Jiayuan Guo - HW3

Chapter 4, Problem 5

(a) If the Bayes decision boundary is linear, we expect QDA to perform better on the training dataset because its higher flexibility will yield a closer fit. But on the test dataset, LDA will perform better because QDA will overfit the Bayes decision boundary.

(b) If the Bayes decision boundary is non-linear, we expect QDA to perform better on both the training and test dataset for its higher flexibility

(c) We expect the test prediction accuracy of QDA relative to LDA to improve.

QDA is more flexible and also has higher variance. When the the sample size n increases, training dataset becomes larger, and the higher variance will not be the major concern

(d) False.

With fewer sample points, complex model like QDA will lead to overfit and cause larger test error than LDA

Chapter 4, Problem 6

(a) The logistic model is

$\hat{p}(X) = \frac{exp(-6+0.05X_1+X_2)}{1+exp(-6+0.05X_1+X_2)}$, X1 = hours studied, X2 = undergrad GPA

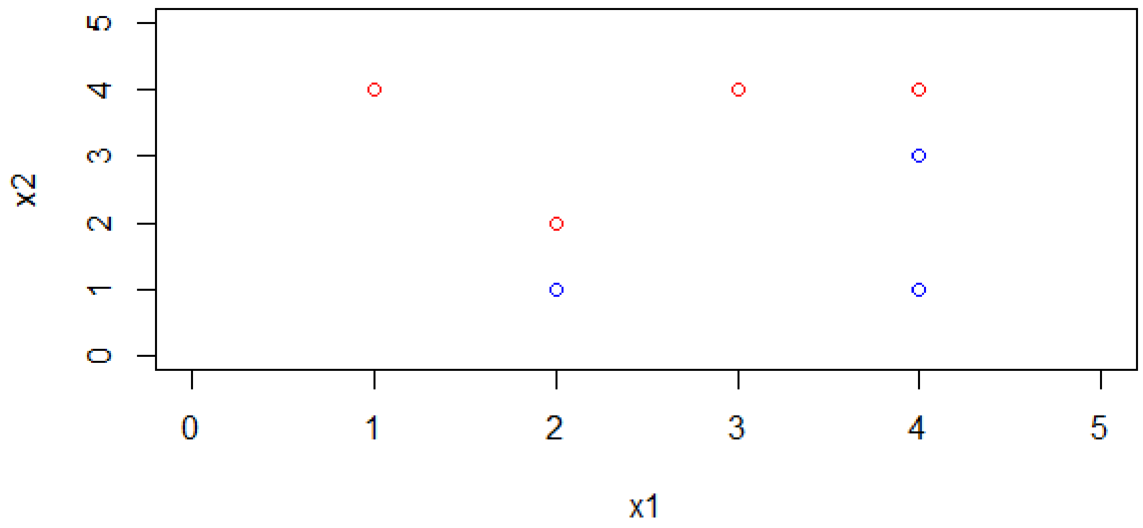
So when X1 = 40, X2 = 3.5, $\hat{p}(X) = 37.75$

(b) For $\hat{p}(X) = 0.5$, we have $0.5 = \frac{exp(-6+0.05X_1+3.5)}{1+exp(-6+0.05X_1+3.5)}$

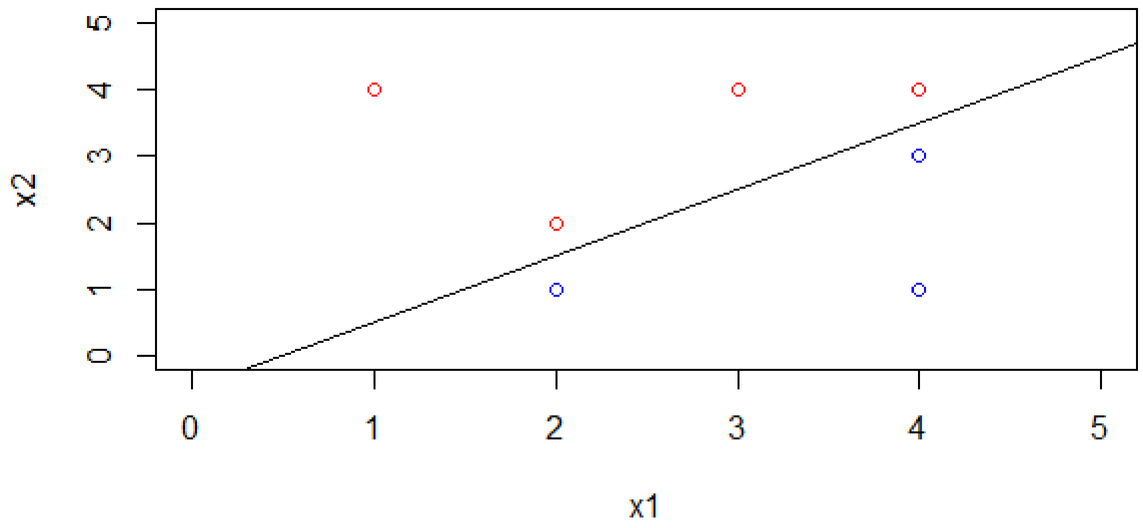
So by calculating this equation, we can get X1 = 50 hours

Chapter 9, Problem 3

(a)



(b)

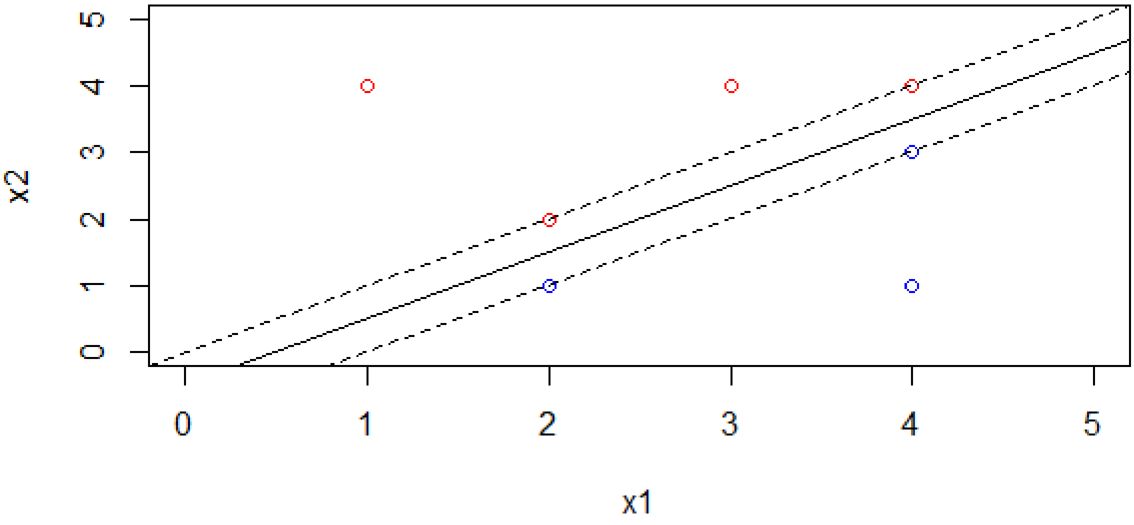


â€‹ For this hyperplane, the equation is $X_2 = -0.5 + X_1$

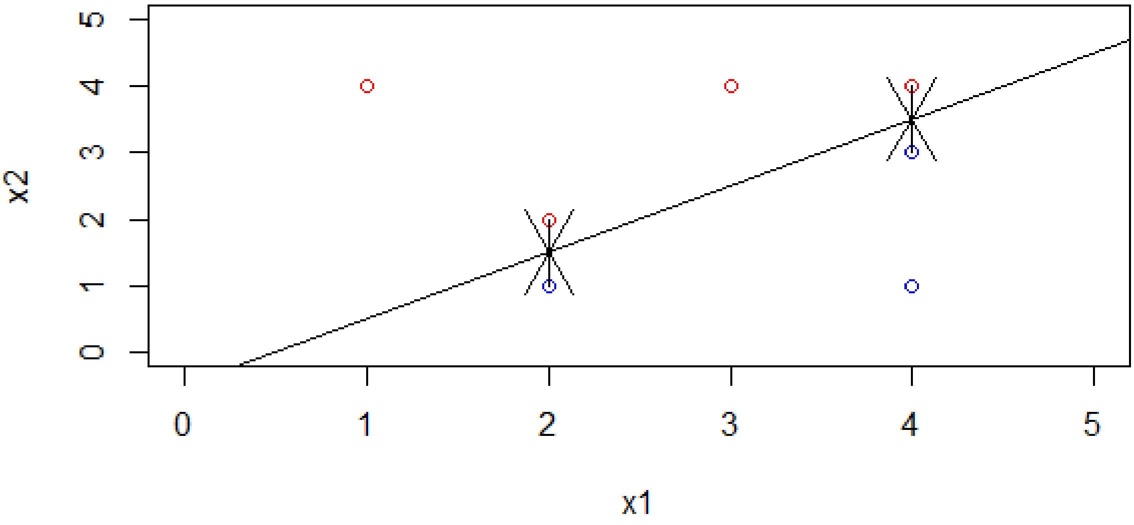
(c) $\beta_0 = 0.5, \beta_1 = -1, \beta_2 = 1$

The classification rule is $0.5 - X_1 + X_2 > 0$ for red and $0.5 - X_1 + X_2 < 0$ for blue

(d)



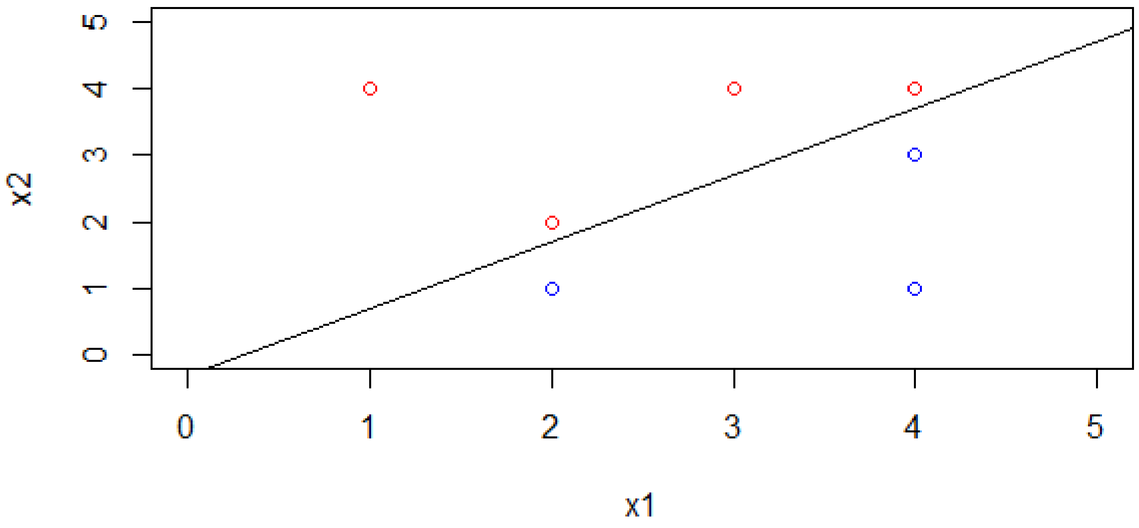
(e)



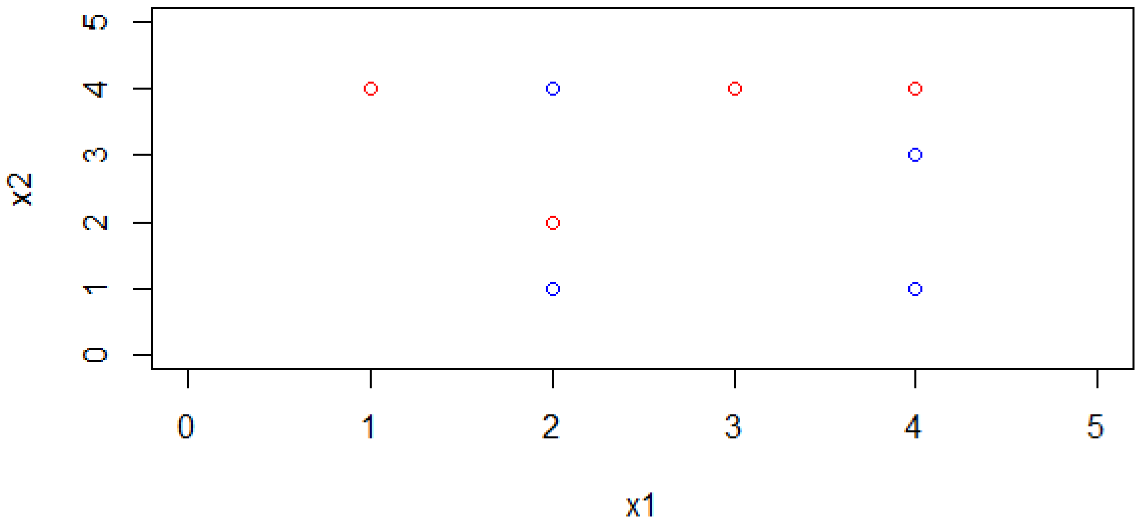
â€‹ The support vectors are the points (2,1), (2,2), (4,3) and (4,4).

(f) If we moved the observation #7 (4,1), the maximal margin hyperplane will not be changed because it is not a support vector.

(g) Hyperplane X2 = -0.3 + X1 is not the optimal hyperplane seperating



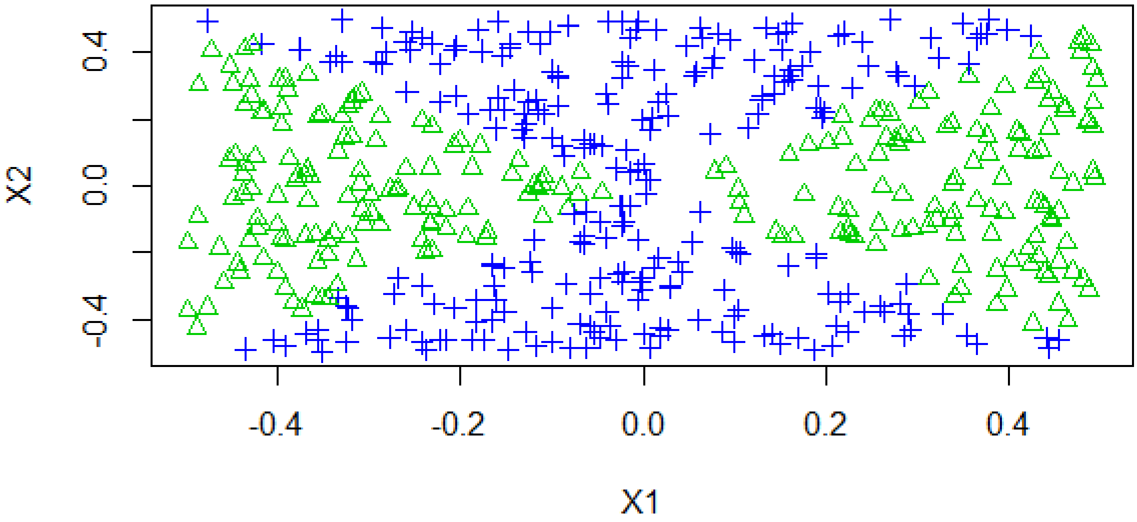
(h)



The additional observation is (2,4) blue

Chapter 9, Problem 5

(b)

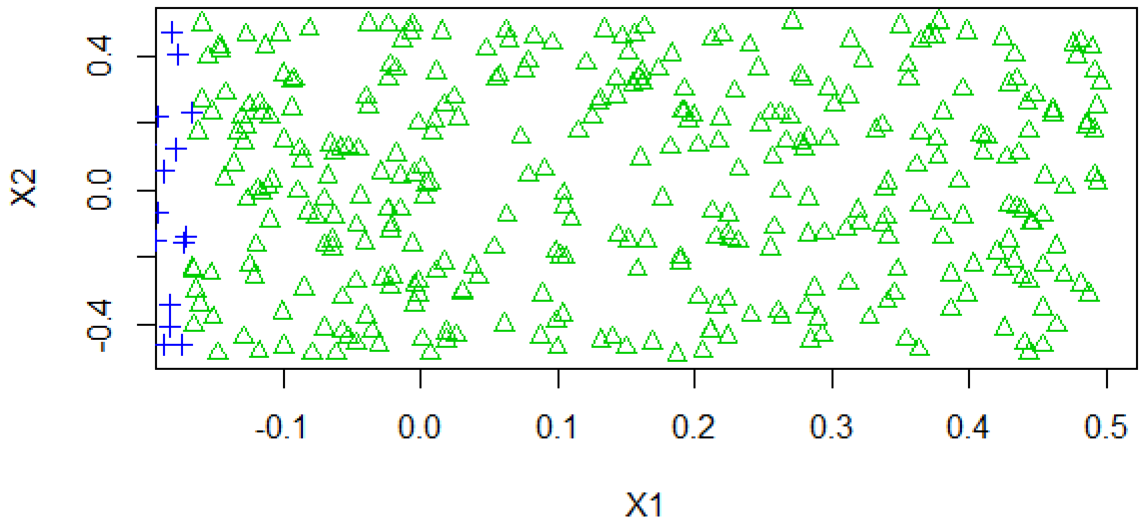


(c) Logistic Regression:

```
Output:
Call:
glm(formula = y ~ x1 + x2, family = "binomial")

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.179   -1.139   -1.112    1.206    1.257

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.087260   0.089579  -0.974    0.330
x1           0.196199   0.316864   0.619    0.536
x2          -0.002854   0.305712  -0.009    0.993

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 692.18  on 499  degrees of freedom
Residual deviance: 691.79  on 497  degrees of freedom
AIC: 697.79

Number of Fisher Scoring iterations: 3
```

Both variables are not statistically significant.

(d)

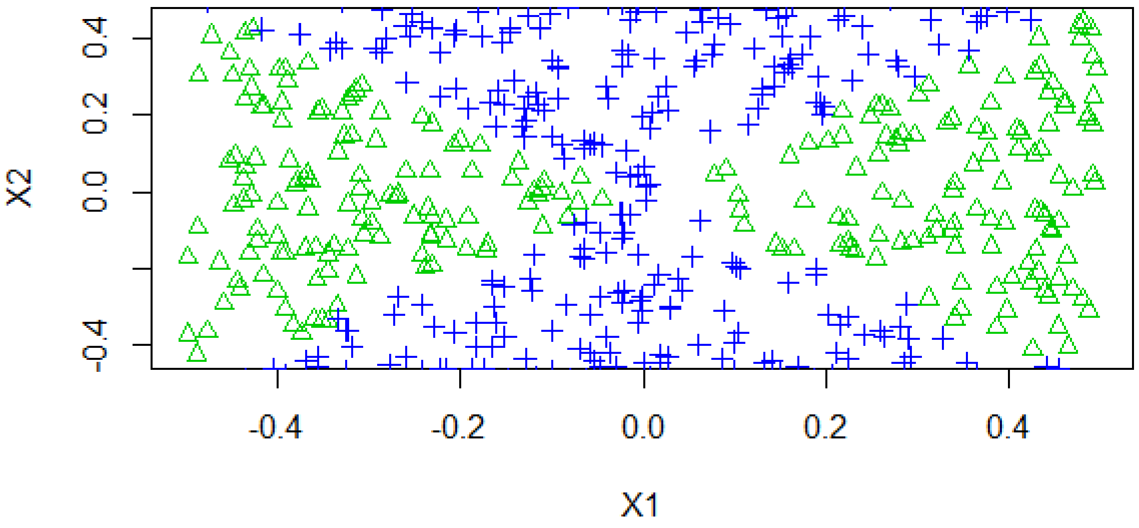


As it shows in figure, this boundary is linear.

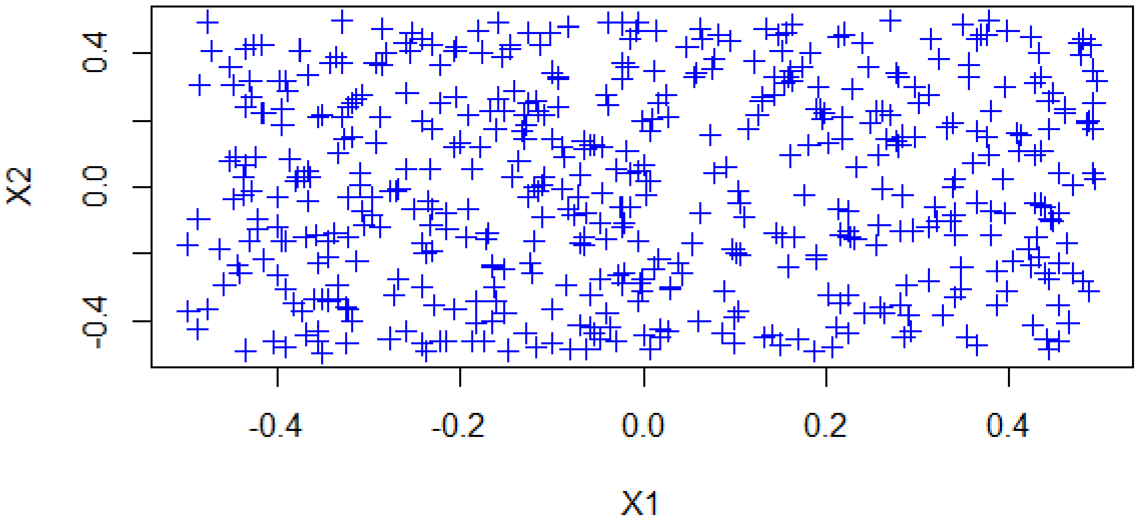
(e) Logistic regression with non-linear

```
Output:
Call:
glm(formula = y ~ poly(x1, 2) + poly(x2, 2) + I(x1 * x2), family = "binomial")

Deviance Residuals: 
       Min          1Q      Median          3Q         Max  
-8.240e-04  -2.000e-08  -2.000e-08   2.000e-08   1.163e-03  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -102.2     4302.0  -0.024    0.981
poly(x1, 2)1   2715.3   141109.5   0.019    0.985
poly(x1, 2)2  27218.5   842987.2   0.032    0.974
poly(x2, 2)1   -279.7    97160.4  -0.003    0.998
poly(x2, 2)2 -28693.0   875451.3  -0.033    0.974
I(x1 * x2)     -206.4    41802.8  -0.005    0.996

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.9218e+02  on 499  degrees of freedom
Residual deviance: 3.5810e-06  on 494  degrees of freedom
AIC: 12

Number of Fisher Scoring iterations: 25
```

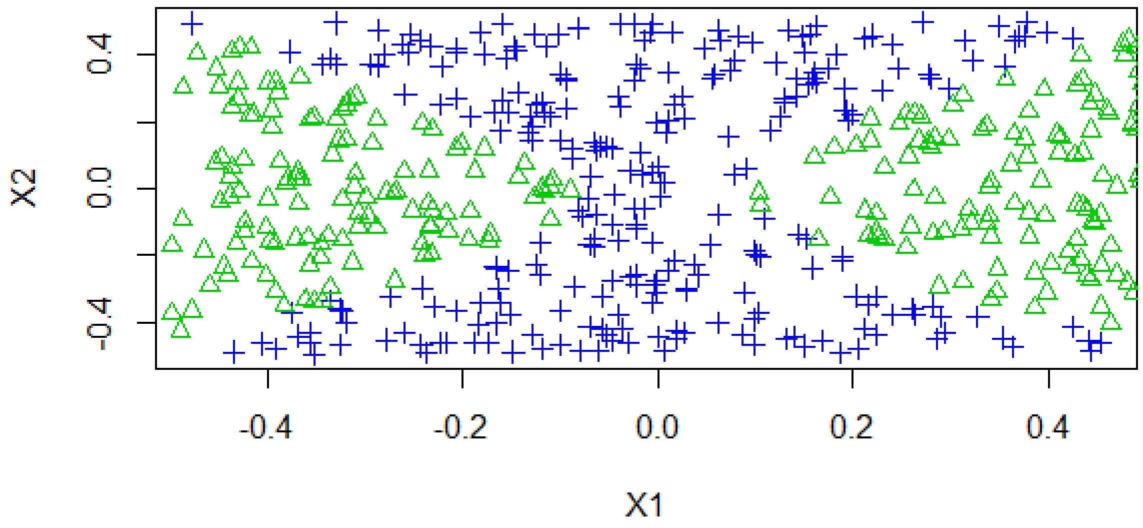
(f)



(g) Support vector classifier:



(h) SVM with non-linear kernel



(i)Conclusion: SVMs with non-linear kernel and logistic regression with interaction terms are good for finding non-linear boundary. Logistic regression with non-interactions and SVMs with linear kernels don't perform well to find the decision boundary.

However, SVM requires only tune gamma but it requires some manual tuning to find the right interaction terms when using logistic regression.

6. Using the Boston data set from the MASS package, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression, LDA, QDA and KNN models using various subsets of the predictors. Also try penalized logistic regression (ridge and lasso), as well as SVM using the optimal choices of tuning parameters for each method. Describe your findings.

Information about variables in Boston data:

```
Boston {MASS}   R Documentation
Housing Values in Suburbs of Boston

Description

The Boston data frame has 506 rows and 14 columns.

Usage

Boston
Format

This data frame contains the following columns:

crim
per capita crime rate by town.

zn
proportion of residential land zoned for lots over 25,000 sq.ft.

indus
proportion of non-retail business acres per town.

chas
Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox
nitrogen oxides concentration (parts per 10 million).

rm
average number of rooms per dwelling.

age
proportion of owner-occupied units built prior to 1940.

dis
weighted mean of distances to five Boston employment centres.

rad
index of accessibility to radial highways.

tax
full-value property-tax rate per \$10,000.

ptratio
pupil-teacher ratio by town.

black
1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

lstat
lower status of the population (percent).
```

```
medv
median value of owner-occupied homes in \$1000s.

Source

Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. J. Environ.
Economics and Management 5, 81–102.

Belsley D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics. Identifying Influential Data
and Sources of Collinearity. New York: Wiley.
```

Logistic regression:

```
Call:  glm(formula = crim01 ~ . - crim01 - crim, family = binomial,
    data = Boston)

Coefficients:
(Intercept)           zn        indus         chas          nox           rm
 -34.103704    -0.079918    -0.059389     0.785327    48.523782    -0.425596
        age          dis          rad          tax      ptratio        black
   0.022172     0.691400     0.656465    -0.006412     0.368716    -0.013524
      lstat         medv
   0.043862     0.167130

Degrees of Freedom: 505 Total (i.e. Null);  492 Residual
Null Deviance:      701.5
Residual Deviance: 211.9    AIC: 239.9
```

```
> mean(pred.glm != crim01.test)
[1] 0.125
```

The test error rate of logistic regression is 12.5%.

LDA

```
> mean(pred.lda$class != crim01.test)
[1] 0.1513158
```

The test error rate of LDA is 15.13%.

QDA

```
> mean(pred.qda$class != crim01.test)
[1] 0.1842105
```

The test error rate of QDA is 18.42%.

KNN

```
> print(min_error_rate)
[1] 0.06578947
> print(K)
[1] 3
```

When k=3, we get the minimum test error rate of KNN, which is 6.57%