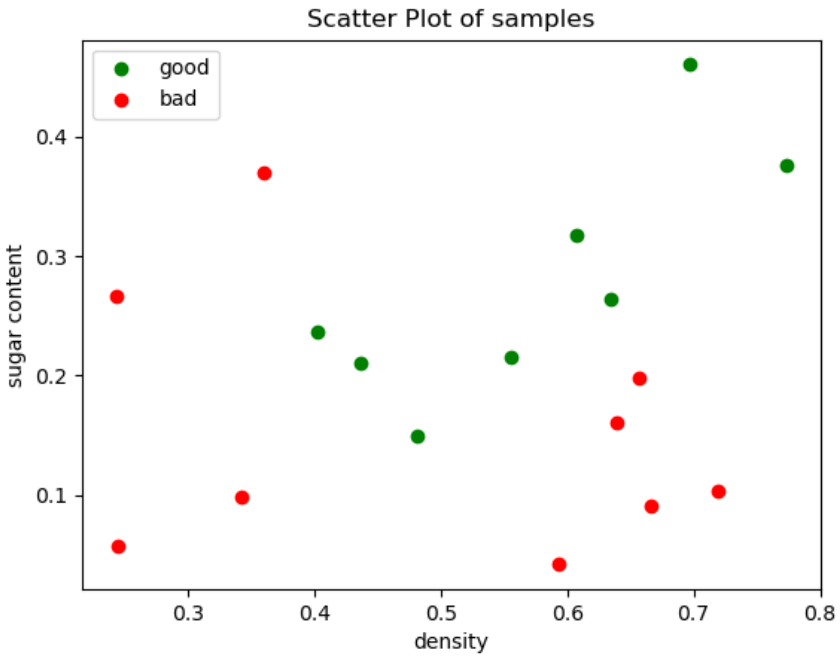


图 3 样本数据

在高斯核中，还可以设置不同的带宽，得到不同的具体的高斯核函数。在 libsvm 中，可以通过调整 gamma 参数来调整高斯核的带宽。带宽的选取没有固定可靠的办法，在实际工程中一般使用经验和试错

法来确定。按照参数调整的一般做法，本作业对几种不同数量级的gamma参数分别进行了实验，并观察训练后支持向量个数的情况。实验的结果见表1。实验中使用的模型都是软间隔的支持向量机，cost参数均设为100（cost参数即教材130页6.29式中的C，C是一种惩罚损失，不宜过小）。

核函数类型	参数	支持向量数量	边界上支持向量数量
线性核	无	13	13
高斯核	$\gamma = 0.001$	16	16
高斯核	$\gamma = 0.01$	16	16
高斯核	$\gamma = 0.1$	14	12
高斯核	$\gamma = 1$	12	7
高斯核	$\gamma = 10$	7	0
高斯核	$\gamma = 100$	13	0

表1 不同核函数和参数的实验结果

从表1中可以看到，核函数的类型和参数对最终训练出的支持向量机有着较大的影响。使用高斯核后，在参数选取合适的情况下，可以有效减少支持向量的数量。从直观上理解，支持向量越少，表明最大间隔内部和边界上的样本点越少，其它的样本点都被“相当安全”地分到了分隔线的两侧。这表明使用高斯核函数将样本映射到高维空间后，在参数合适的条件下，问题的线性可分性得到了明显改善。

对于软间隔的支持向量机模型，支持向量共有两种可能的情况：一种是在最大间隔的边界上，即表中最右侧一列表示的内容；另一种

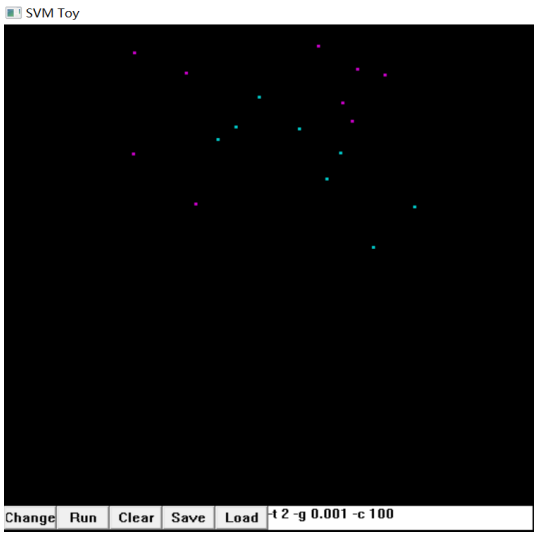
是在最大间隔内部，可能是正确分类的，也可能是被错误分类的，具体分类正确与否要看模型的另一个参数ξ（见教材 132 页相关分析）。

为了达到较好的训练效果，我们希望支持向量的个数尽可能少，同时在最大间隔中的支持向量个数也尽可能少，以减小误分类的可能性。

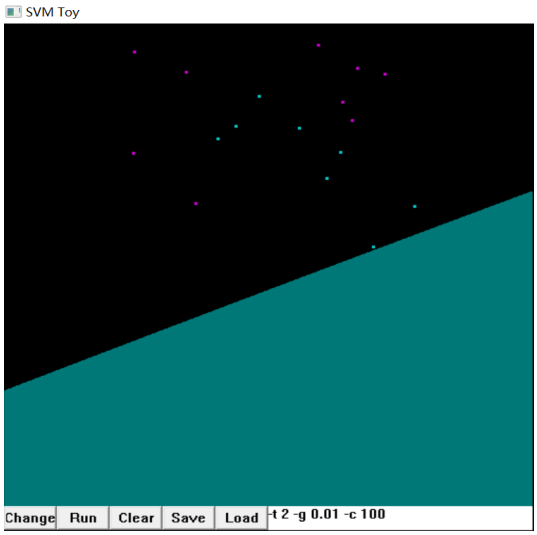
另一方面，训练集中的数据总共有 17 个，但最终的支持向量个数最少也有 7 个。通常来说，支持向量机方法具有稀疏性，即最终对模型有影响的数据应远少于原本的数据，但这一性质在本作业中没有体现出来。分析其原因，可能是因为该数据集的线性可分性本身就不够好，这一点可以从图 3 中看出。此外，本作业使用的数据集总共只有 17 个数据，数量太少，因此冗余的信息较少，所以每一数据都有较高的相对重要性，对模型有较大的影响。如果训练集的样本数量变多，支持向量机的稀疏性应该会得到更好的体现。

此外，从表 1 中还可以看到这样一个普遍规律：如果模型的支持向量数量越少，那么位于最大间隔内部的支持向量数量将有越大的趋势。这表明，随着参数的变化，模型的效果一方面变好，另一方面变差。因此调节参数的过程，就是要使两方面的影响综合后，总的效果最好，通常使用泛化的分类正确率（测试集上的正确率）来评价。

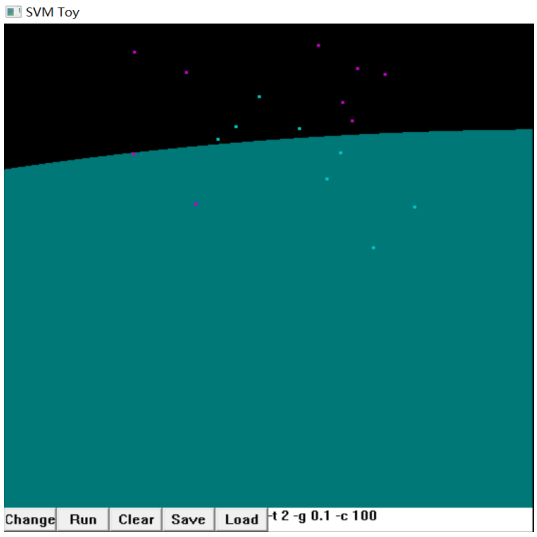
为了对结果有更直观的理解，可以使用 libsvm 中自带的 svm-toy 工具对训练的结果进行可视化，画出得到的分类边界。实验中各参数得到的支持向量机可视化结果见图 4 和图 5 所示。



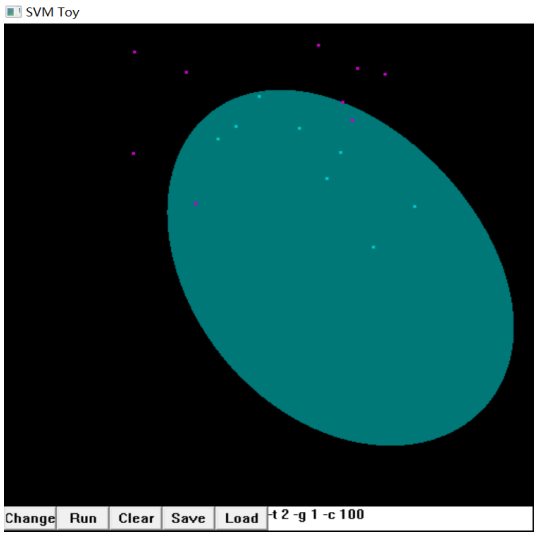
(a) $\gamma = 0.001$



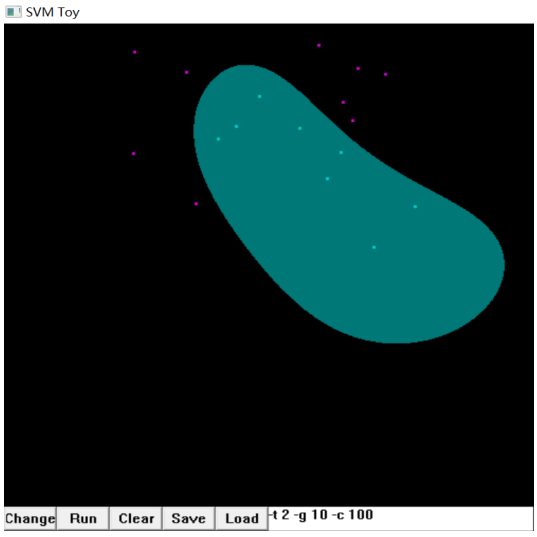
(b) $\gamma = 0.01$



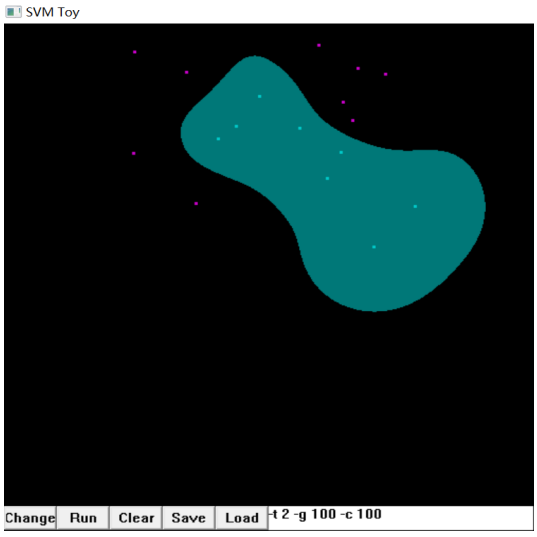
(c) $\gamma = 0.1$



(d) $\gamma = 1$



(e) $\gamma = 10$



(f) $\gamma = 100$

图 4 使用不同参数的高斯核得到的支持向量机可视化结果

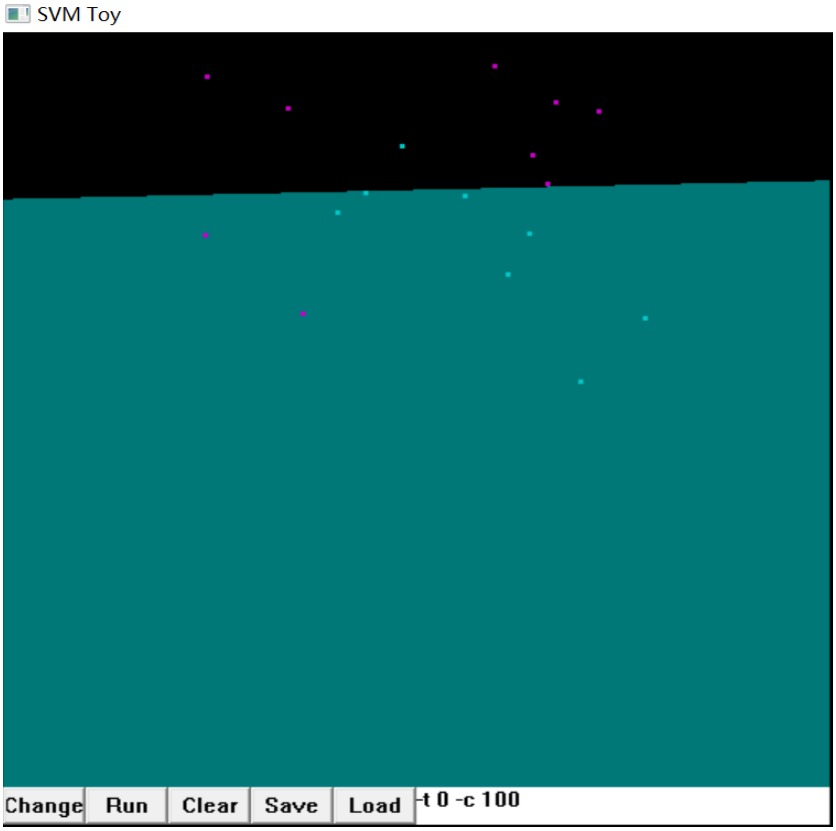
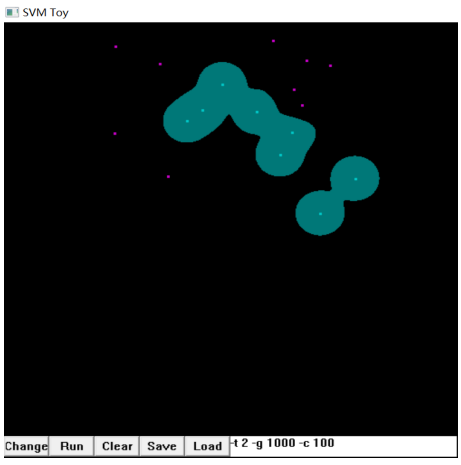
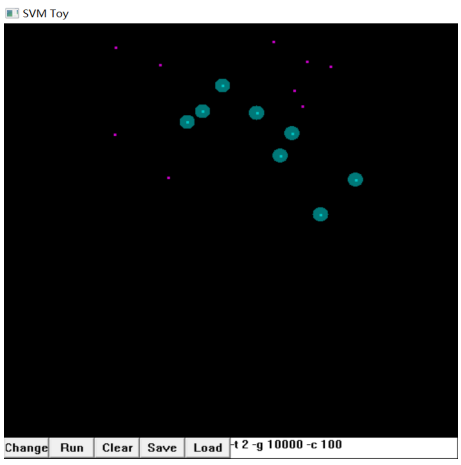


图 5 使用线性核得到的支持向量机可视化结果

从图 4 和图 5 中可以看到，使用参数合适的高斯核函数，能够得到针对本问题性能更好的支持向量机模型。从趋势上来说，参数γ越大，分类效果越好。但是需要注意的是，参数γ过大后，将出现严重的过拟合现象，模型的泛化性能严重下降，如图 6 所示。



(e) $\gamma = 1000$



(f) $\gamma = 10000$

图 6 参数γ过大导致模型严重过拟合

此外，需要注意的是，以上实验是将所有样本均作为训练集进行的，而没有测试集。理论上来讲，应该是在训练集上训练模型，然后在测试集上检验模型效果。然而，在实验中发现，样本的数量过少，总共只有 17 个，划分为训练集和测试集后，对模型的效果检验结果没有显著的统计学意义。

- **改进方向**

- 扩充数据集，使得有足够的数据来划分出训练集和测试集，进而有效地检验训练出的模型的泛化性能。
- 在合适的数量级上，对参数γ进一步优化，以得到更好的效果。
- 尝试其它核函数的效果。
- 自己编写支持向量机的训练程序。
- 自己编写支持向量机可视化的程序。
- ……