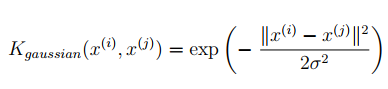
实验六

**function sim = gaussianKernel(x1, x2, sigma)**

计算向量X1和X2的高斯距离

代码：

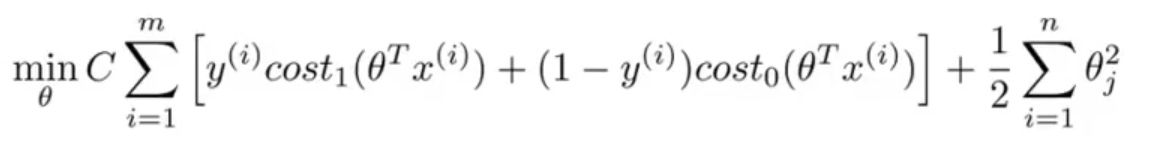
sim = exp( -sum((x1 - x2).^2) / 2 / sigma^2 );



附：linearKernel

sim = x1' \* x2; % dot product

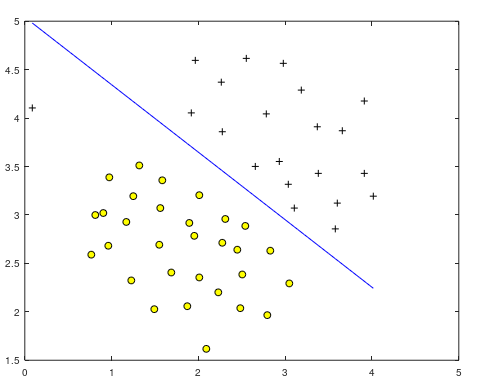
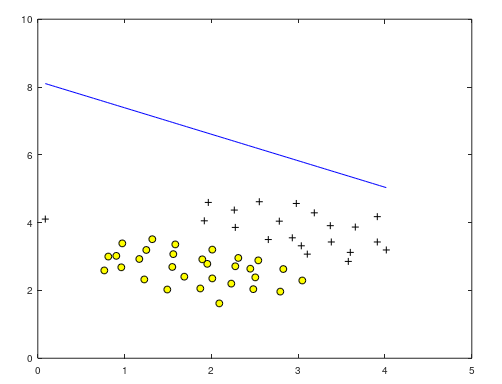
% 核函数表示两个样本的不相似程度，高斯核即两个样本的高斯距离，线性核即两个样本的点积

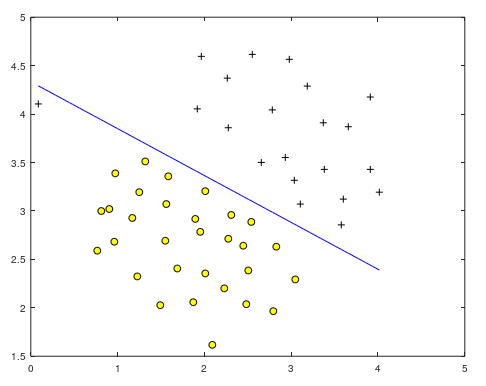
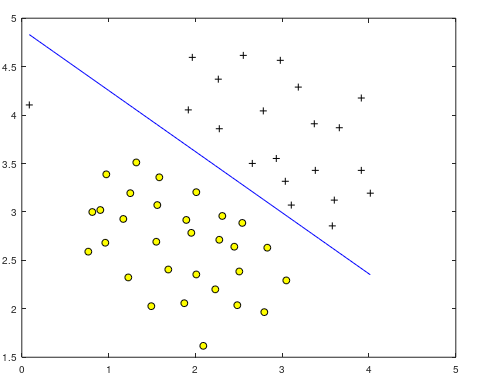


linearSVC的超参数只有C一个，C的作用相当于逻辑回归中1/lambda，C越小，惩罚项比重越大，越容易欠拟合，C越大，惩罚项比重越小，越容易过拟合。（实验过程中发现C变大后训练时间变长，也说明倾向于将模型训练得更加复杂即过拟合）

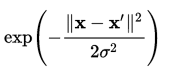
linearSVC在data1上的训练结果：

C分别为0.01 0.1 1 10





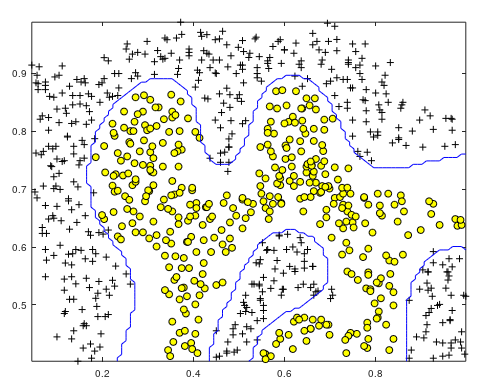
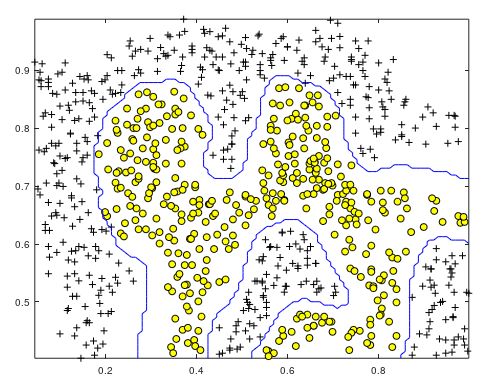
RbfSVC的超参数有两个：C和sigma。更常用的是gamma



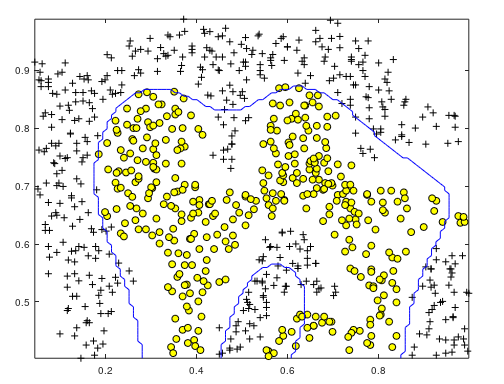
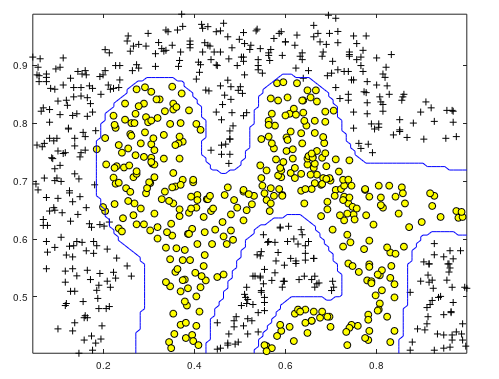
% 参数 gamma与sigma成反比，gamma越小，影响的训练样本越远，可以看作是支持向量影响半径的倒数。

RbfSVC在data2上的训练结果：

C=1，sigma=0.03 C=1，sigma=0.1



C=0.1，sigma=0.03 C=0.1，sigma=0.1



实验发现：C越大，模型越复杂，训练越慢，C越小，训练越快，边界越光滑；sigma越大（gamma越小），支持向量考虑的区域越大，模型拟合复杂边界的能力越差，sigma越小（gamma越大），越趋向于分隔的更细，倾向于过拟合。

**function word\_indices = processEmail(email\_contents)**

在一系列C和sigma的组合中选择交叉验证结果最好的一组，即grid search cross validation

代码：

error = 1

for CTemp = [0.1 0.5 1 5 10]

for sigmaTemp = [ 0.01 0.1 0.5 1 5]

model = svmTrain(X, y, CTemp, @(x1, x2) gaussianKernel(x1, x2, sigmaTemp));

predictions = svmPredict(model, Xval);

errorTemp = mean(double(predictions ~= yval));

if errorTemp < error

error = errorTemp;

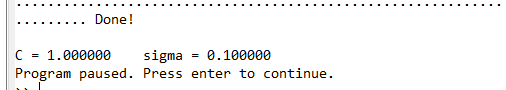
C = CTemp;

sigma = sigmaTemp;

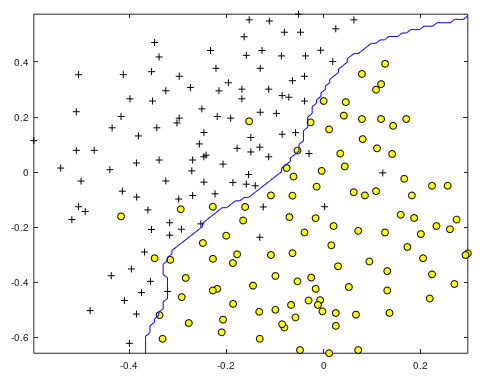
endif

endfor

endfor

data3上的grid search结果：

最终训练结果



**function word\_indices = processEmail(email\_contents)**

处理并判断email\_content中的每个词是否出现在vocabList以及出现的位置（index），返回一个有所有出现的词的index练成的向量

（处理单词过程中用到的porterStemmer是用于返回某一个单词变体的原始单词，如去掉ing，去掉s等，但感觉效果并不是很好，比如把EmailSample1里的anyone变成了anyon，this变成了thi）

（需补充的代码为判断处理过后的单词str是否在vocabList中，如果存在就在word\_indices向量后再加上出现的index）

代码：

for index = 1:length(vocabList)

if strcmp(vocabList{index}, str)

word\_indices = [word\_indices ; index];

break;

endif

end

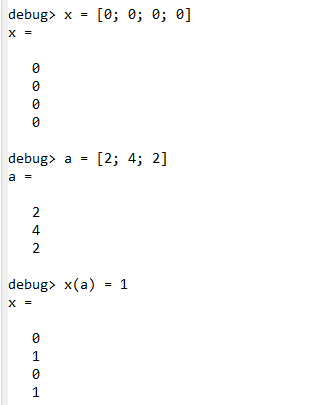
**function x = emailFeatures(word\_indices)**

返回一个长度同vocabList的向量x，对于vocabList里的所有单词的index，如果在word\_indices里出现，则记为1，其余为0

代码：

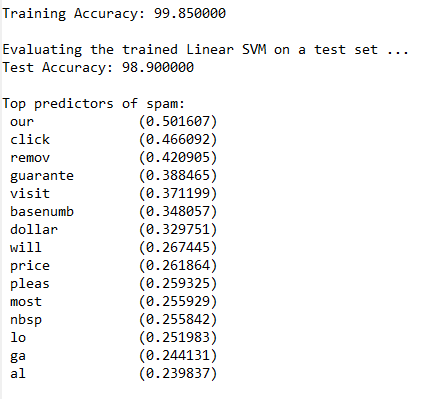
x(word\_indices) = 1;

此处用到的技巧如下图



Ex6\_spam结果

线性核svm训练结果以及对判断垃圾邮件权重最大的15个单词（垃圾邮件最有可能含有的单词）



四封样例邮件分别的预测结果

