# 2020111142_谢嘉薪_Ass3

Xie Jiaxin, 2020111142

# 练习 & 第三次作业

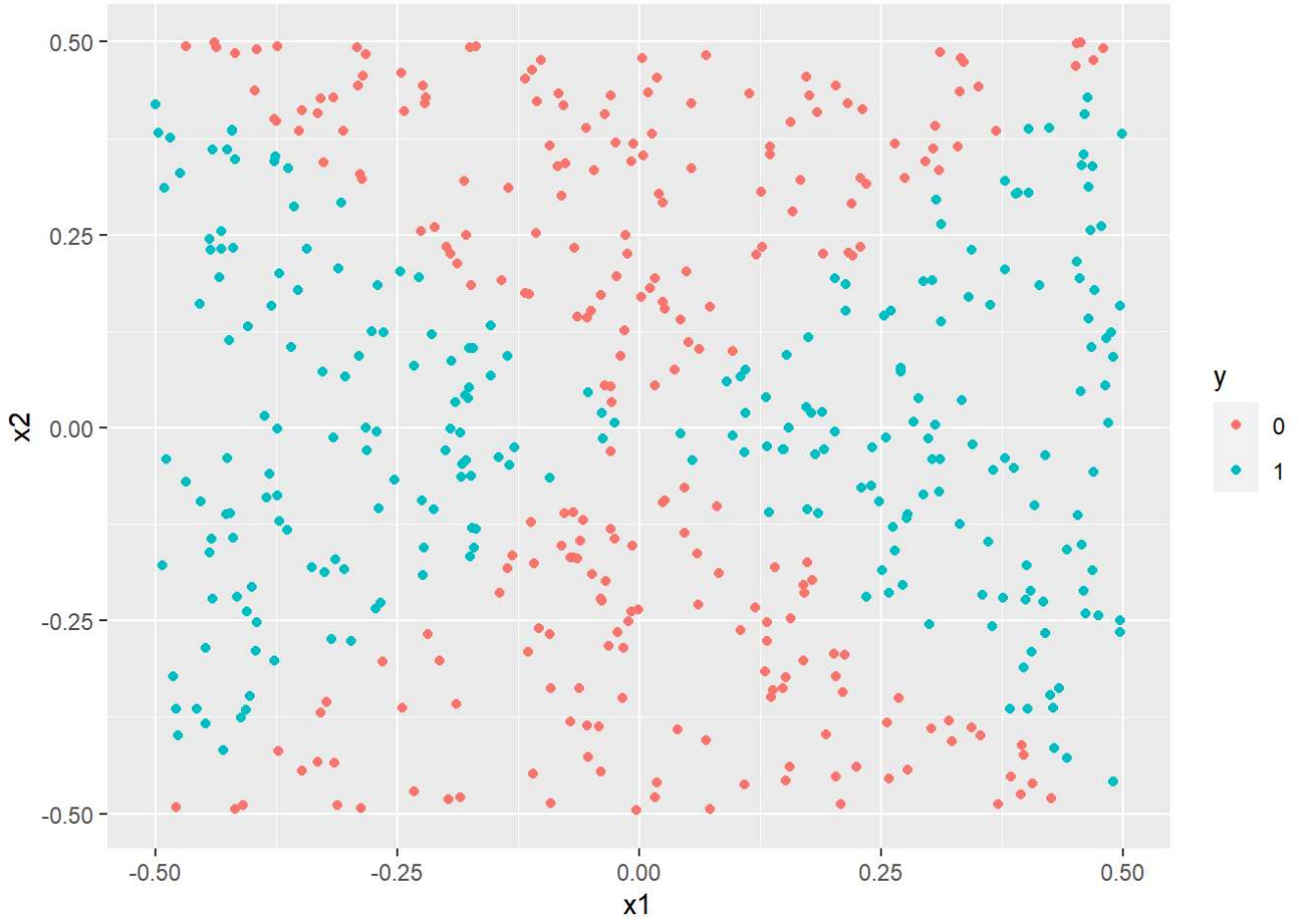
我们已经看到，可以用非线性核拟合SVM，以便使用非线性决策边界执行分类。我们现在将看到，我们还可以通过使用特征的非线性变换执行逻辑回归来获得非线性决策边界。

a. 生成一个$n = 500$且$p = 2$的数据集，使得观测值属于两个类，它们之间具有二次决策边界。例如，可以按如下方式执行：

```
x1 <- runif(500) - 0.5
x2 <- runif(500) - 0.5
y <- 1 * (x1^2 - x2^2 > 0)
```

b. 绘制观测值并按标签赋颜色。在x轴上标注X1，在y轴上标注X2。

```
x <- cbind(x1,x2)
data_b <- data.frame(x = x, y = as.factor(y))
library(ggplot2)
ggplot(data=data_b,aes(x1,x2))+
  geom_point(aes(color=y))
```



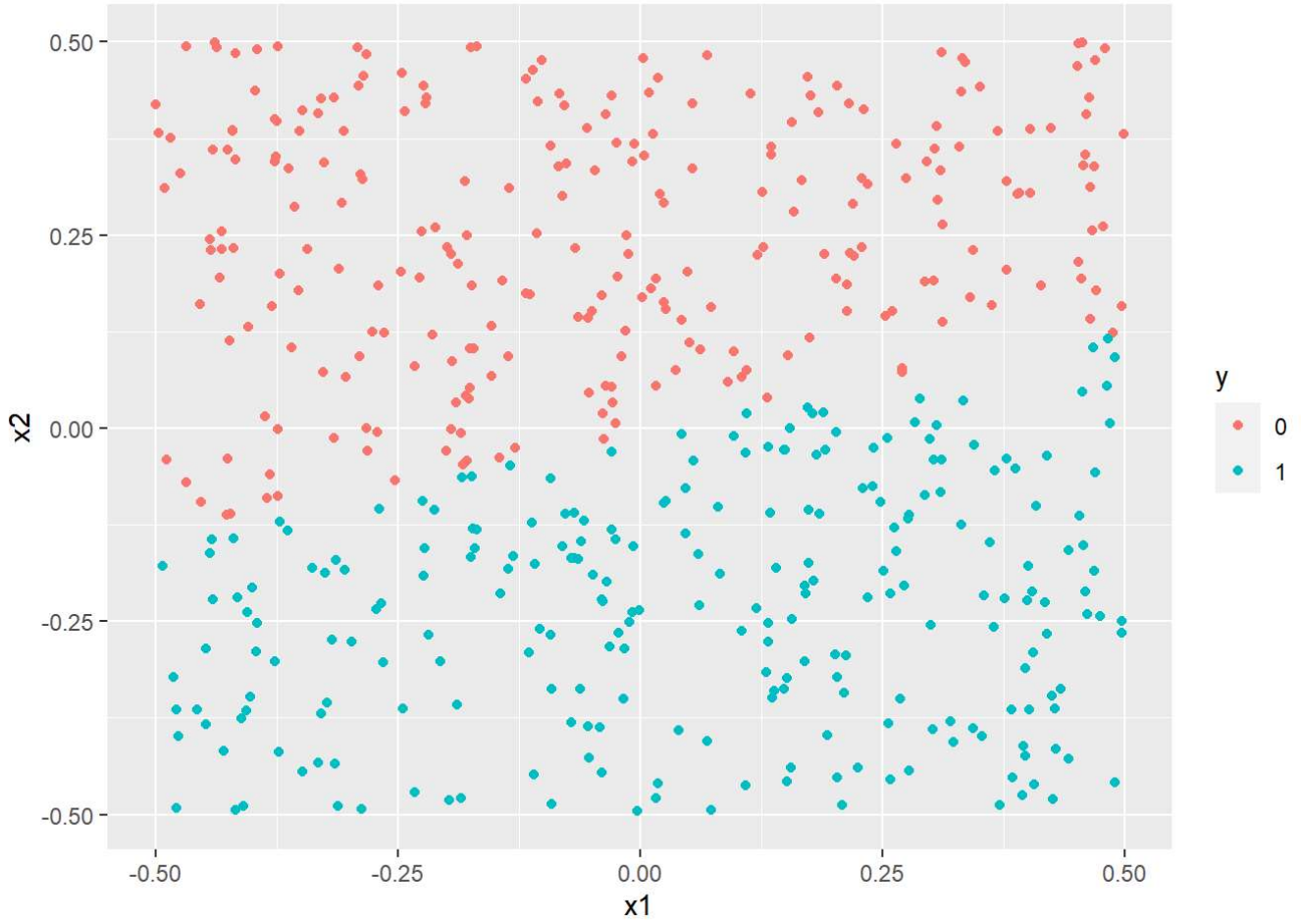

c. 使用X1和X2作为预测变量，对数据拟合逻辑回归模型。

```
data_c <- data.frame(x = x, y = as.factor(y))
glm.fits <- glm(y ~ x1 + x2, data_c, family = binomial)
summary(glm.fits)
```

```
##
## Call:
## glm(formula = y ~ x1 + x2, family = binomial, data = data_c)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.285  -1.154  -1.067   1.176   1.280
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.00260    0.08979  -0.029    0.977
## x1           0.10944    0.31158   0.351    0.725
## x2          -0.42961    0.31411  -1.368    0.171
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 693.14  on 499  degrees of freedom
## Residual deviance: 691.09  on 497  degrees of freedom
## AIC: 697.09
##
## Number of Fisher Scoring iterations: 3
```

d. 将该模型应用于训练数据集，以获得每个训练观测值的预测类标签。绘制观测值，并根据预测的类标签着色。决策边界应为线性。

```
pre_d <- predict(glm.fits,type = "response")
pre_d[pre_d>0.5] <- 1
pre_d[pre_d<=0.5] <- 0
data_d <- data.frame(x = x, y = as.factor(pre_d))
ggplot(data=data_d,aes(x1,x2))+
  geom_point(aes(color=y))
```




```
table(data_b[,"y"],pre_d)
```

```
##    pre_d
##       0   1
##   0 135 116
##   1 121 128
```

```
mean(data_b[,"y"]==pre_d)
```

```
## [1] 0.526
```

e. 使用X1和X2的非线性函数作为预测因子，用逻辑回归模型拟合数据（例如，$X1^2$，$X1 * X2$，$log(X2)$,以此类推）。

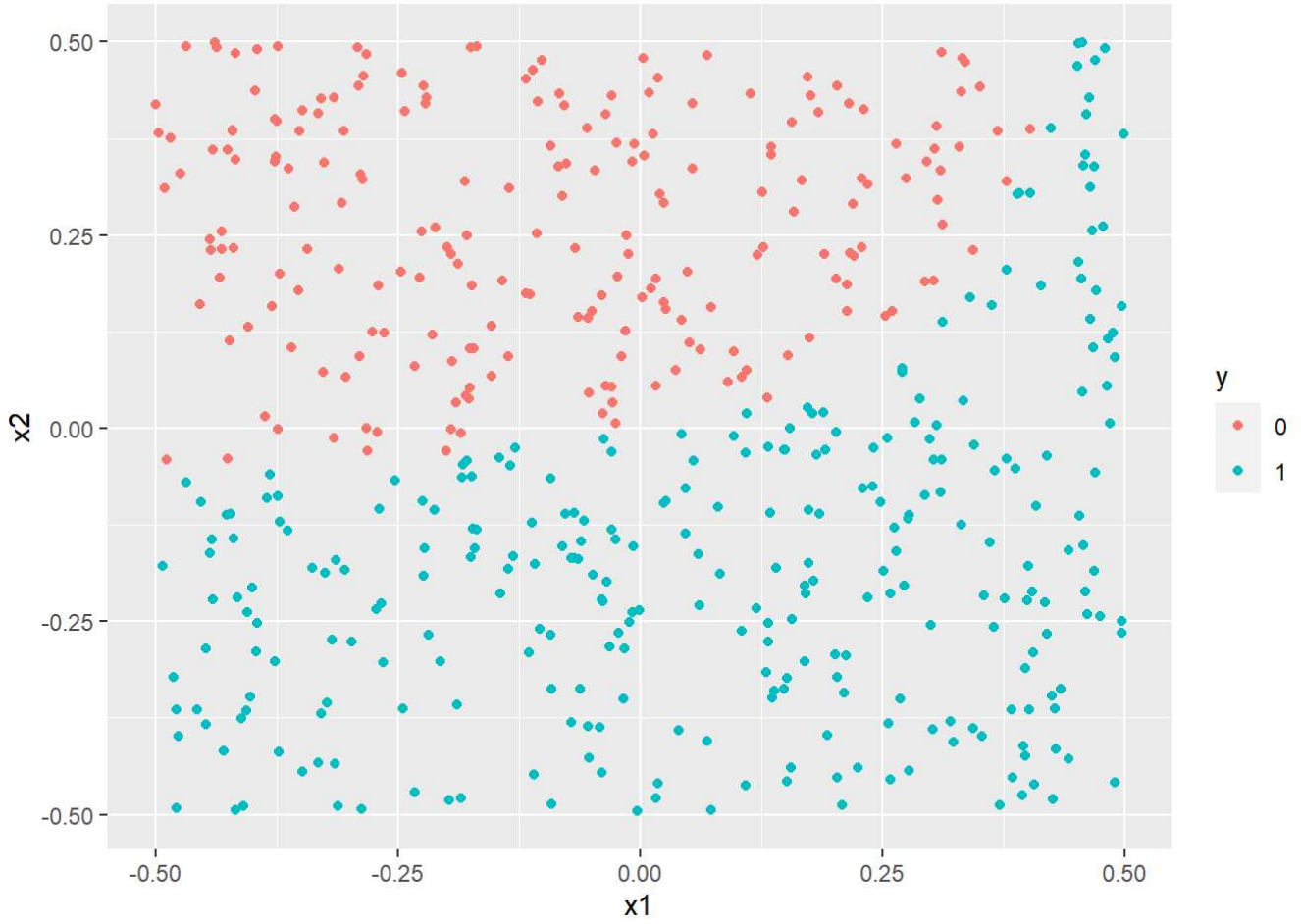
```
data_e <- data.frame(x = x, y = as.factor(y))

glm.fits_e <- glm(y ~ x1^2 + x1*x2, data_e, family = binomial)
summary(glm.fits_e)
```

```
##
## Call:
## glm(formula = y ~ x1^2 + x1 * x2, family = binomial, data = data_e)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3359  -1.1663  -0.9959   1.1667   1.3515
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.0004761  0.0899333   0.005    0.996
## x1           0.0909583  0.3128227   0.291    0.771
## x2          -0.4318930  0.3145375  -1.373    0.170
## x1:x2        0.8337974  1.0367844   0.804    0.421
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 693.14  on 499  degrees of freedom
## Residual deviance: 690.44  on 496  degrees of freedom
## AIC: 698.44
##
## Number of Fisher Scoring iterations: 3
```

f. 将该模型应用于训练数据，以获得每个训练观测值的预测类标签。绘制观测值，根据类标签着色。决策边界应明显为非线性的。如果不是，那么重复(a)-(e)，直到找到一个预测的类标签明显是非线性的例子。

```
pre_f <- predict(glm.fits_e,type = "response")
pre_f[pre_f>0.5] <- 1
pre_f[pre_f<=0.5] <- 0
data_f <- data.frame(x = x, y = as.factor(pre_f))
ggplot(data=data_f,aes(x1,x2))+
  geom_point(aes(color=y))
```




```
table(data_b[,"y"],pre_f)
```

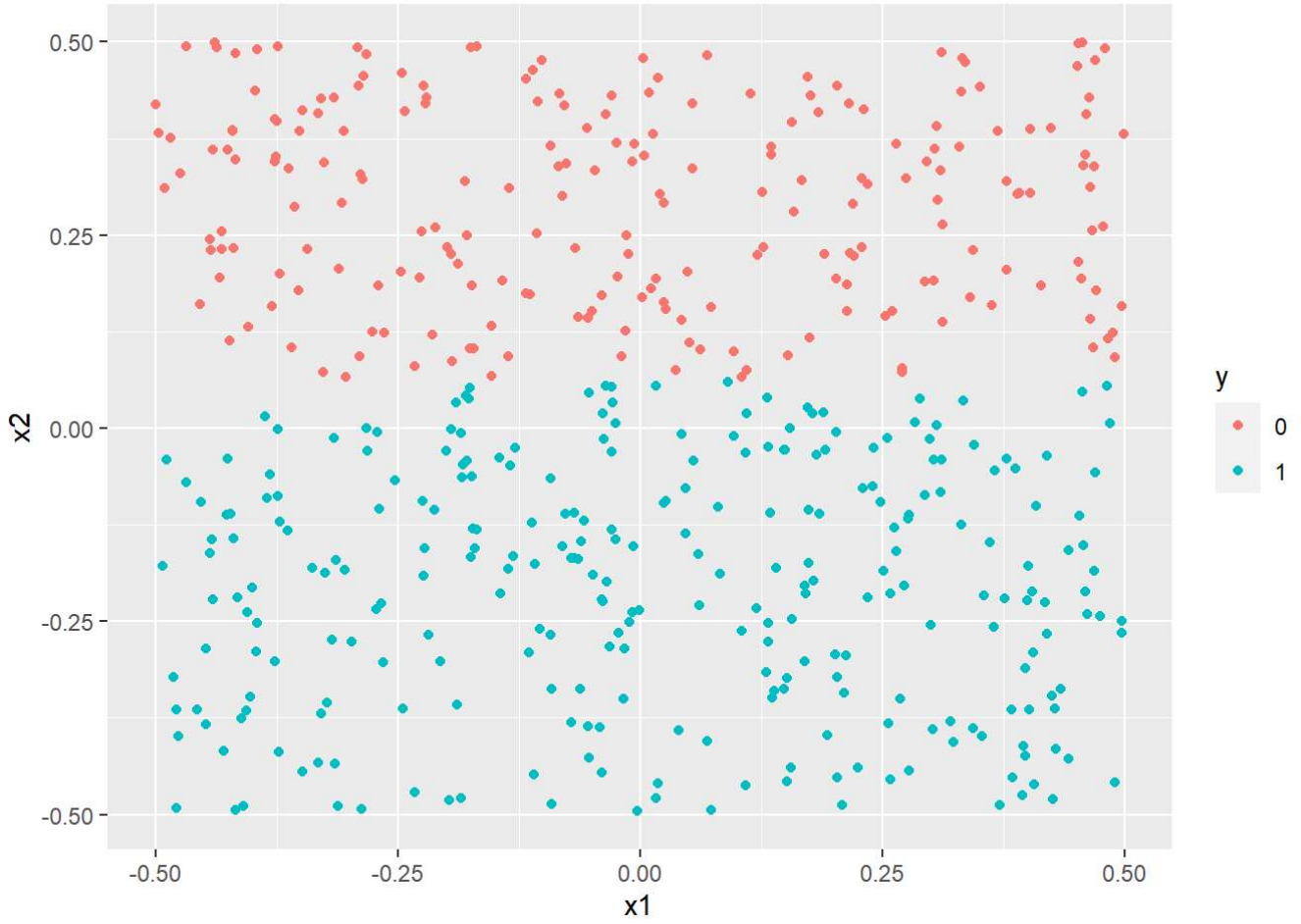
```
##    pre_f
##       0   1
##   0 130 121
##   1  82 167
```

```
mean(data_b[,"y"]==pre_f)
```

```
## [1] 0.594
```

g. 用支持向量分类器拟合数据，并将X1和X2作为预测变量。获得每个训练观测值的分类预测。绘制观测值，并根据预测的类标签着色。

```
library(e1071)
data_g <- data.frame(x = x, y = as.factor(y))
svmfit <- svm(y ~ ., data_g, kernel = "linear", gamma = 1, cost = 10)
pre_g <- predict(svmfit)
data_pre <- data.frame(x=x,y=as.factor(pre_g))
ggplot(data=data_pre,aes(x1,x2))+
  geom_point(aes(color=y))
```




```
table(data_g[,"y"],pre_g)
```

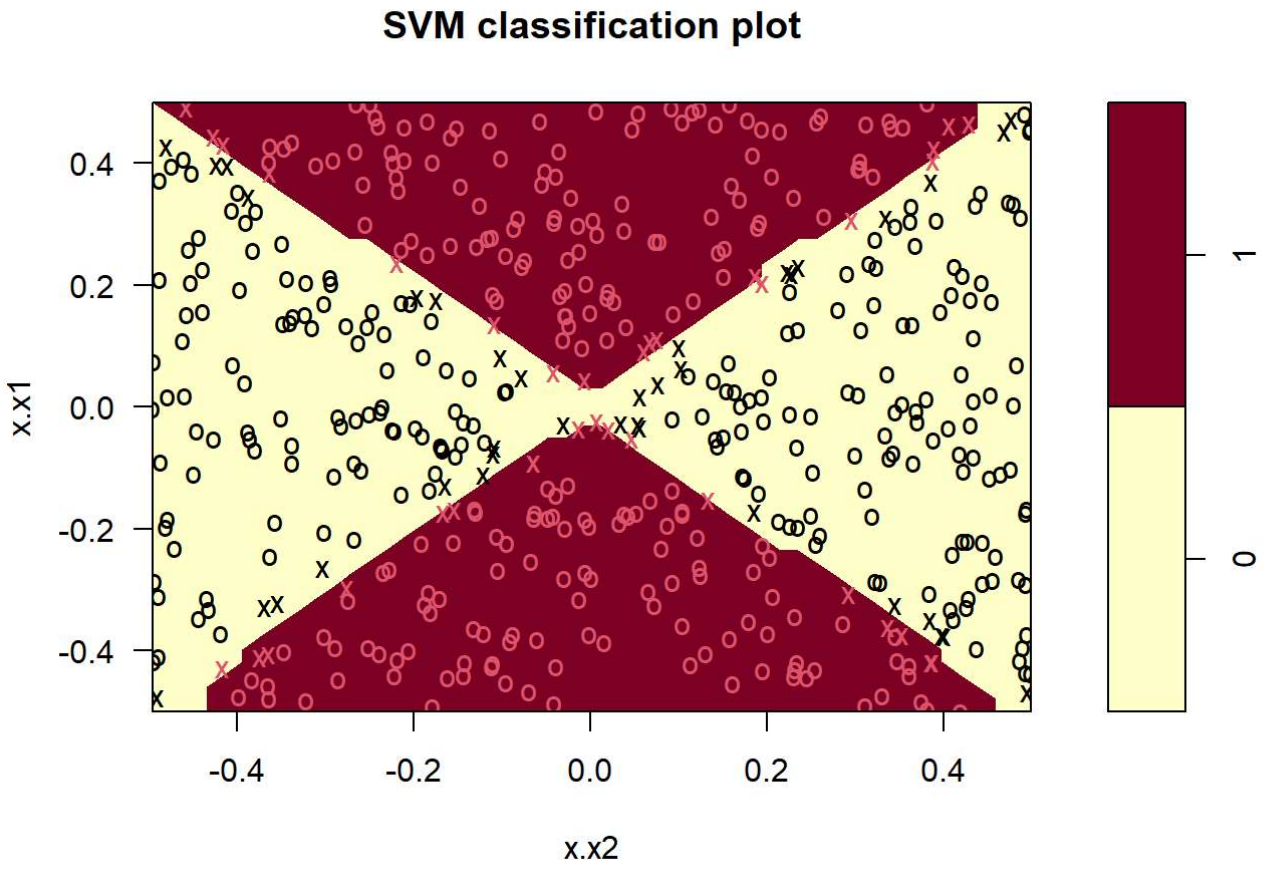
```
##    pre_g
##       0   1
##   0 131 120
##   1  91 158
```

```
mean(data_g[,"y"]==pre_g)
```

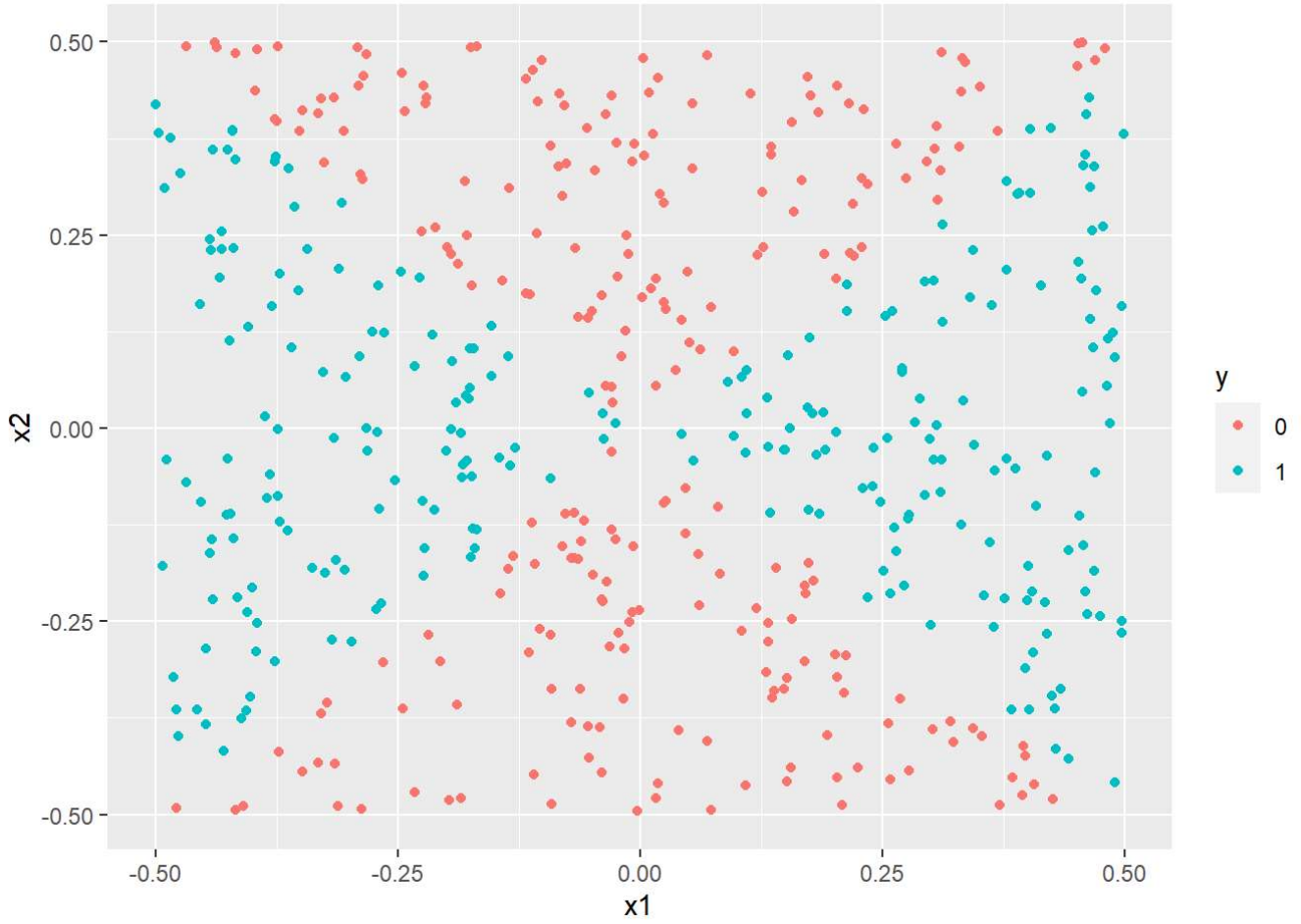
```
## [1] 0.578
```

h. 使用非线性核的SVM拟合数据。获得每个训练观测值的分类预测。绘制观测值，并根据预测的类标签着色。

```
svmfit_h <- svm(y ~ ., data_g, kernel = "radial", gamma = 1, cost = 10)
plot(svmfit_h,data_g)
```




```
pre_h <- predict(svmfit_h)
data_h <- data.frame(x=x,y=as.factor(pre_h))
ggplot(data=data_h,aes(x1,x2))+
  geom_point(aes(color=y))
```




```
table(data_g[,"y"],pre_h)
```

```
##    pre_h
##       0   1
##   0 251   0
##   1   2 247
```

```
mean(data_g[,"y"]==pre_h)
```

```
## [1] 0.996
```

i. 描述你获得的结果。

- 使用X1和X2作为预测变量拟合出来的logistic模型只能获得线性决策边界，且分类效果极差
- 使用X1和X2的非线性函数作为预测因子拟合出来的logistic模型可以获得非线性分类边界。理由：非线性函数可将低维特征映射到高维空间，x1和x2在二维空间非线性可分，但在三维空间线性可分，高维空间的线性决策边界映射在二维平面便是非线性的
- 支持向量机采用线性核函数也只能获得线性决策边界，且分类效果极差
- 支持向量机采用高斯核函数可以获得非线性决策边界,且分类效果最好