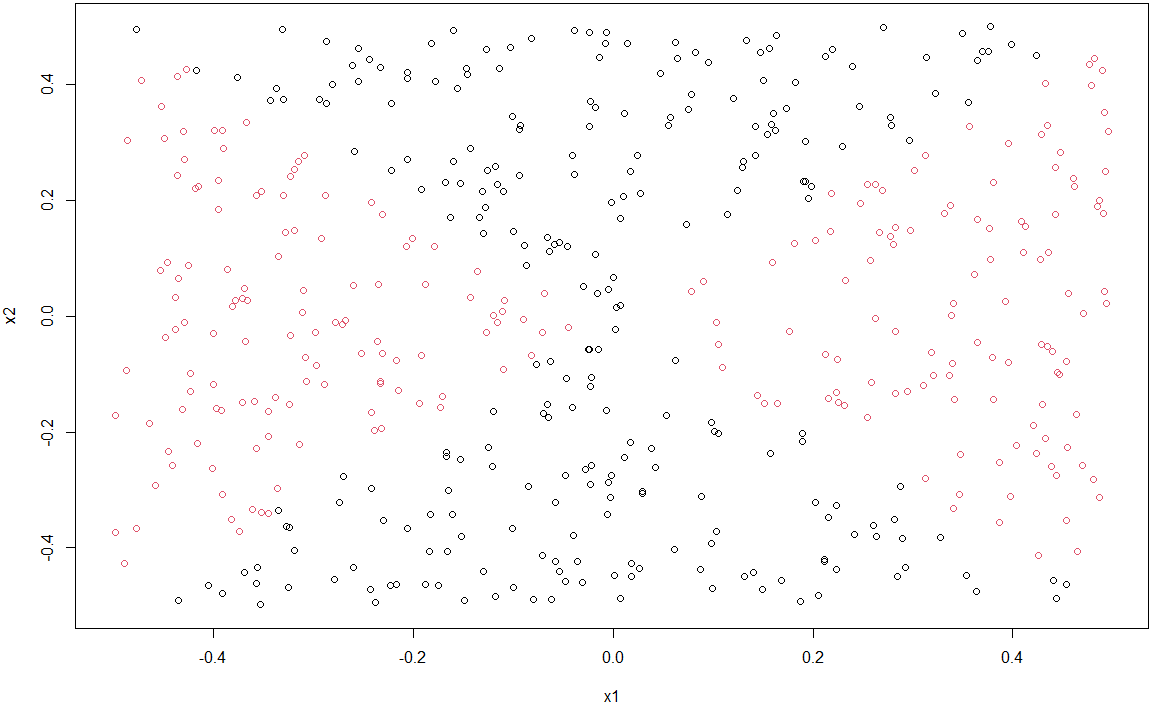
5. We have seen that we can fit an SVM with a non-linear kernel in order to perform classification using a non-linear decision boundary. We will now see that we can also obtain a non-linear decision boundary by performing logistic regression using non-linear transformations of the features.

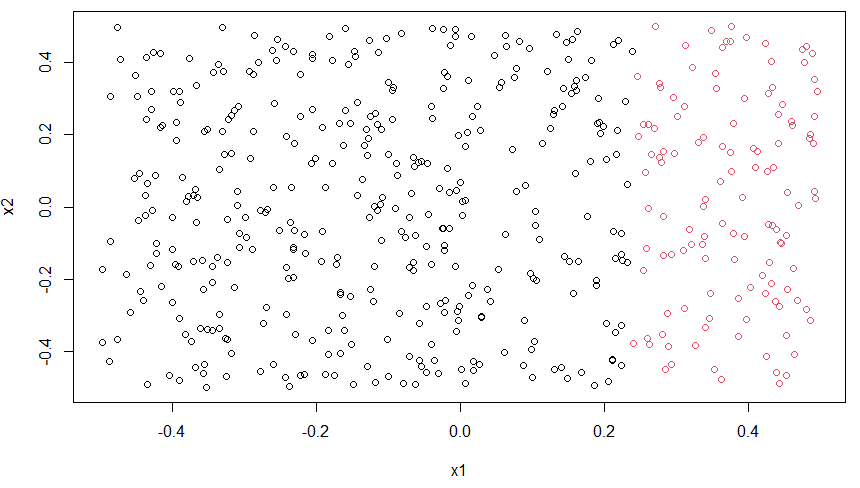
1. Generate a data set with n = 500 and p = 2, such that the observations belong to two classes with a quadratic decision boundary between them. For instance, you can do this as follows:  
   skip
2. Plot the observations, colored according to their class labels. Your plot should display X1 on the x-axis, and X2 on the y-axis.  
   

The black represent y=0, red represent y=1.

1. 一張含有 文字 的圖片

   自動產生的描述Fit a logistic regression model to the data, using X1 and X2 as predictors.  
     
   By fitting the data with logistic regression model, we can see the result as right, all of the predictors are not significant.
2. Apply this model to the training data in order to obtain a predicted class label for each training observation. Plot the observations, colored according to the predicted class labels. The decision boundary should be linear.  
     
   Because the logistic regression give a probability as output, I use the threshold value which give the lowest error rate(0.489 or 0.49).

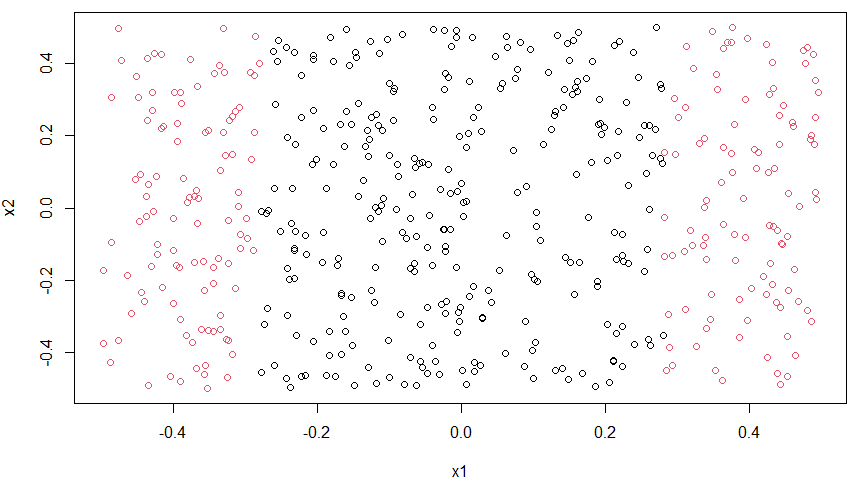
The plot blows show the observation and colored according to the predicted class, and the boundary is linear.



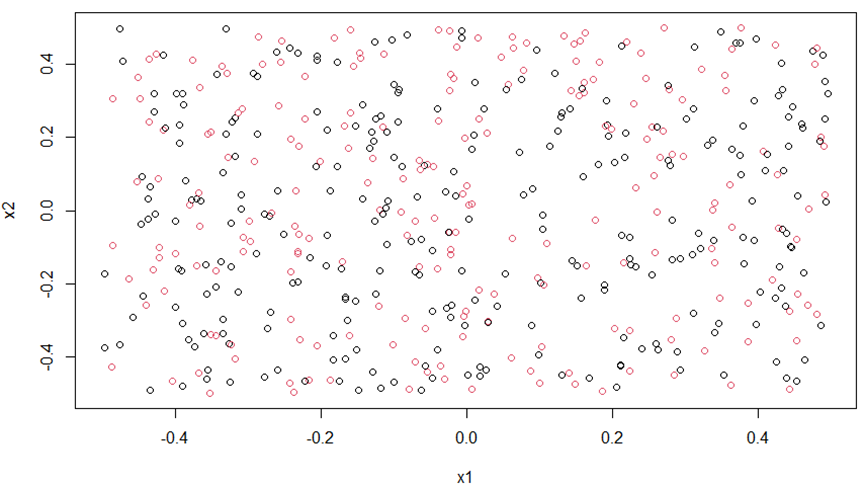
1. Now fit a logistic regression model to the data using non-linear functions of X1 and X2 as predictors (e.g. X2 1 , X1×X2, log(X2), and so forth).

Three models were used to predict

1. y~I(x1^2),

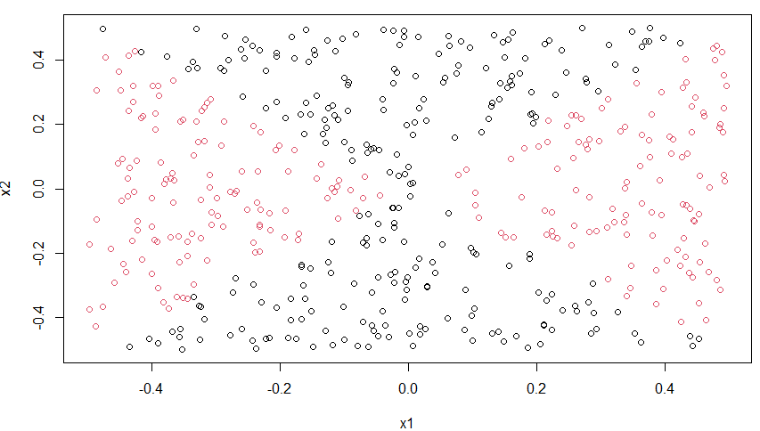
This model have two boundary. The error rate is 22.8%一張含有 文字, 瓶 的圖片

自動產生的描述

1. y~I(x1^2)+log(x2)

一張含有 文字, 標誌 的圖片

自動產生的描述This model seems don’t have the boundary, and have the highest error rate  
54.2%

1. y~I(x1^2)+I(x2^2)

This one is the real function which we use to generate the data so it is not surprising to see it have 0 error rate!

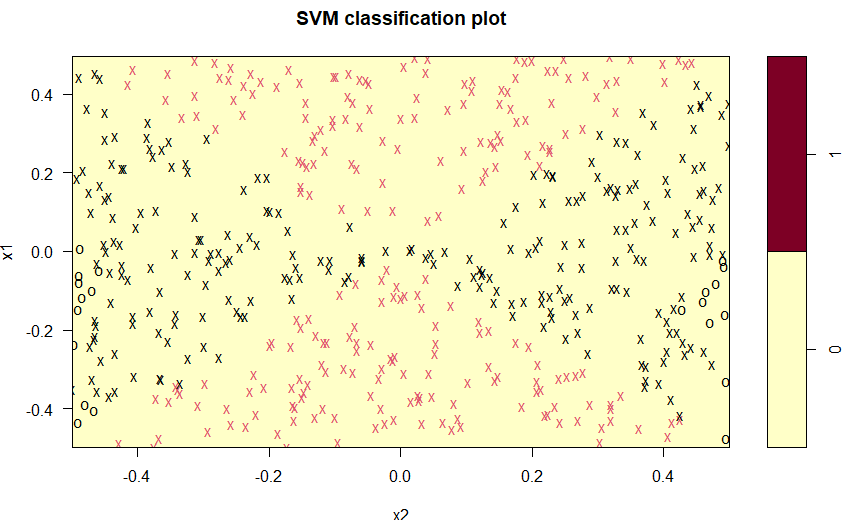
一張含有 文字, 標誌, 室外 的圖片

自動產生的描述

1. Apply this model to the training data in order to obtain a predicted class label for each training observation. Plot the observations, colored according to the predicted class labels. The decision boundary should be obviously non-linear. If it is not, then repeat (a)-(e) until you come up with an example in which the predicted class labels are obviously non-linear.   
   The non-linear boundary was seen in the model 3 in part (e).
2. Fit a support vector classifier to the data with X1 and X2 as predictors. Obtain a class prediction for each training observation. Plot the observations, colored according to the predicted class labels.

The plot at left is generate by svm model’s plot method, and the right plot is drawn by myself. We can see that linear svm model label all observation to 0. Separate the data by a straight line is impossible.

一張含有 大自然, 雨 的圖片

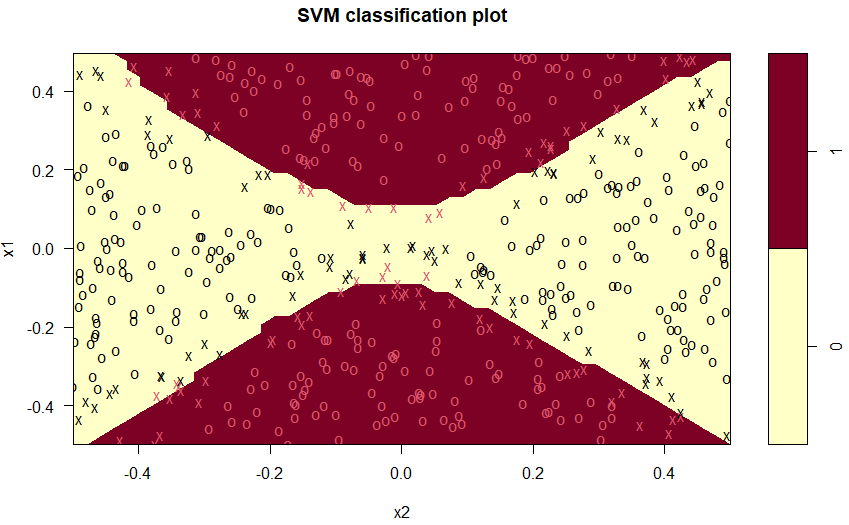
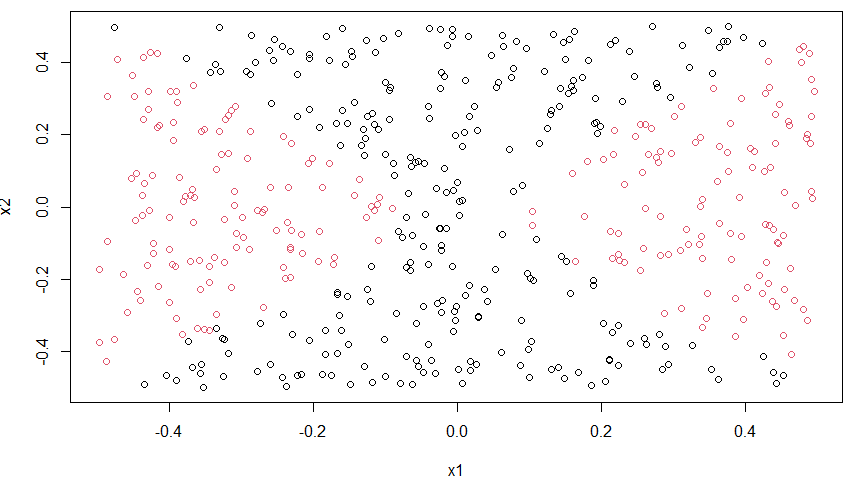
自動產生的描述

1. Fit a SVM using a non-linear kernel to the data. Obtain a class prediction for each training observation. Plot the observations, colored according to the predicted class labels.

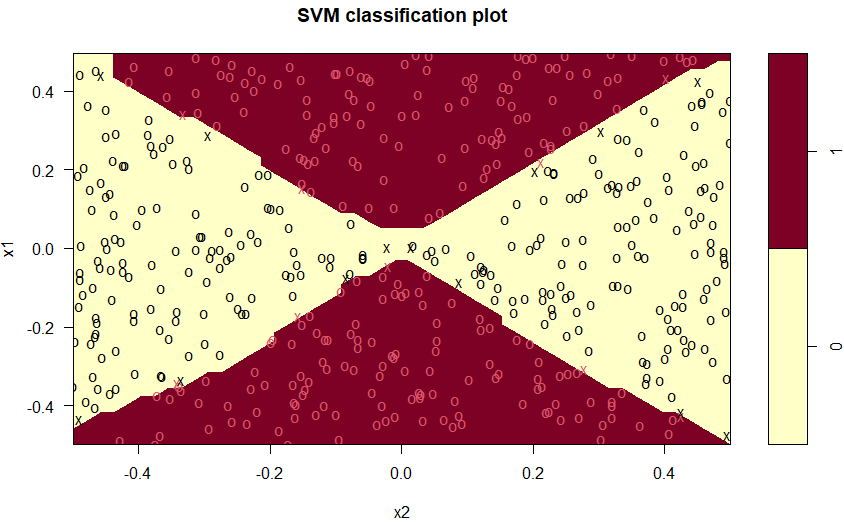
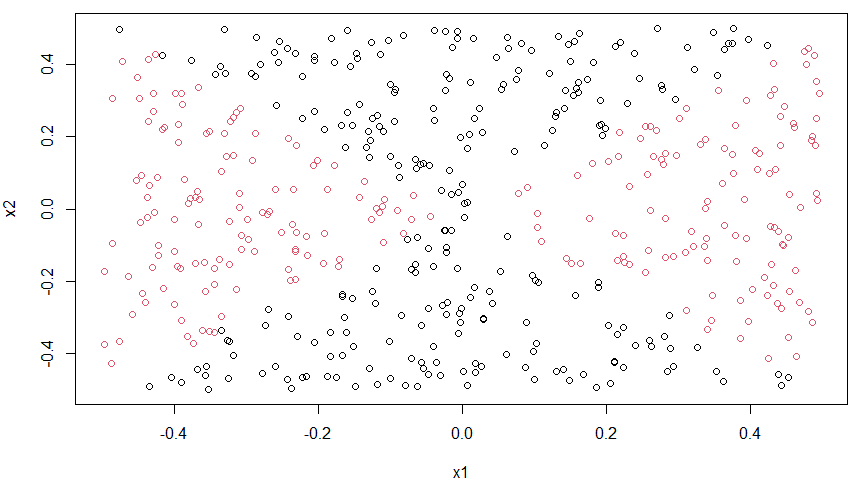
By trying to set kernel to “radial” and “polynomial”, we get different result.

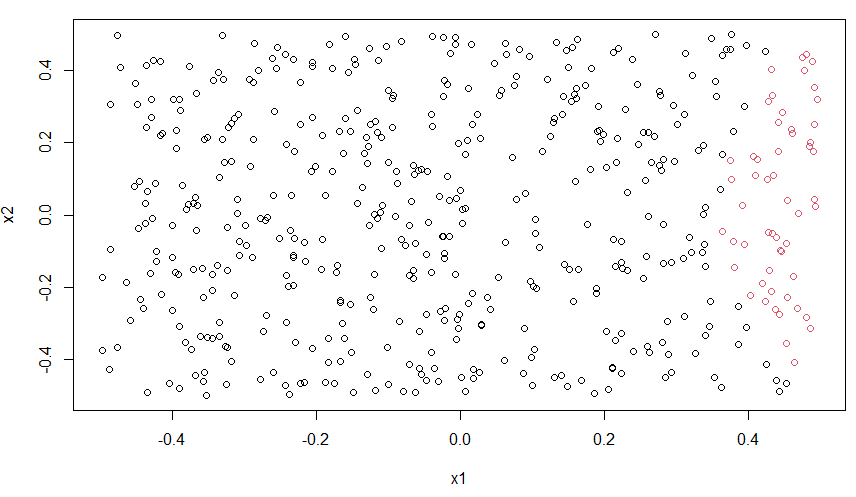
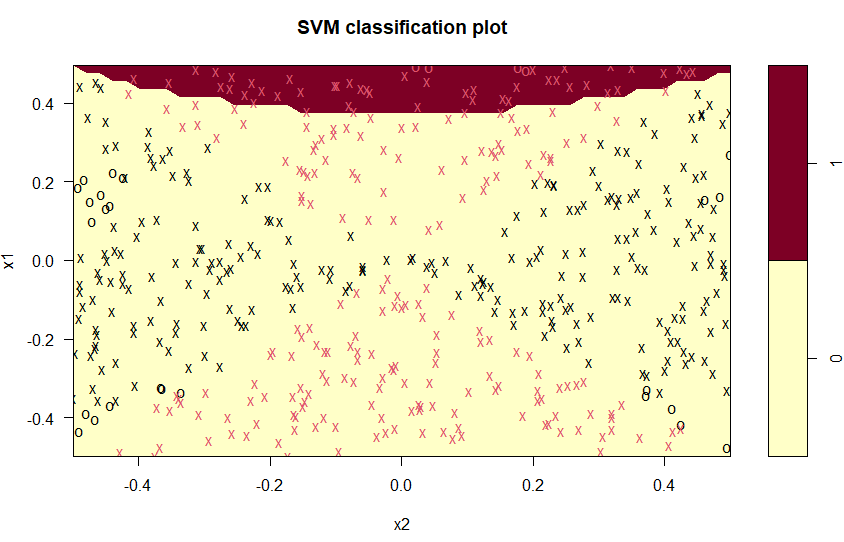
1. Radial

This model provide the boundary which vary close to the real boundary, which have the low error rate at 2.8%.(under cost =10 gamma =1)



By increasing the cost parameter to 10000, the model can perfectly predict the training data(0 error rate), but this might increase the risk to overfit the data.



1. Polynomial  
   The polynomial svm don’t work as well as the radial model, and have the error rate 37%.  
    
2. Comment on your results.

This exercise ask us to predict the data by both logistic regression and SVM.The data isn’t linearly separable so it leads to poor result if we try to simply use the logistic or Support Vector Classifier. However, if we adjust the input by doing some data transformation or change the model to support vector machine, both model work vary well.

7. In this problem, you will use support vector approaches in order to predict whether a given car gets high or low gas mileage based on the Auto data set.

1. Create a binary variable that takes on a 1 for cars with gas mileage above the median, and a 0 for cars with gas mileage below the median.

The median is 22.75, and have 196 for high, 196 for low.

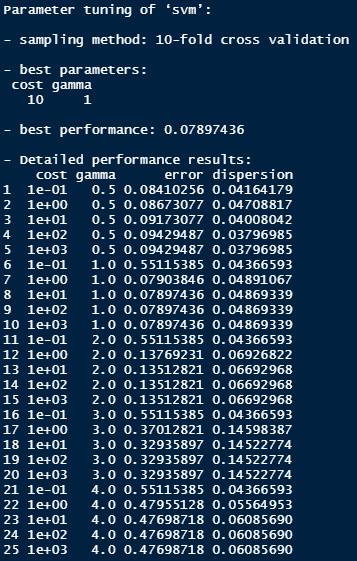
1. 一張含有 文字 的圖片

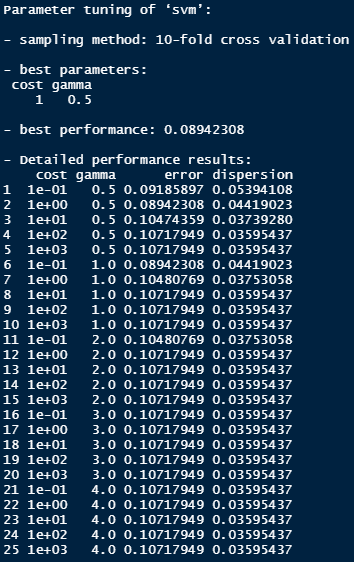
   自動產生的描述Fit a support vector classifier to the data with various values of cost, in order to predict whether a car gets high or low gas mileage. Report the cross-validation errors associated with different values of this parameter. Comment on your results. Note you will need to fit the classifier without the gas mileage variable to produce sensible results.  
     
   Using the tune function to do 10-fold cross validation. The best parameters is cost=0.1, which only have 8.67% of cross validation error rate.

Use this best model to predict the full data set. The error rate is 8.16%.

一張含有 文字, 瓶 的圖片

自動產生的描述

1. Now repeat (b), this time using SVMs with radial and polynomial basis kernels, with different values of gamma and degree and cost. Comment on your results.
   1. radial:  
      By performing 10-fold cross validation, the best model’s parameter is gamma=1, cost=10 the training error rate is 7.9%. Apply this model on full data set and construct the confusion matrix. The error rate is 0%  
      一張含有 文字 的圖片

      自動產生的描述
   2. polynomial  
      By performing 10-fold cross validation, the best model’s parameter is gamma=0.5, cost=1 the training error rate is 8.94%. Apply this model on full data set and construct the confusion matrix. The error rate is 0.7%

一張含有 文字 的圖片

自動產生的描述

1. Make some plots to back up your assertions in (b) and (c).  
   The linear model

|  |  |
| --- | --- |
|  |  |
|  |  |

Radial

|  |  |
| --- | --- |
|  |  |
|  |  |

Polynomial

|  |  |
| --- | --- |
|  |  |
|  |  |

By those plot we can see the patterns about the position of support vector are very close, they are all in the bottom-left on the scatter plot. However I couldn’t found any plot which have a brown area in it, but I’m sure I am using the right function and run the code correctly.