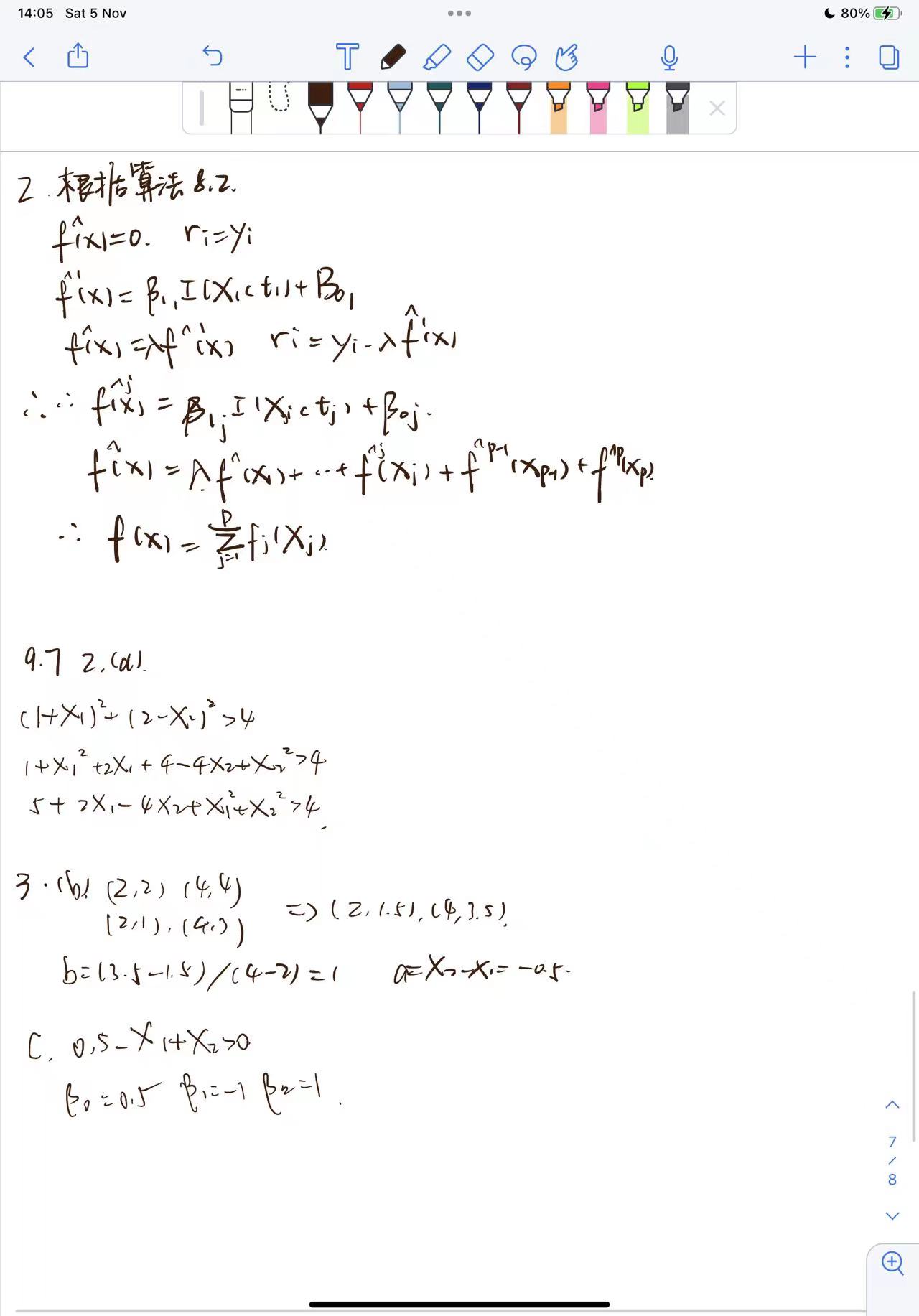
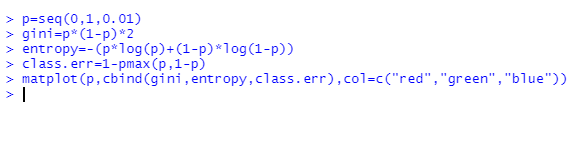
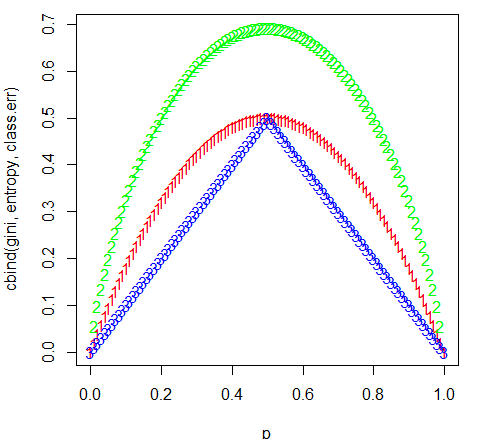
8.4

2

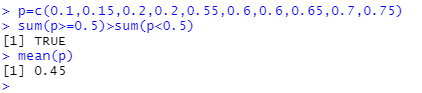


3





5.

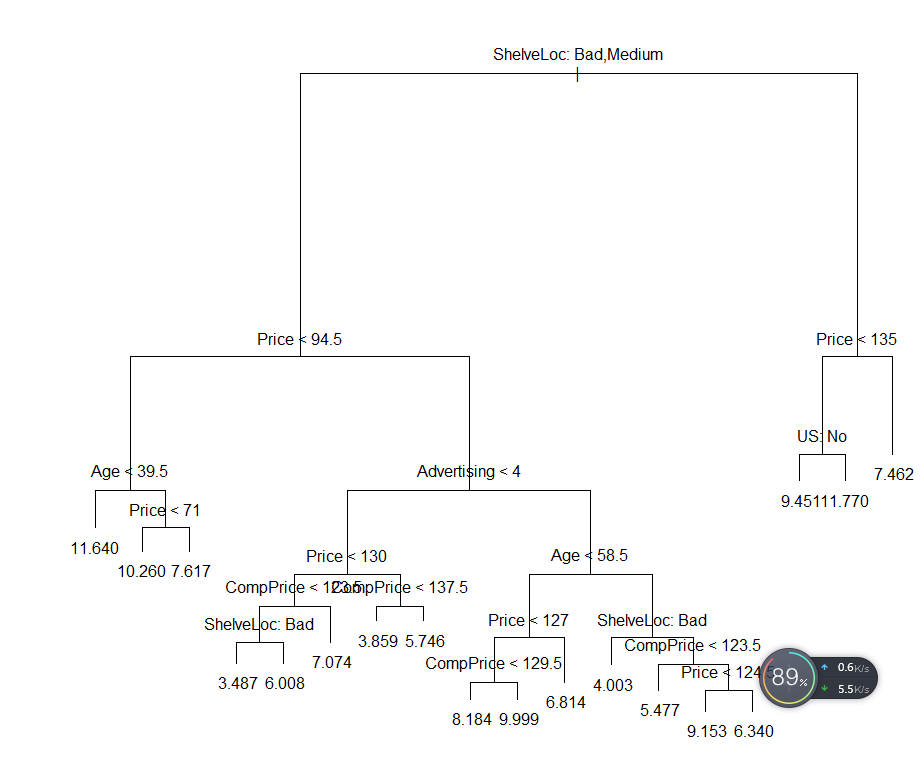


多数投票： 以0.5为界限，红色的比绿色的多，所以分类结果选择红色

平均概率： 算p的均值得到0.45，所以选择绿色

8（a.b）

****

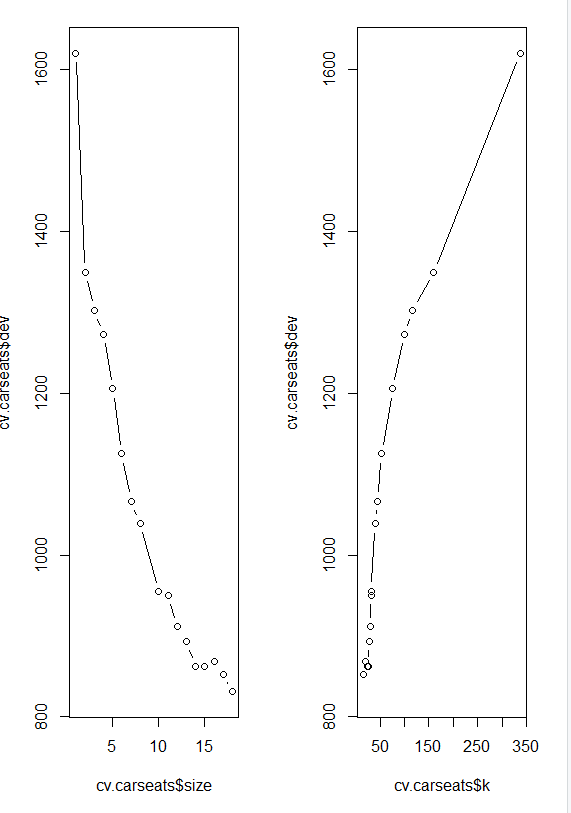


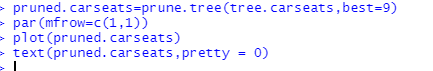
所以测试错误率为4.92

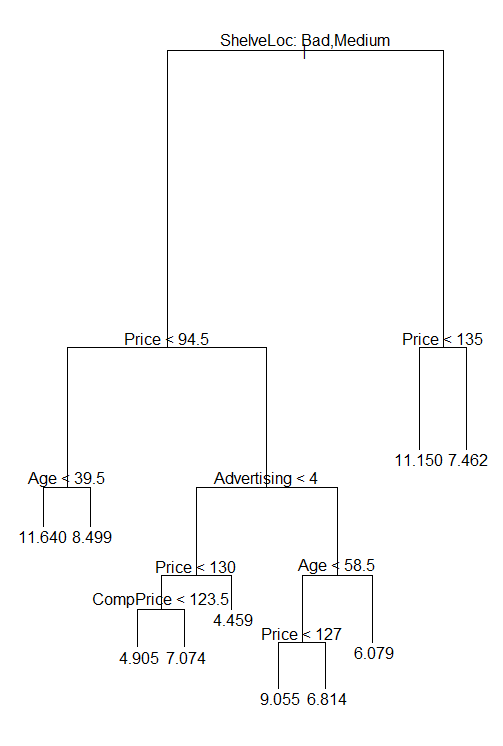
8(c)







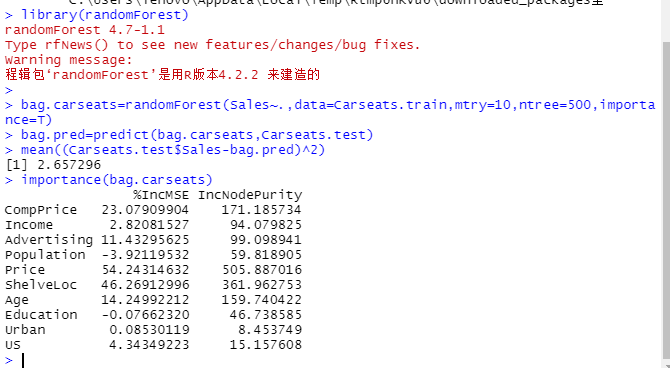






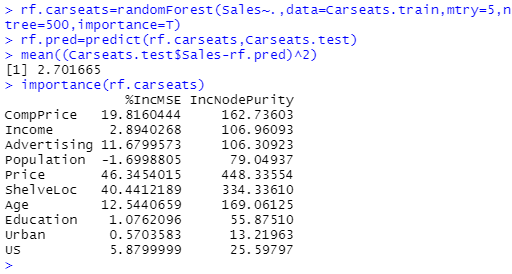
这次的均方误差是4.918，

8（d）



测试错误率是2.66，重要的变量： compPrice，Price，ShelveLoc，Age

8（e）



测试错误率为2.70，重要变量： Price,ShelveLoc,CompPrice

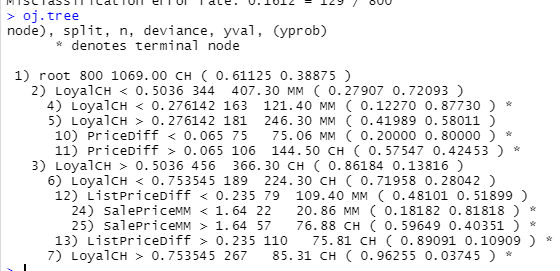
9（a，b）



最终使用四个变量：‘’LoyalCH“，“PriceDiff”，“ListPriceDiff”，“SalePriceMM”

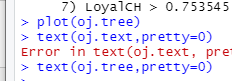
终端节点是7个，训练错误率0.1612

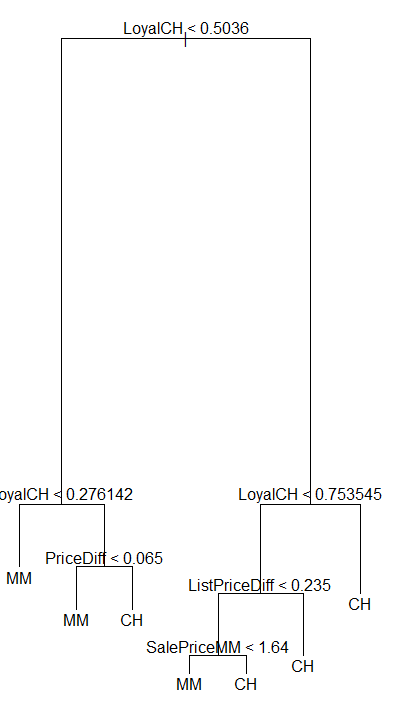
9（c）



选择12结点。 分割值是0.24，这个结点下有79个值，该节点以下区域内的所有点的偏差，值为109.40，这一点的预测值sales=MM， 有48.1%的概率选择CH作为sales的值，有51.9%的概率选择CH作为Sales 的值。

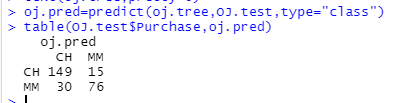
9（d）





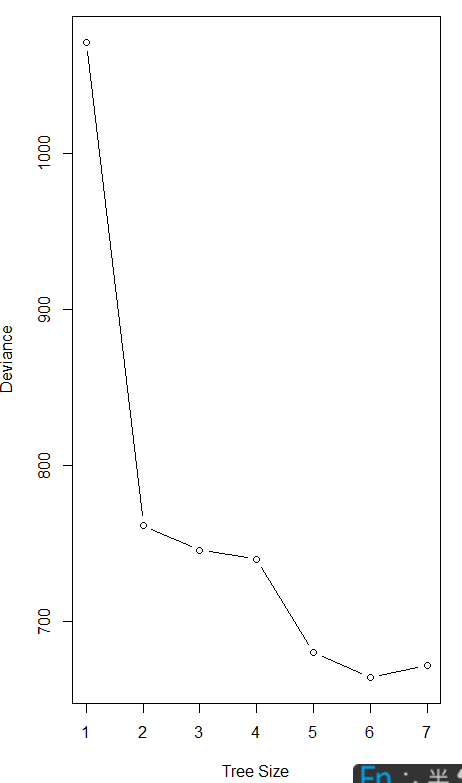
LoyalCH是树中最重要的变量，实际上前3个节点包含了LoyalCH。如果LoyalCH<0.27，则树预测MM。如果LoyalCH>0.76，则树预测CH。对于LoyalCH的中间值，还依赖于PriceDiff，listpricediff和saleprice的值。

9(e)



9(f.h)

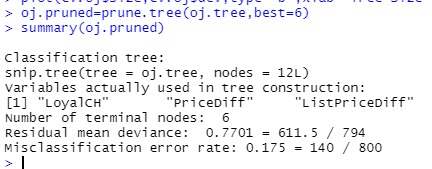




9(h)

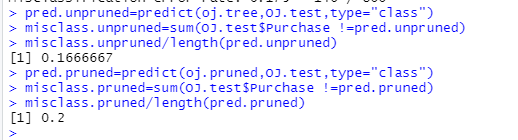
规模为6的树对应的交叉验证分类错误率最低

9（i，j）



修剪后的树与原始树的错误分类误差为0.175

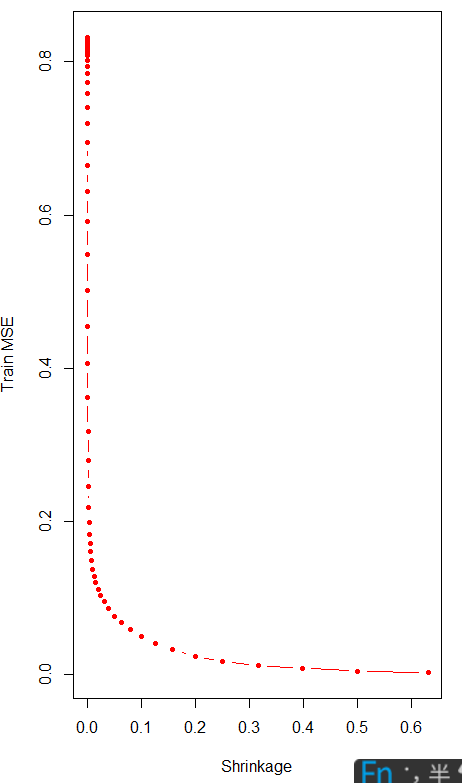
9（k）



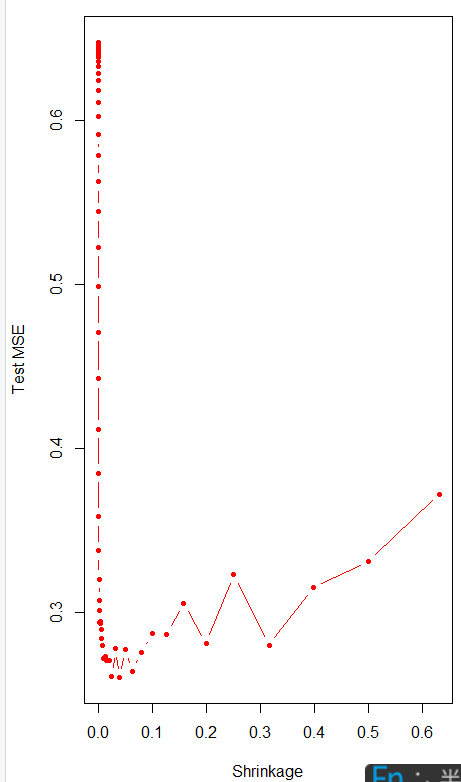
剪枝后的错误率高

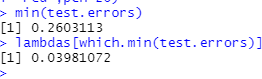
10（a，b，c）



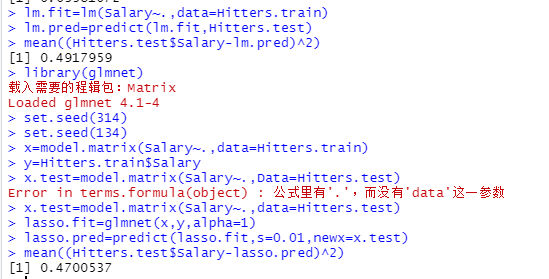


10（d）



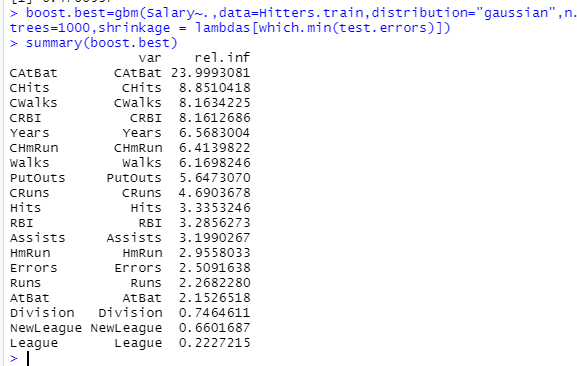


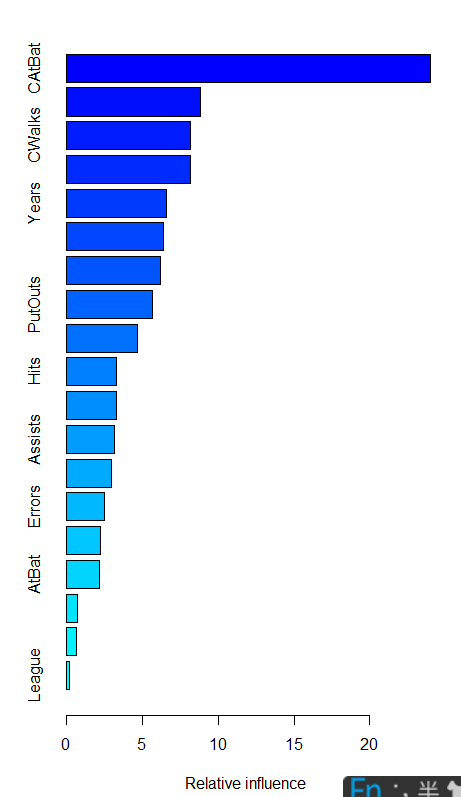
10（e）



线性模型和像Lasso这样的正则化都比提升法具有更高的检验MSE。

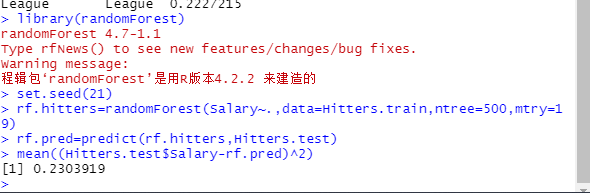
10（f）





提升法重要的变量有： cAtBat，CWalks

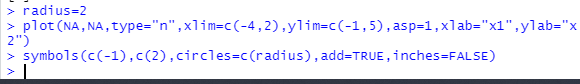
10（g）

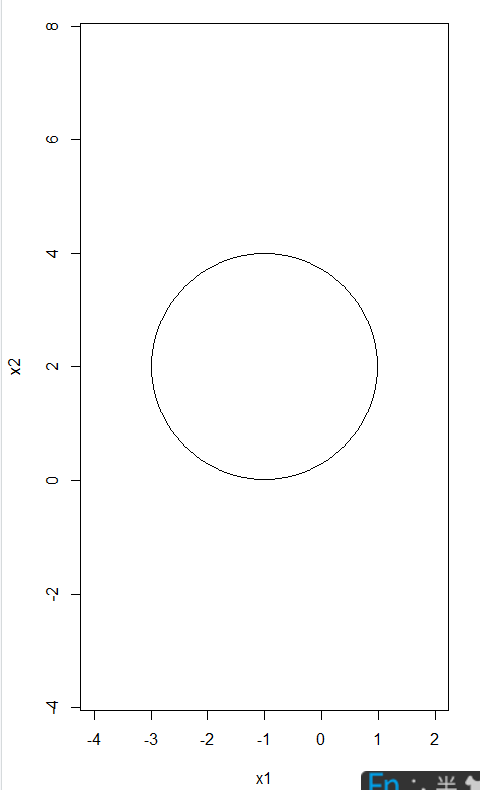


测试均方误差是0.23

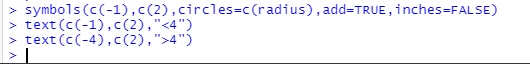
9.7

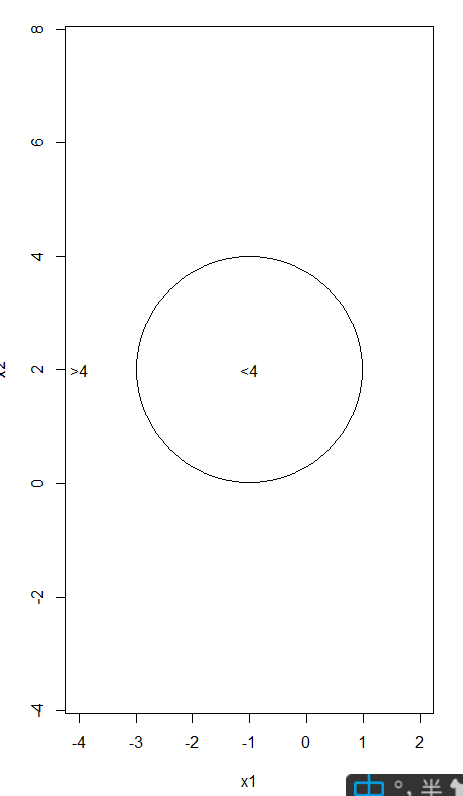
2（a）



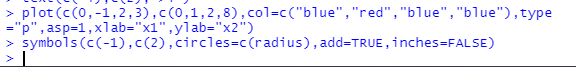


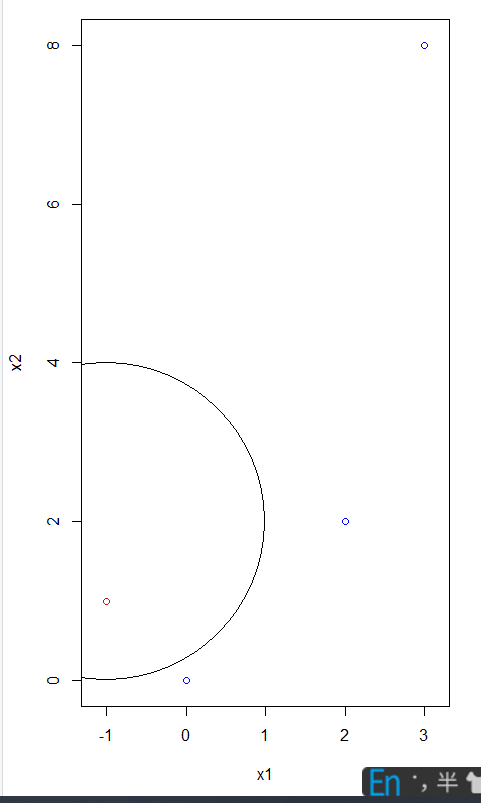
9（b）



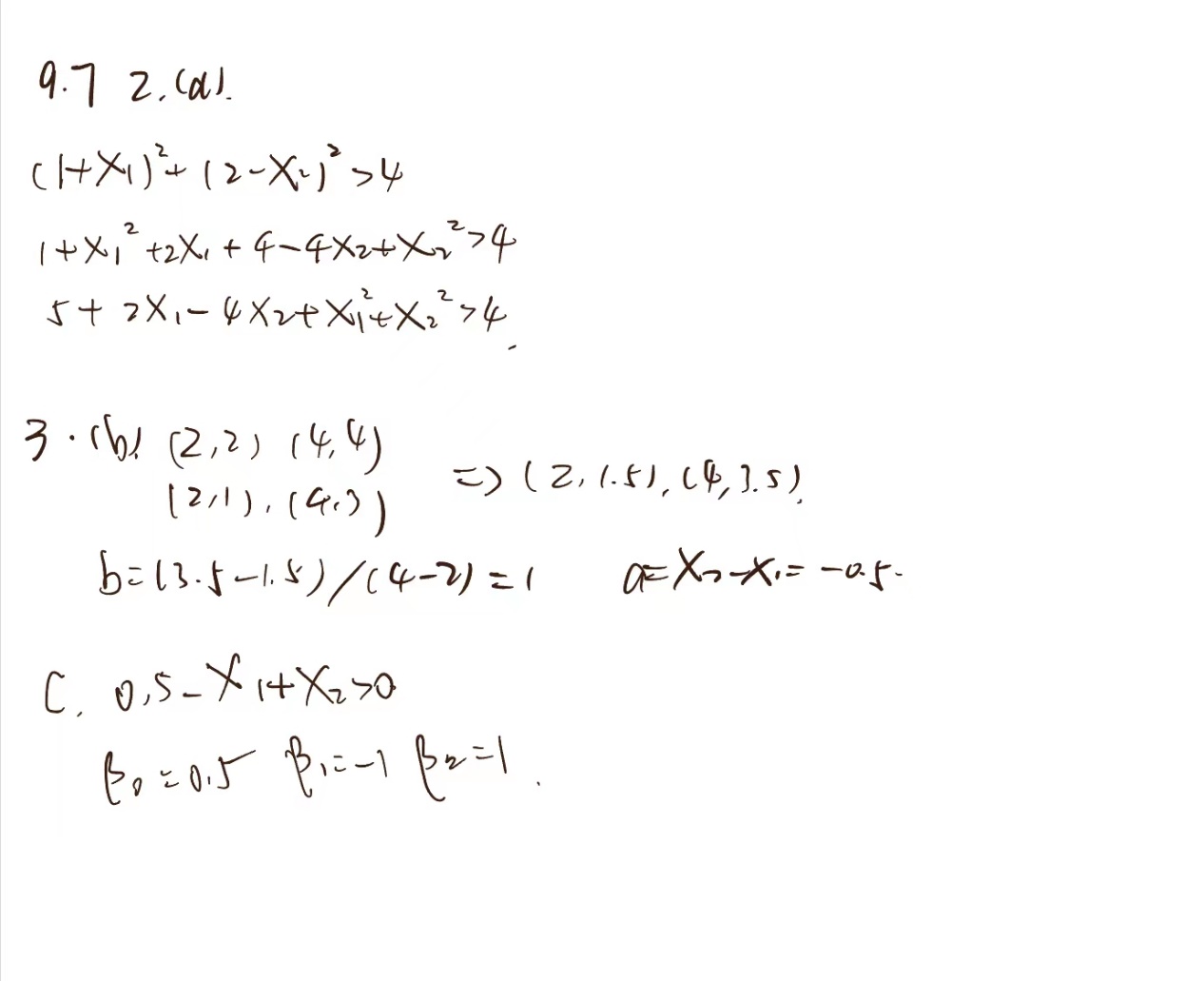


9（c）

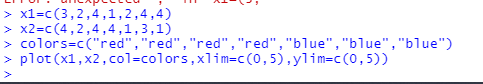


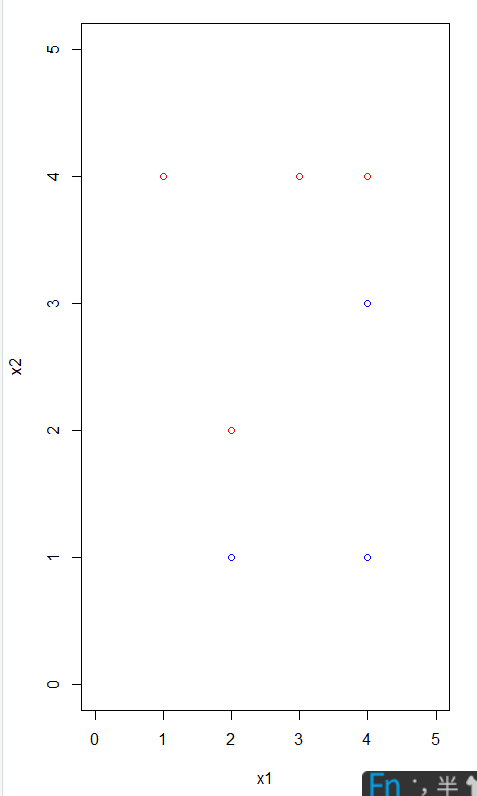


9（d）

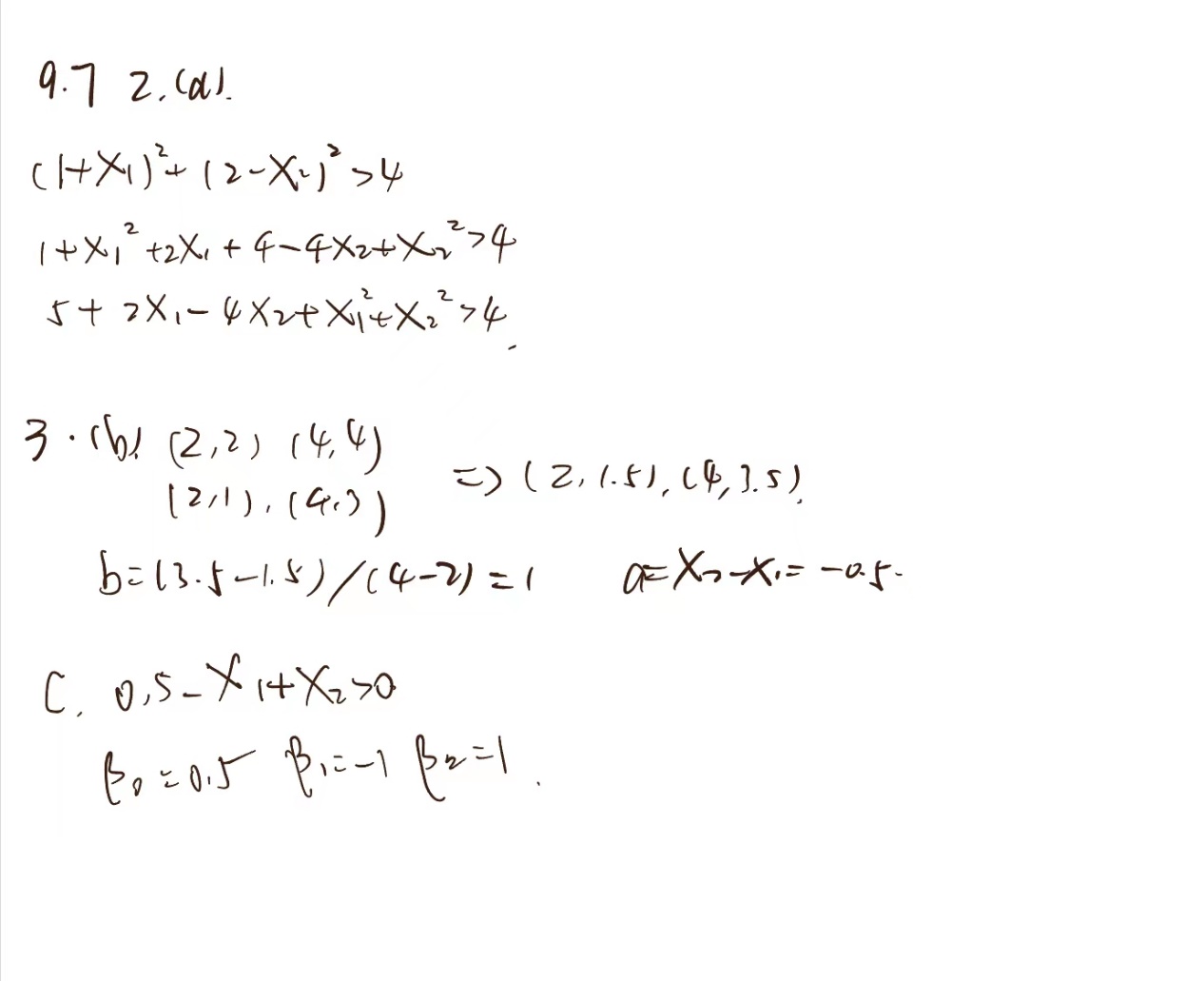


3（a）

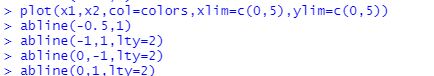


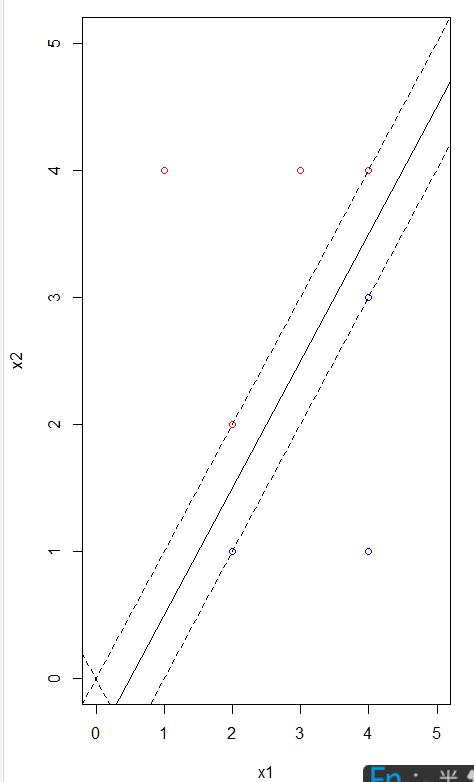


3（b，c）

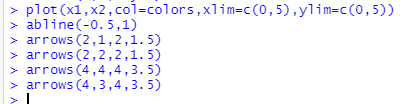


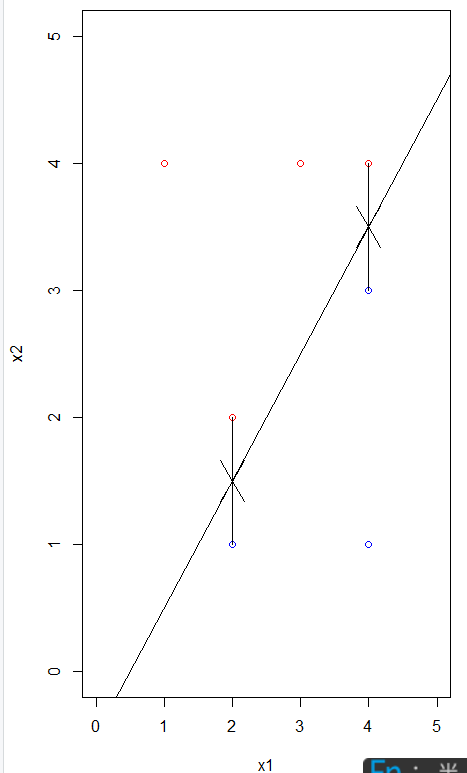
3（d）





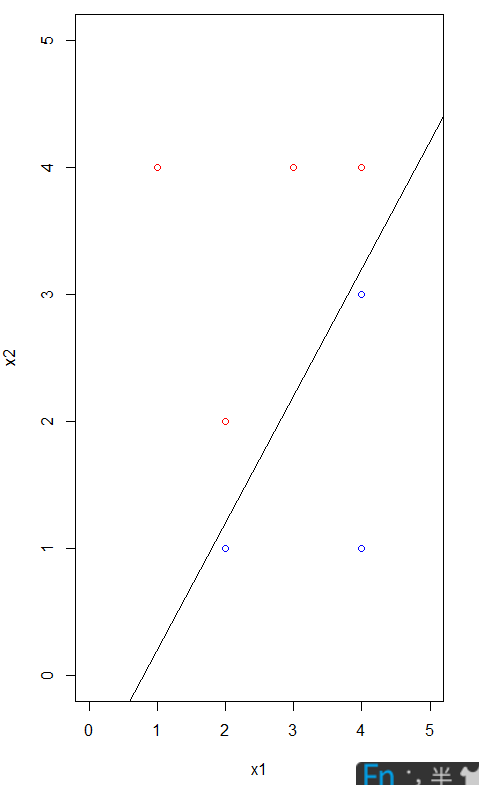
3（e）





3（g）

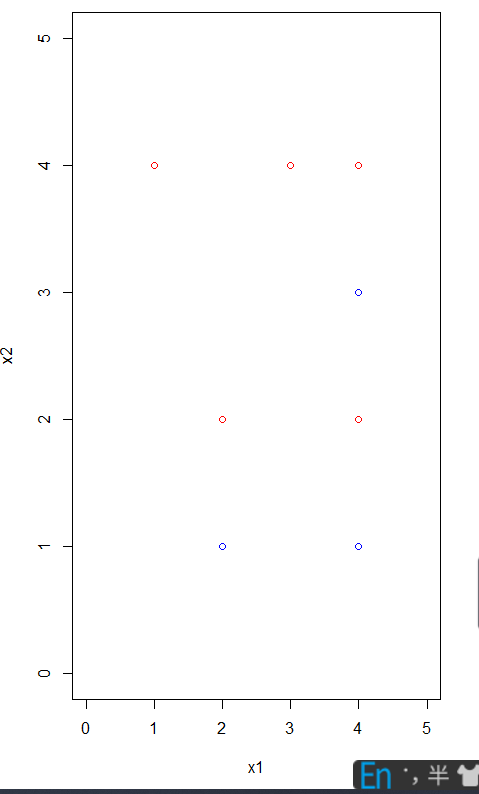




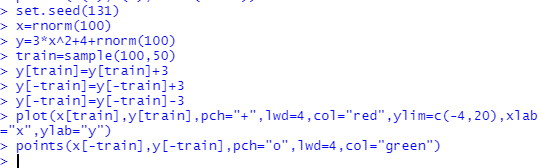
**-0.8-X1+X2>0**

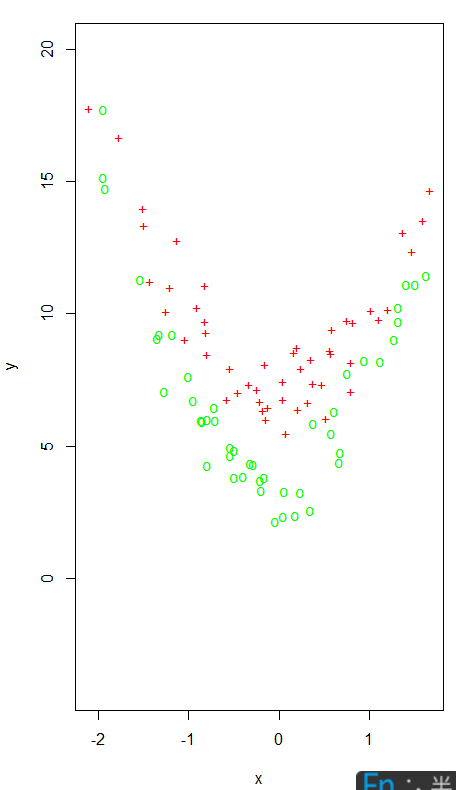
3(h)

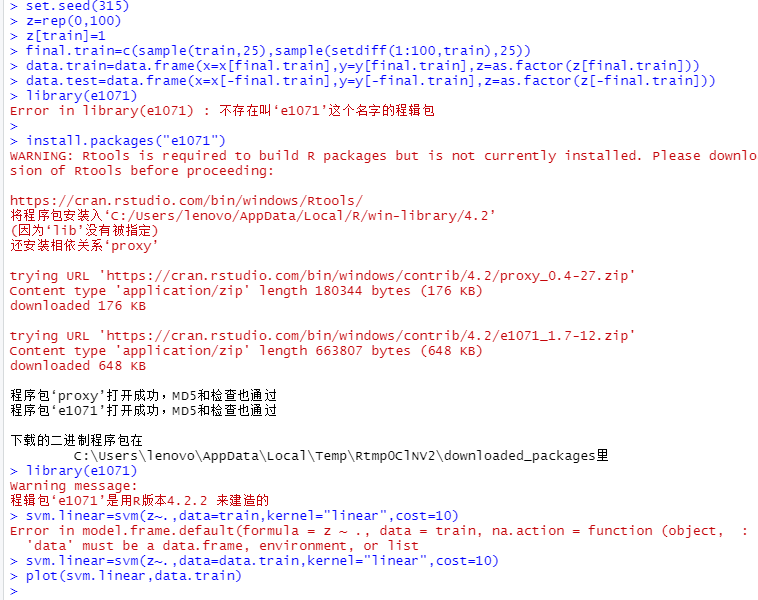


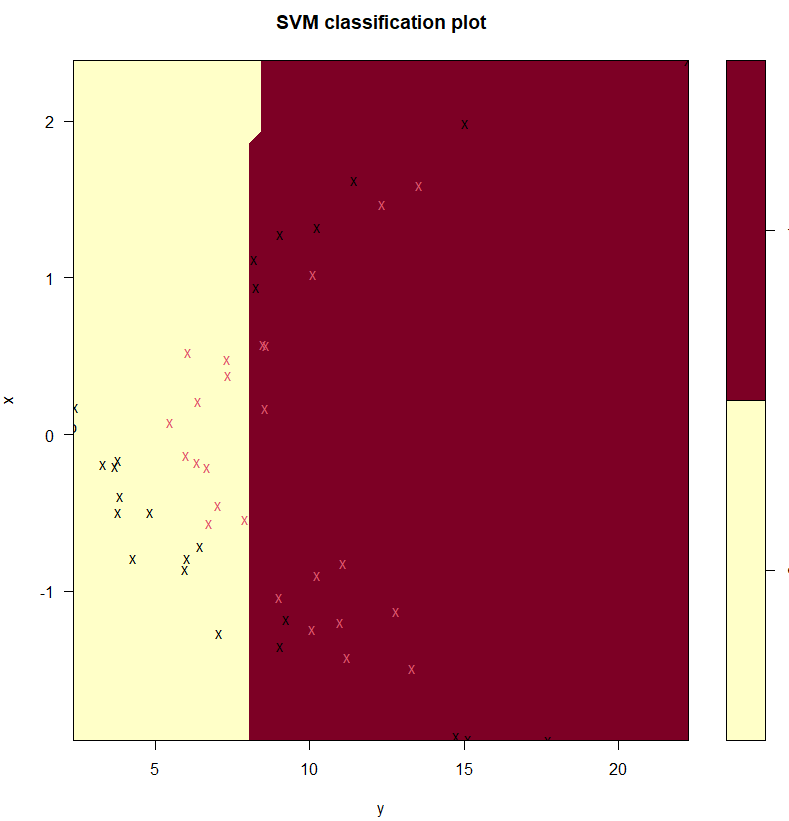


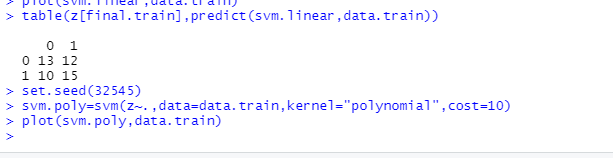
4

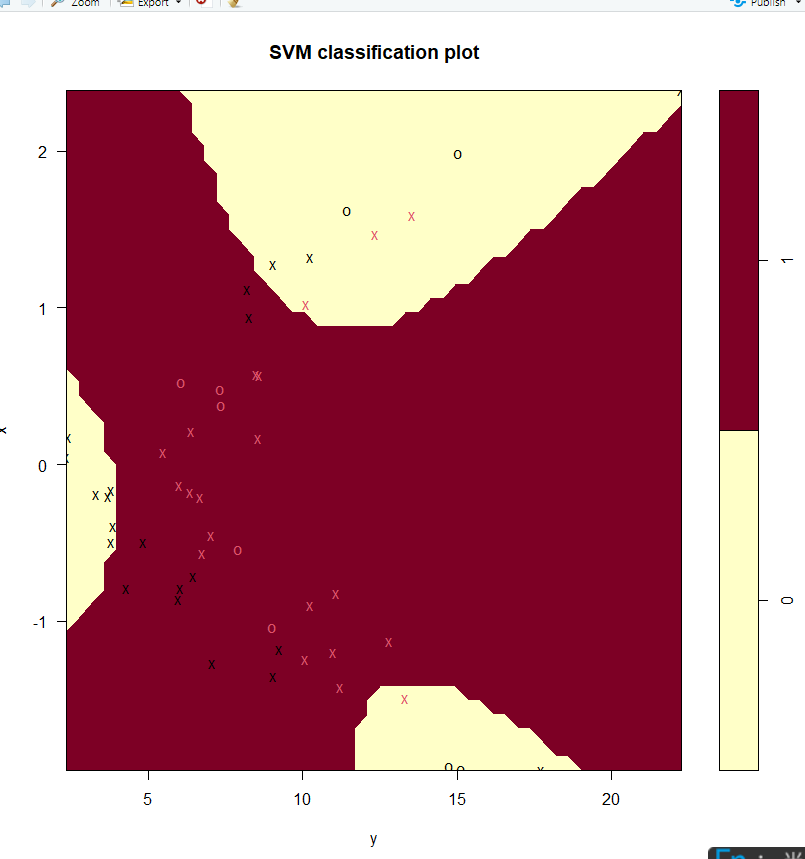


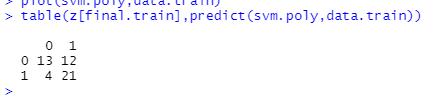






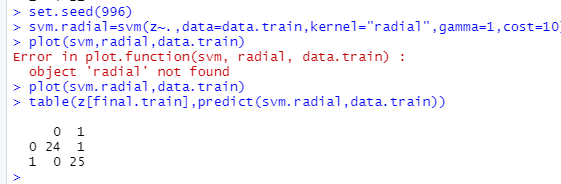






这是一个默认的3次多项式核。它使15列数据错误。

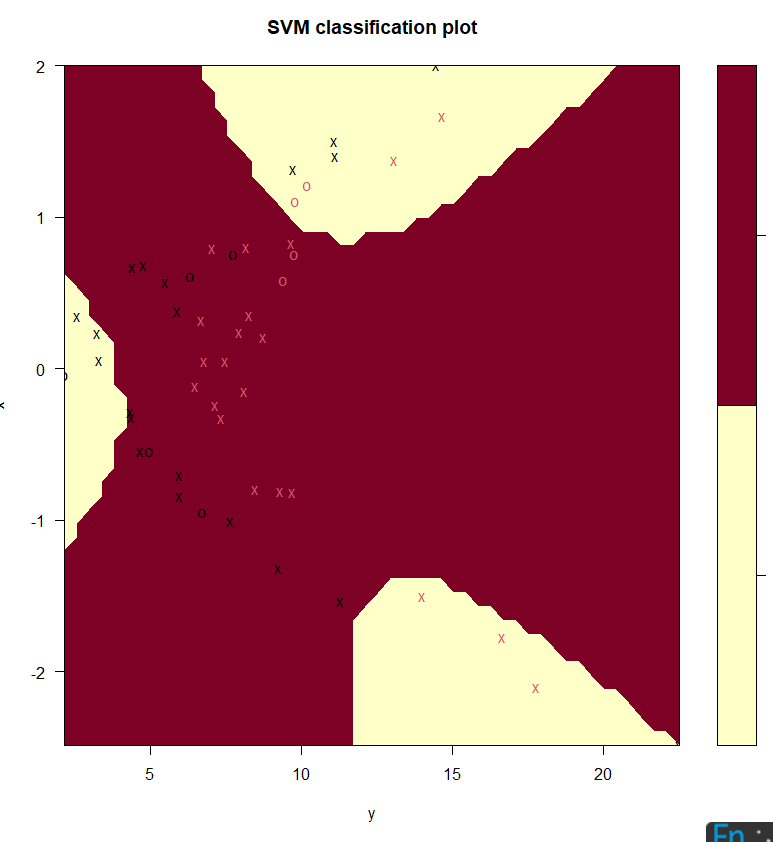
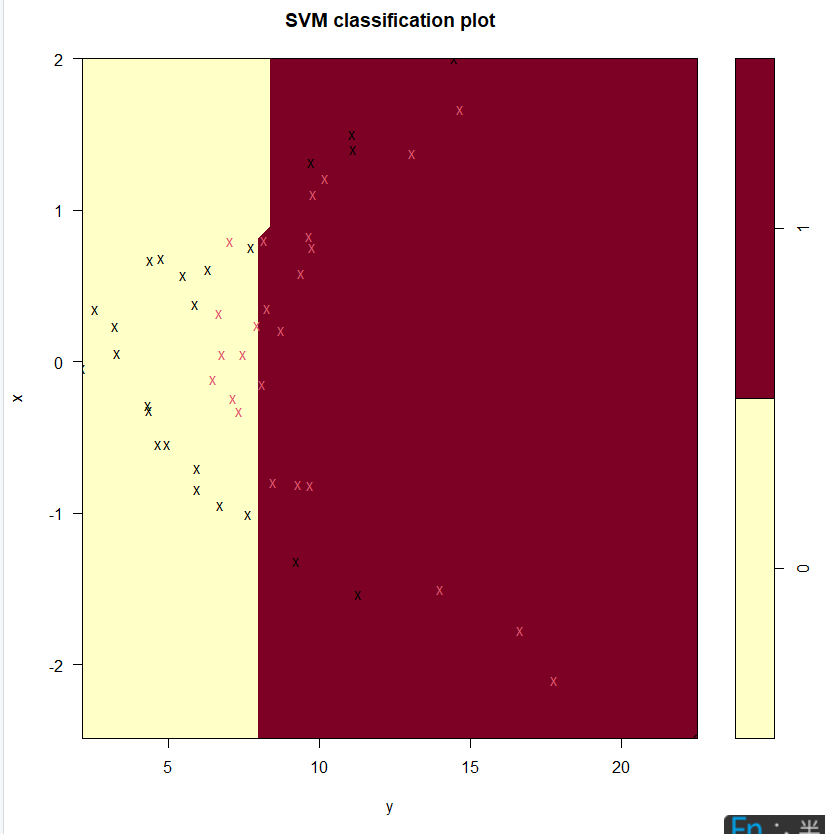
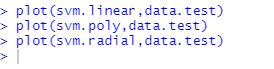
最后，我们训练了一个径向基核为1的支持向量机。

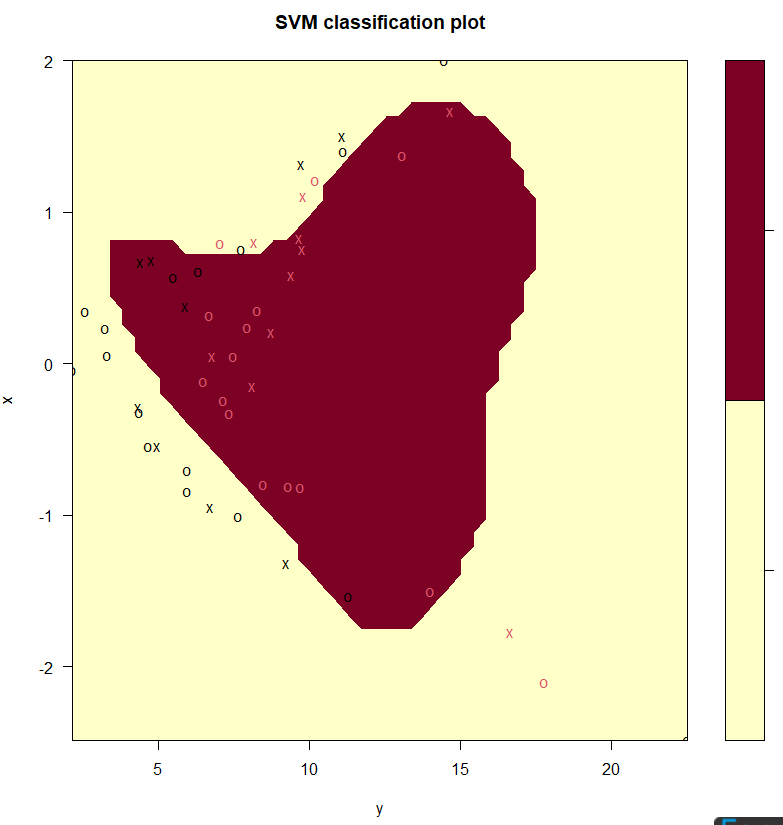


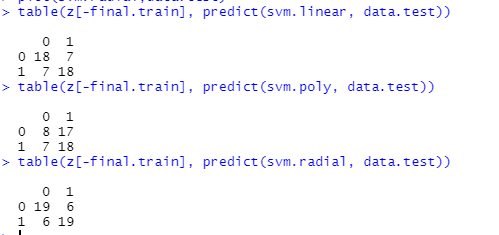


这个分类器对训练数据进行了完美的分类!

接下来是对测试集的分析

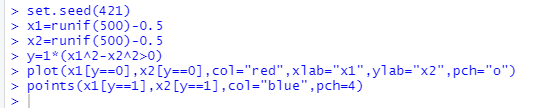


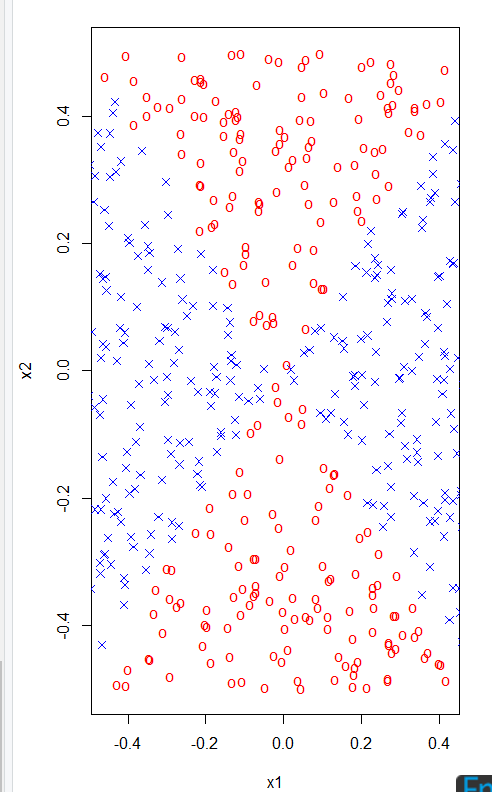




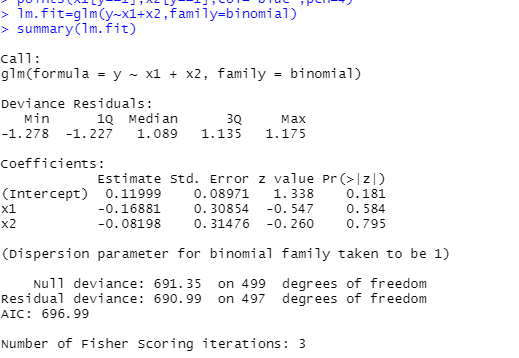
从表中可以看出，线性、多项式和径向基核分别对6、14和0个测试点进行了错误分类。径向基核是最好的，具有零的测试误分类误差。

5（a，b）

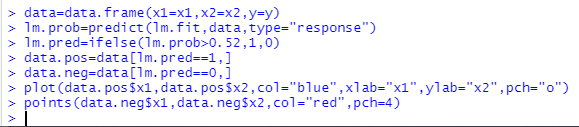


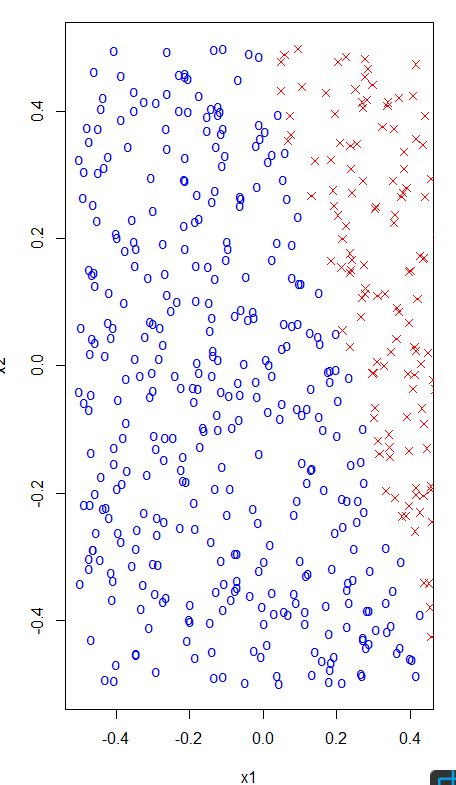


5(c)



5(d)

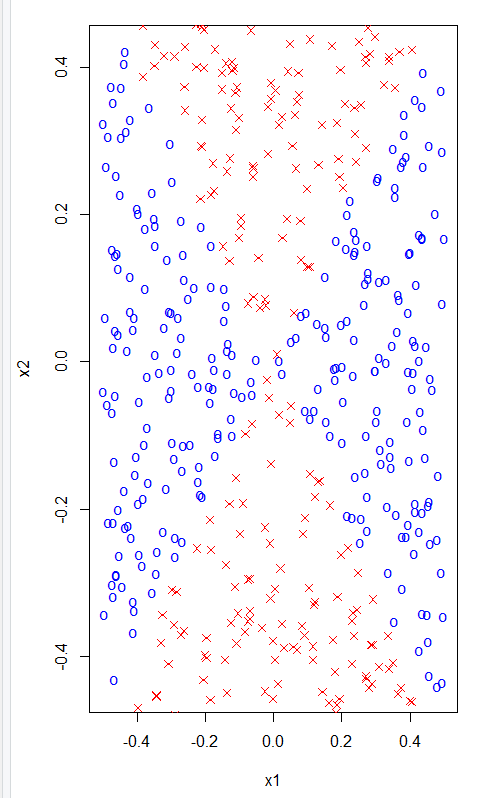




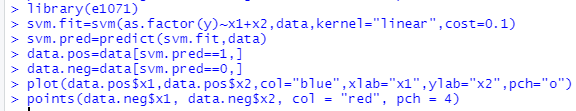
在阈值为0.5情况下，不是线性决策边界，所以扩大阈值为0.52，可以发现决策边界编程线性的了。

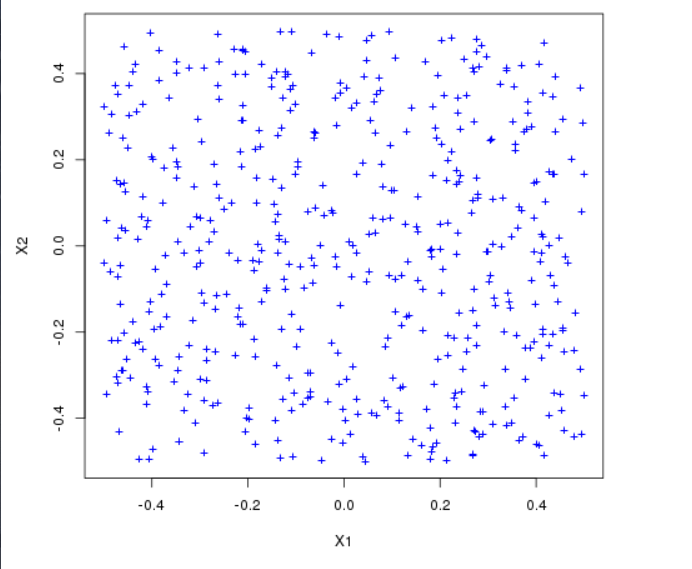
5（e，f）





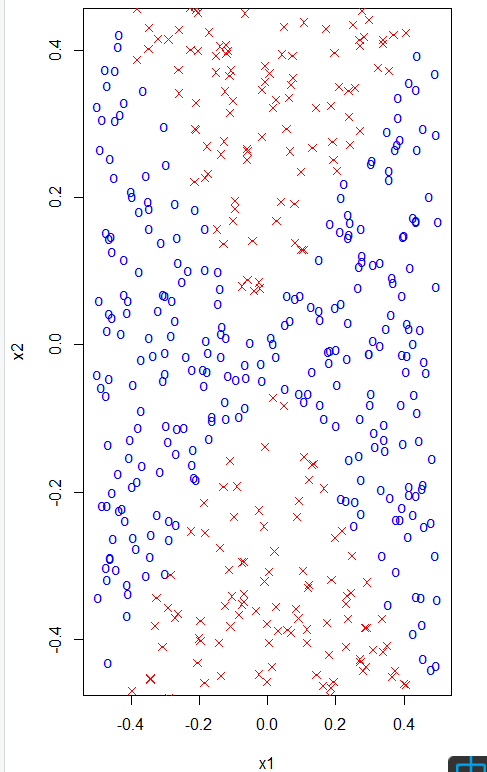
5（g）





5（h）

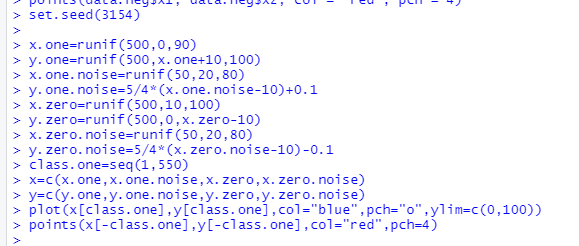


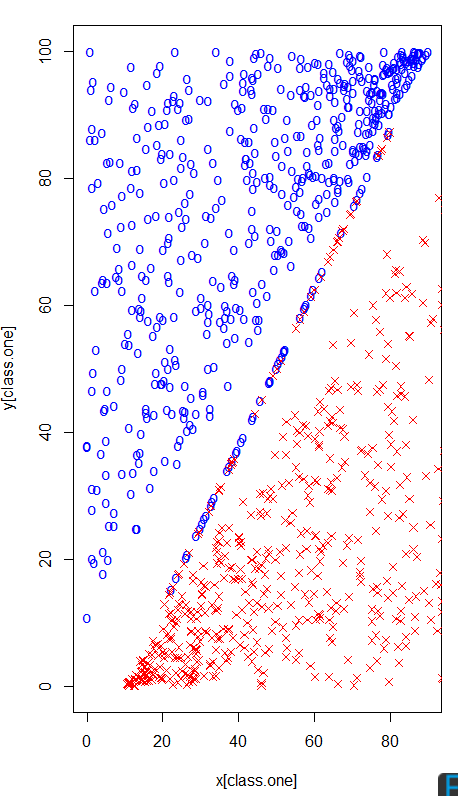


5（i）

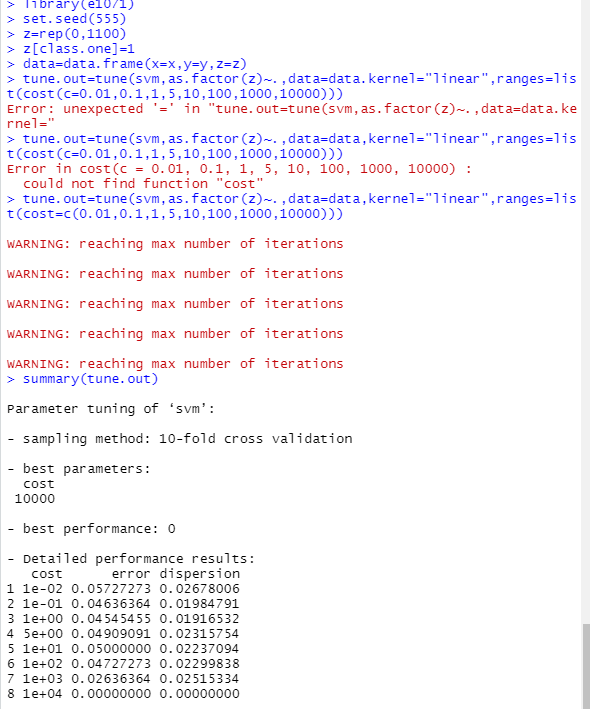
具有非线性核的支持向量机擅长寻找非线性边界。非交互逻辑回归和线性核支持向量机都无法找到决策边界。

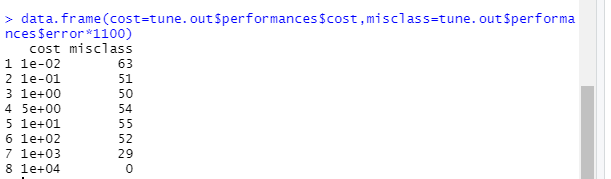
6（a）



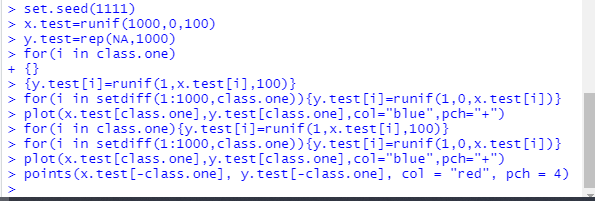


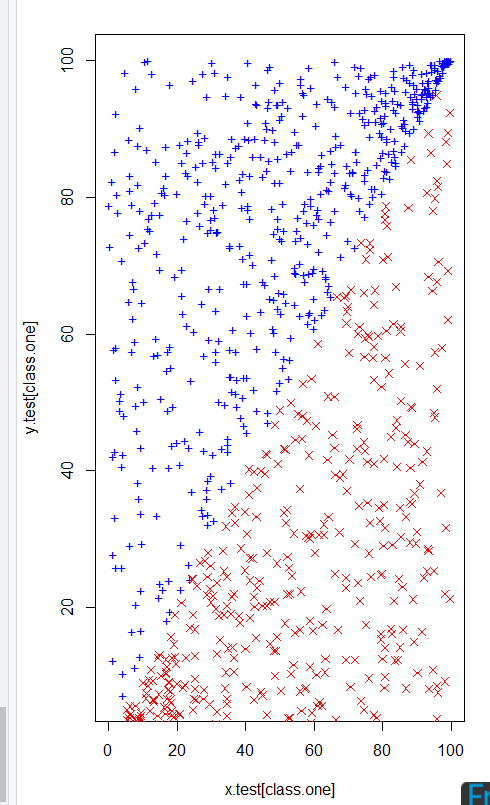
6（b）

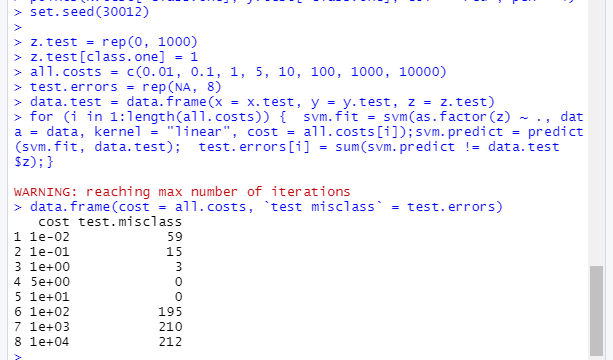




6（c）







6（d）

再一次验证了线性核的过拟合现象。试图正确地拟合分类噪声点的代价很大，从而导致训练数据过拟合。

7





