支持向量机实验报告

1. **实验目的**

使用LIBSVM，在西瓜数据集3.0α 上分别用高斯核和拉普拉斯核训练一个SVM，并比较其支持向量的差别。

1. **实验工具**

Python3.8，LIBSVM。

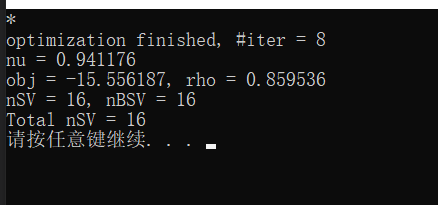
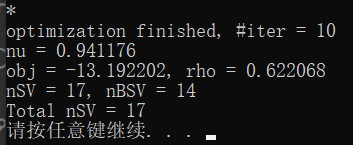
1. **重要实验代码说明**



1. **实验结果**

Sigma = 1的输出结果：

高斯核： 拉普拉斯核：

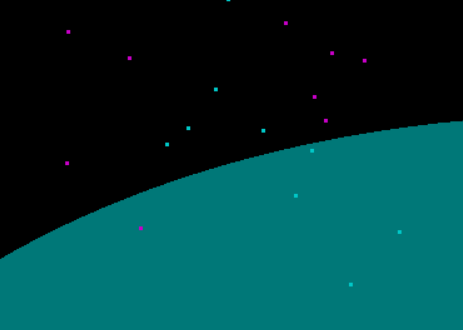
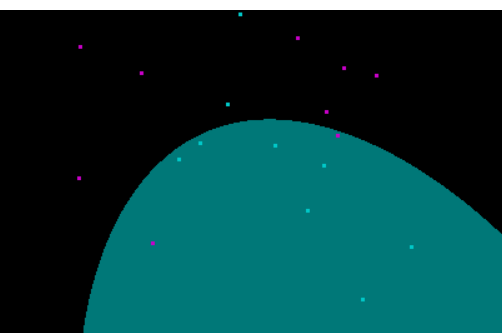
**从上述结果来看，高斯核和拉普拉斯核的支持向量个数近似相等，与样本个数（17个）也近似相同，说明很多点在软间隔范围内，分类很粗糙。因此需要调整sigma核调整cost。**

**1、不同cost下的高斯核和拉普拉斯核结果（sigma=1）**

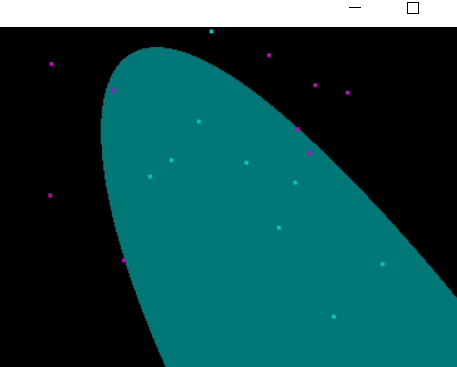
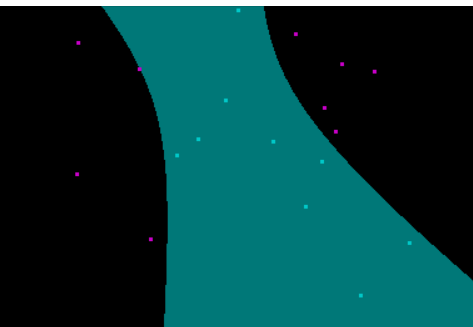
|  |  |  |  |
| --- | --- | --- | --- |
|  | 高斯核 | 拉普拉斯核 | Cost |
| 支持向量的个数 | 16 | 17 | 1 |
| 15 | 13 | 10 |
| 13 | 12 | 100 |
| 10 | 12 | 500 |
| 9 | 12 | 1000 |
| 9 | 12 | 2000 |
| 7 | 12 | 4000 |
| 7 | 12 | 8000 |
| 6 | 12 | 10000 |

高斯核的结果展示：

Cost = 10： cost = 100

Cost = 1000 cost = 4000

从结果来看：

1、cost变大，两个核对样本的分类效果变好，但cost过大容易出现过拟合的现象。

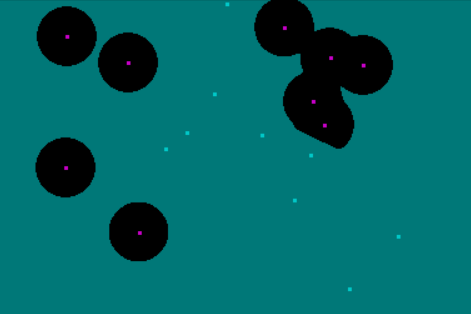
2、cost变大过程中，高斯核的支持向量个数在不断下将，但拉普拉斯核的支持向量数在cost小于100时下载速度和高斯相同，但随后便稳定在了12个，而高斯核还在不断下降。这说明了当cost大于100时，拉普拉斯核分类的边界没有太大变动。可能的原因有两个，一是拉普拉斯核相较与高斯核，不容易过拟合；二是训练集数据使用拉普拉斯核效果较好，两类样本点能够较好的区分开，因此边界可以稳定。

**2、不同cost下的高斯核和拉普拉斯核结果（cost=1000）**

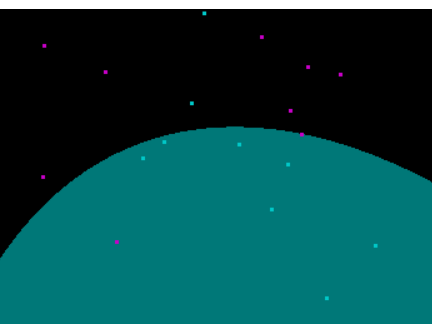
|  |  |  |  |
| --- | --- | --- | --- |
|  | 高斯核 | 拉普拉斯核 | Sigma |
| 支持向量的个数 | 17 | 17 | 0.01 |
| 16 | 17 | 0.05 |
| 10 | 17 | 0.1 |
| 6 | 12 | 0.5 |
| 9 | 12 | 1 |
| 13 | 12 | 2 |
| 13 | 12 | 5 |
| 16 | 12 | 25 |
| 16 | 12 | 50 |

高斯核的结果展示：

Sigma = 0.01 sigma = 0.1

Sigma = 2 sigma = 5

对拉普拉斯核的进一步测试

|  |  |  |
| --- | --- | --- |
|  | 拉普拉斯核 | Sigma |
| 支持向量的个数 | 16 | 0.2 |
| 15 | 0.25 |
| 13 | 0.3 |
| 13 | 100 |
| 15 | 150 |
| 16 | 200 |
| 16 | 500 |
| 17 | 1000 |

从结果上看：

1、当sigma较小时出现极为严重的过拟合现象，而当sigma较大时，拟合效果很差。

原因如下：

当sigma趋于0时，核函数（高斯和拉普拉斯）都趋于0，所有映射后的点彼此之间的距离均相等，不存在聚类现象，这样一来每个样本点都被单独分类，于是出现了很多小圈圈。

当sigma趋于无穷时，核函数（高斯和拉普拉斯）都趋于1，两个不同的点经过映射后，成为高维空间上的同一个点（相互之间距离为0）。这样一来，所有的样本点将被划分成同一个类，无法区分开来。

2、sigma由小变大时，高斯核支持向量的个数变化较为平滑，而拉普拉斯核在sigma较小时出现急剧下降，然后保存稳定，当sigma很大时才逐渐上升，这也许能进一步说明拉普拉斯核有较好的过拟合抗性。而当sigma较大时缓慢上升说明欠拟合抗性也比高斯核好。

1. **结论**
2. **从以上两个测试可以看出，拉普拉斯核比高斯核有更好的过拟合和欠拟合抗性，得到的分类超平面对sigma和cost参数变化不敏感。在实际应用中可以较快的调节到合适的参数。但另一个方面就是拉普拉斯核上限低，经过调节参数，高斯核的支持向量个数可以降到很低，即分类效果优秀，拉普拉斯核无法做到这一点。**
3. **从查阅到的资料都是高斯核，LIBSVM无拉普拉斯核的情况看，高斯核分类效果上限高的特点在使其在实际应用中更为广泛。**

**六、改进**

**1、由于拉普拉斯核没有进行可视化，得到的结论也有可能出现偏差，进一步的实验可以考虑使用sklearn库进行拉普拉斯核的可视化。**

**2、实验样本较少，可能结论不够准确，同时也无法进行测试集检验，接下来可以使用其他样本数较大的样本集进行试验。**