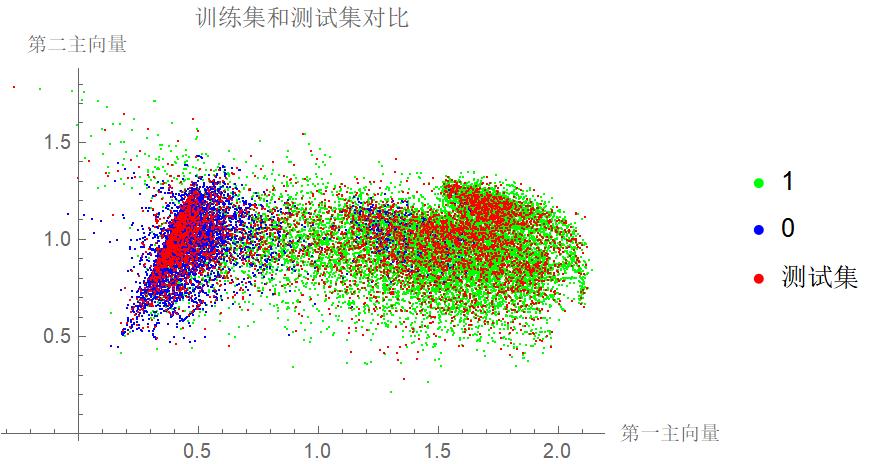
第三题是分类问题。

## 数据降维

我们首先要将测试集按照（对训练集进行主成分分析得到的主向量）展开。将测试集映射到主向量张成的特征空间中。



可以看到，测试集的分布范围。

我们决定采用支持向量机的模型进行分类。

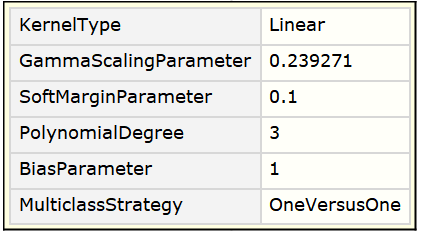
在这个问题中使用支持向量机的好处在于：

* 算法原理简单直观
* 仅仅使用一部分支持向量来做超平面的决策，无需依赖全部数据。
* 有大量的核函数可以使用

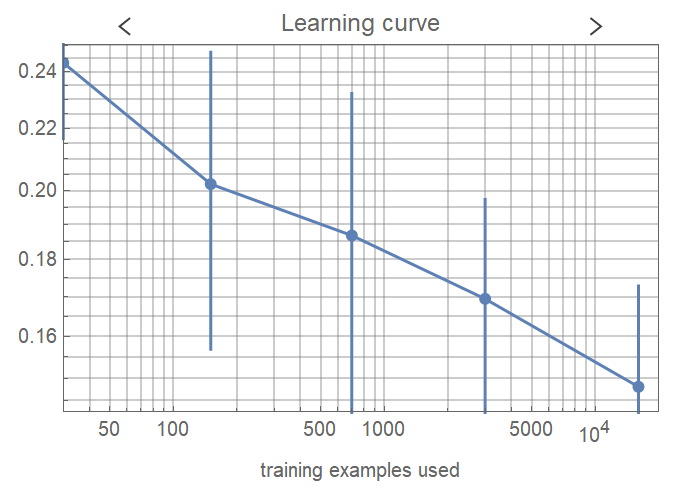
## 线性核

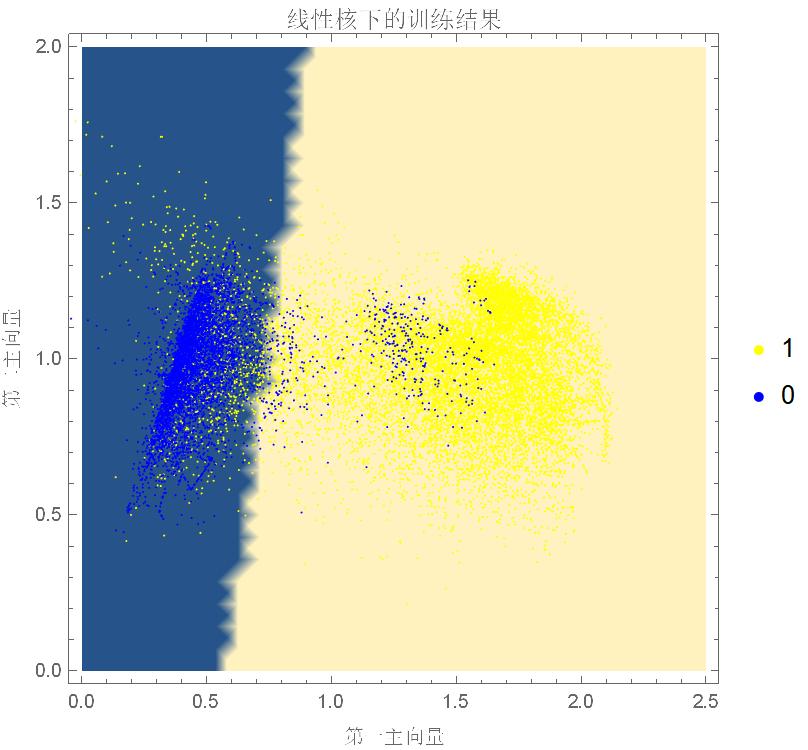
我们首先采用线性核进行分类。

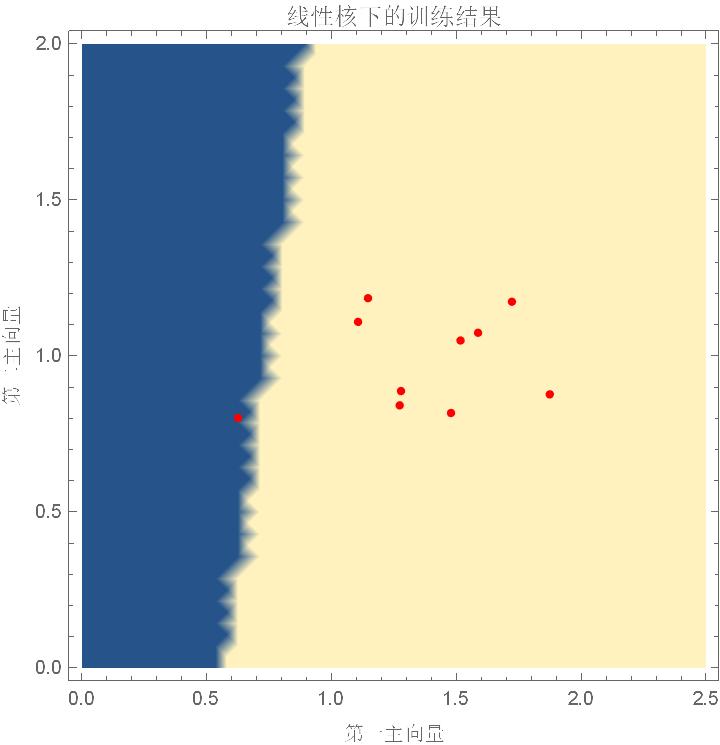
分类器的信息如下：



学习曲线：



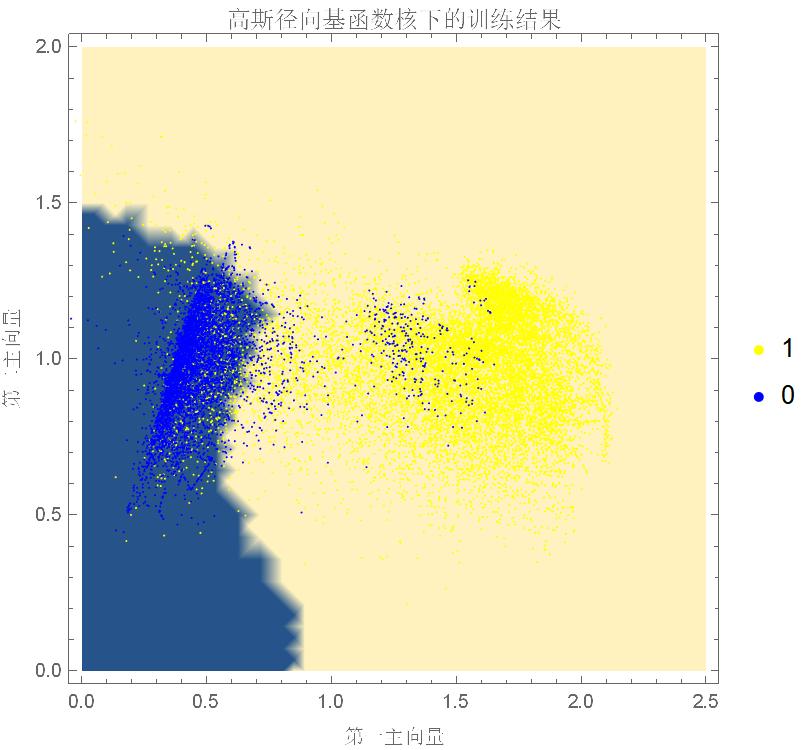


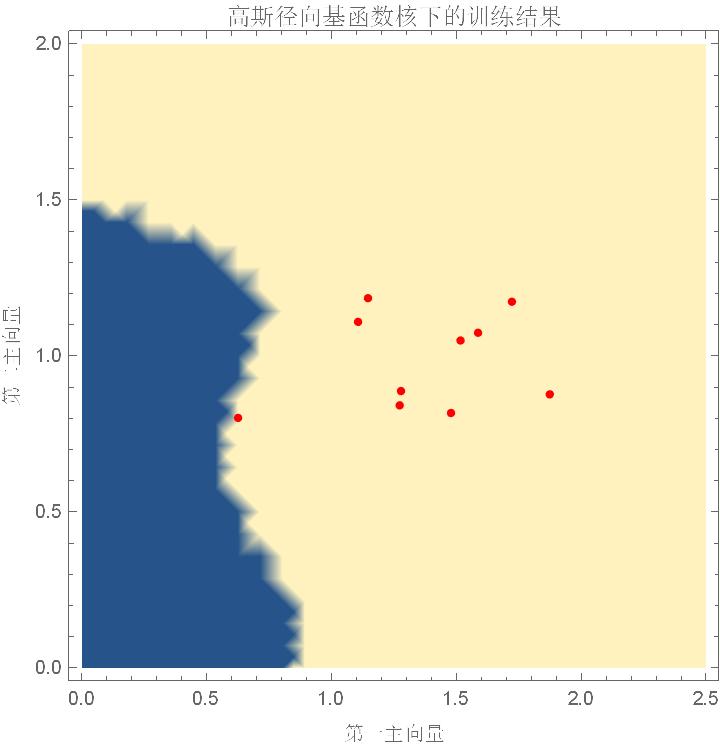


前10个数据在线性核下的分类结果为，

{1, 1, 1, 1, 0, 1, 1, 1, 1, 1}

## RBF核

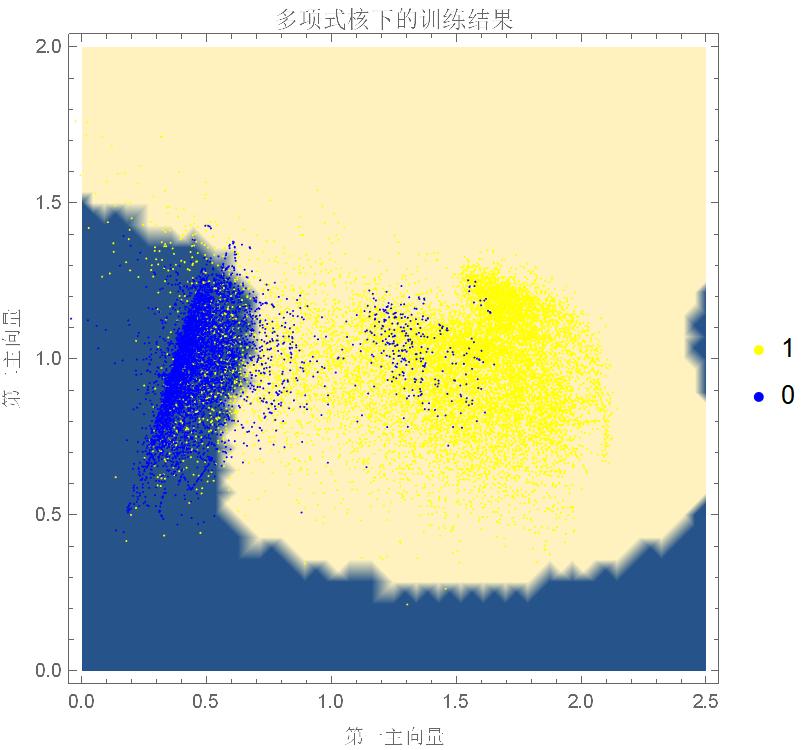


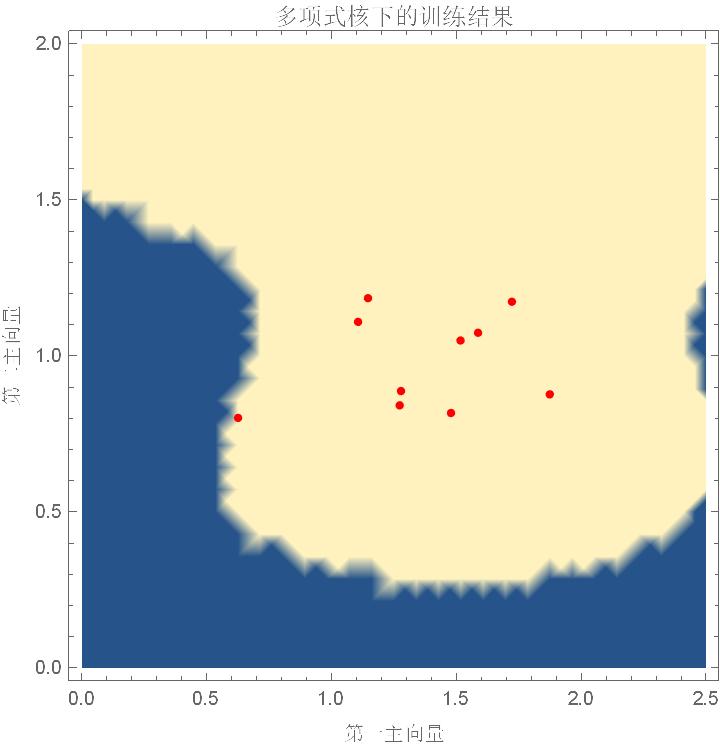


分类结果：

{1., 1., 1., 1., 1., 1., 1., 1., 1., 1.}

## 多项式核





分类结果

{1., 1., 1., 1., 1., 1., 1., 1., 1., 1.}

## 模型评价

在支持向量机模型下，我们采取了多个核对分类器进行训练。训练结果发现，线性核下的支持向量机Loss函数收敛最快，区域分割线为直线。该模型下前10个测试集数据只有一个被分为“1”类。

而采用高斯径向基函数核（RBF）以及多项式核的话，学习曲线收敛较慢，而且区域分割线较为复杂。

从训练集的分布来看，这个问题大致可以归为线性分类问题。因此不难理解线性核在该问题中的良好表现。

对于模型的改进建议是采取更多的线性分类模型，比较不同模型间准确率与收敛曲线，找出最适合的分类模型。