Homework1_sol

Lei Shi

12/6/2021

R Markdown

Question 1

(a) For a logistic model we have the success probability

$$\log(\frac{p}{1-p}) = b_0 + b_1 * X_1 + b_2 * X_2,$$

which in turn gives

$$p = \frac{\exp(b_0 + b_1 * X_1 + b_2 * X_2)}{1 + \exp(b_0 + b_1 * X_1 + b_2 * X_2)}.$$

```
b0 <- -4
b1 <- 0.05
b2 <- 1
x1_1 <- 5
x1_2 <- 3.5
p <- exp(b0 + b1*x1_1 + b2*x1_2) / (1 + exp(b0 + b1*x1_1 + b2*x1_2))
cat("The probability of getting an A for this student is: ", round(p, 3))
```

```
## The probability of getting an A for this student is:  0.438
```

(b) Recall definition of the odds:

$$\text{odds} = \frac{p}{1-p}.$$

For logisitic model we have

$$\text{odds} = \frac{p}{1-p} = \exp(b_0 + b_1 * X_1 + b_2 * X_2).$$

```
p_odds <- exp(b0 + b1*x1_1 + b2*x1_2) # = p/(1-p)
cat("The odds of getting an A for this student is: ", round(p_odds, 3))
```

```
## The odds of getting an A for this student is:  0.779
```

(c) Note that $p = 50$ gives a zero log-odds:

$$0 = b_0 + b_1 * X_1 + b_2 * X_2.$$

Solving for X_1 we have

```
x_1_pred <- -(b0 + b2*x1_2)/b1
cat("This student needs to study for", x_1_pred, "hours to get an A.")
```

```
## This student needs to study for 10 hours to get an A.
```

Question 2

For (a) (b) (c) see the plot:

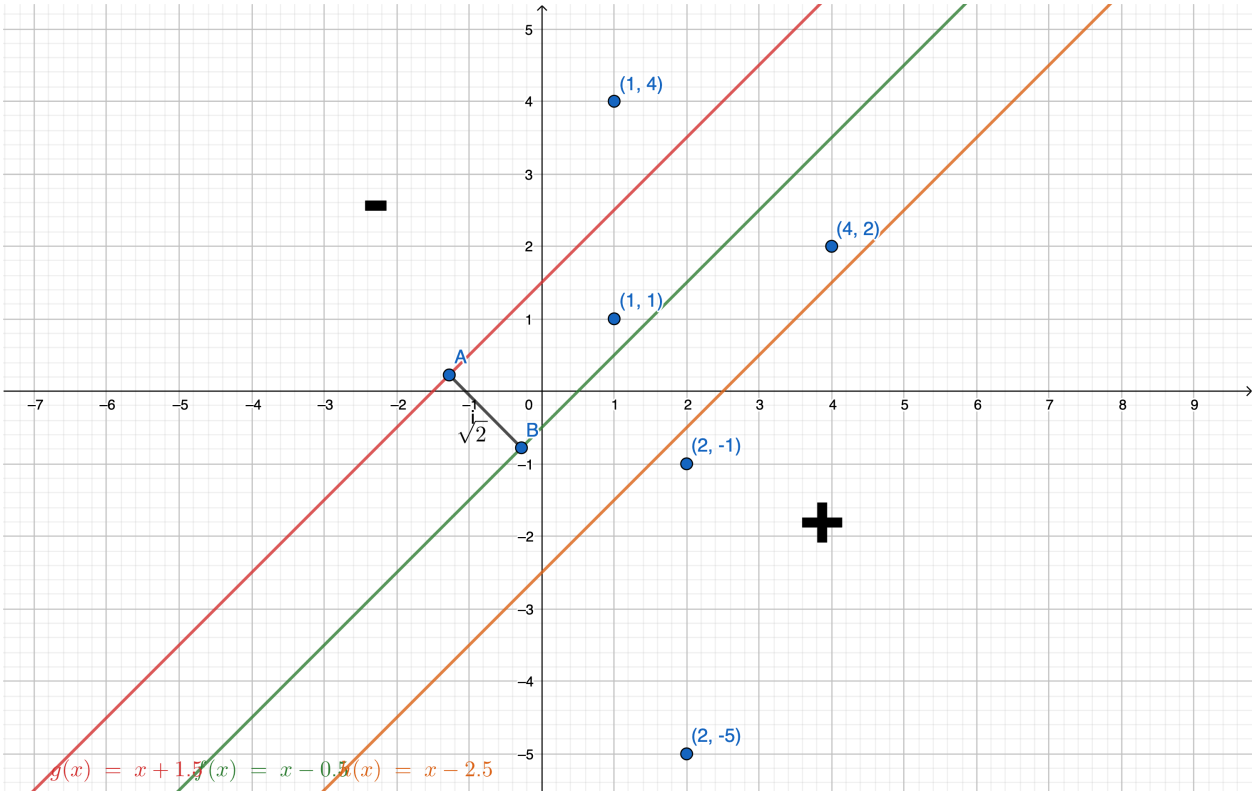

Figure 1: SVM

For part (b), simply notice that the slope of the line is 1 and a distance of $\sqrt{2}$ corresponds to a difference in intercepts: $\pm\sqrt{2}/\cos(45°) = \pm 2$.

The above plot is generated by GeoGebra, a super cool math graphing website. Feel free to try it out: https://www.geogebra.org/?lang=en.

(d) From (b), we can see that $1/\|w\| = \sqrt{2}$(recall how we calculate the margin in the optimization). Therefore, we can obtain $w = (1/2, 1/2)$ after rescaling the coefficients. The margin is given by the formula:

$$\xi = \max\left\{1 - Y\left(\frac{1}{2}X_1 - \frac{1}{2}X_2 - \frac{1}{4}\right), 0\right\}.$$

Now simply plug the test points into the definition:

Points	True label	ξ
(1,4)	-1	0
(1,1)	-1	3/4
(2, -5)	+1	0
(2, -1)	+1	0
(4, 2)	-1	7/4

Question 3

Data importing and preprocessing

```
data <- read.csv("heart_disease.csv")
data <- data %>% filter(Thalassemia != "?") %>% droplevels()
name_lst <- c("Chest_Pain_Type", "Fasting_Blood_Sugar", "Resting_ECG",
"Exercise_Induced_Angina", "ST_Depression_Exercise",
"Peak_Exercise_ST_Segment", "Thalassemia")
factor_lst <- c("Chest_Pain_Type", "Fasting_Blood_Sugar", "Resting_ECG",
"Exercise_Induced_Angina", "Peak_Exercise_ST_Segment" )

x <- data %>% select(all_of(name_lst))
x[factor_lst] <- lapply(x[factor_lst], factor)
y <- as.factor(data$Diagnosis_Heart_Disease)
data_run <- data.frame(y, x)
```

Linear SVM:

```
svm.linear <- svm(y ~ ., data = data_run, kernel = "linear")
summary(svm.linear)
```

```
##
## Call:
## svm(formula = y ~ ., data = data_run, kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##
## Number of Support Vectors:  137
##
##  ( 69 68 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```
support.index <- svm.linear$index
coefs <- svm.linear$coefs
w <- data.frame(value = t(svm.linear$SV)%*%coefs)
w
```

```
##                                value
## Chest_Pain_Type1           0.769309770
## Chest_Pain_Type2          -0.001731588
## Chest_Pain_Type3           0.155706830
## Chest_Pain_Type4          -0.923285012
## Fasting_Blood_Sugar1      -0.065955421
## Resting_ECG1               0.000000000
## Resting_ECG2              -0.529802769
```

```
## Exercise_Induced_Angina1  -0.548814524
## ST_Depression_Exercise    -0.532332166
## Peak_Exercise_ST_Segment2 -0.483938860
## Peak_Exercise_ST_Segment3  0.274404461
## Thalassemia6.0            -0.693203452
## Thalassemia7.0            -1.169822820
```

We measure the importance of the variables according to the magnitude of the svm coefficients. Based on the output, Thalassemia 7.0, chest pain type 1 and 4(corresponding to typical angina and asymptomatic angina), Thalassemia 6.0 seem to have a more important affect on heart disease.

Of course coefficients might not be a good quantification of variable importance: the linear SVM model might not be correct, and the coefficients might not be statistically significant.

Gaussian SVM:

```
svm.gaussian <- svm(y ~ ., data = data_run)
summary(svm.gaussian)
```

```
##
## Call:
## svm(formula = y ~ ., data = data_run)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##
## Number of Support Vectors:  159
##
##  ( 80 79 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```
# In sample prediction
y.pred.gaussian <- predict(svm.gaussian, x)
table(y.pred.gaussian, y)
```

```
##                y
## y.pred.gaussian   0   1
##               0 143  34
##               1  20 104
```

Polynomial

```
svm.poly <- svm(y ~ ., data = data_run, kernel = "polynomial")
summary(svm.poly)
```

```
##
## Call:
```

```
## svm(formula = y ~ ., data = data_run, kernel = "polynomial")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  polynomial
##        cost:  1
##      degree:  3
##      coef.0:  0
##
## Number of Support Vectors:  235
##
##  ( 118 117 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```
# In sample prediction
y.pred.poly <- predict(svm.poly, x)
table(y.pred.poly, y)
```

```
##            y
## y.pred.poly   0   1
##           0 161  81
##           1   2  57
```

Sigmoid

```
svm.sigmoid <- svm(y ~ ., data = data_run, kernel = "sigmoid")
summary(svm.sigmoid)
```

```
##
## Call:
## svm(formula = y ~ ., data = data_run, kernel = "sigmoid")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  sigmoid
##        cost:  1
##      coef.0:  0
##
## Number of Support Vectors:  165
##
##  ( 82 83 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```r
# In sample prediction
y.pred.sigmoid <- predict(svm.sigmoid, x)
table(y.pred.sigmoid, y)
```

```
##                y
## y.pred.sigmoid   0   1
##              0 144  38
##              1  19 100
```