lab1-asn3

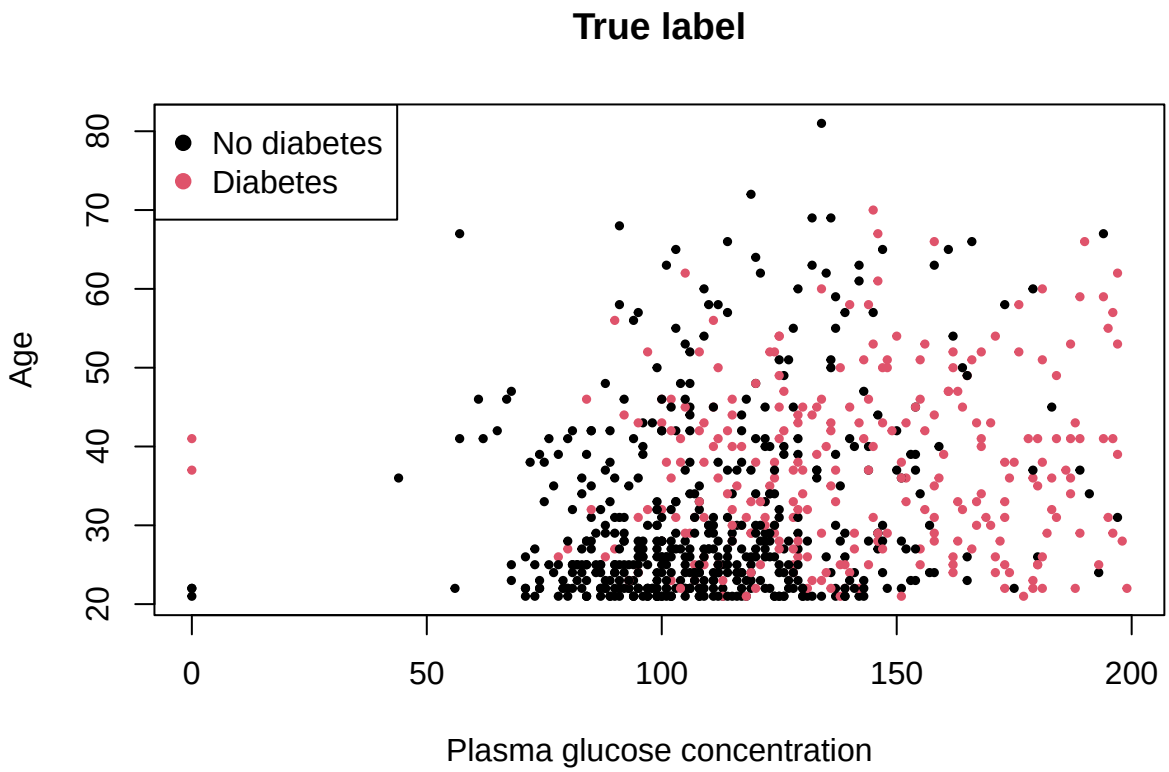
Lepeng Zhang

2023-11-13

Assignment 3. Logistic regression and basis function expansion

1

```
rawdata <- read.csv("pima-indians-diabetes.csv", header = F)

data1 <- rawdata[,c(2,8,9)]
plot(data1$V2, data1$V8, col = data1$V9 + 1, pch = 19, cex = 0.5, xlab = "Plasma glucose concentration"
legend("topleft", legend = c("No diabetes", "Diabetes"), col = c(1, 2), pch = 19)
```



The scatterplot shows that it is quite hard to form a convincing decision boundary. In the black point-clustering area (considering no diabetes), there also exists many red points (with diabetes). This will result in *false negative*, which should avoid as much as possible in this diagnosis problem.

2

```
m1=glm(V9~., data1, family = "binomial")
m1$coefficients
```

```
## (Intercept)          V2          V8
## -5.91244906  0.03564404  0.02477835
```

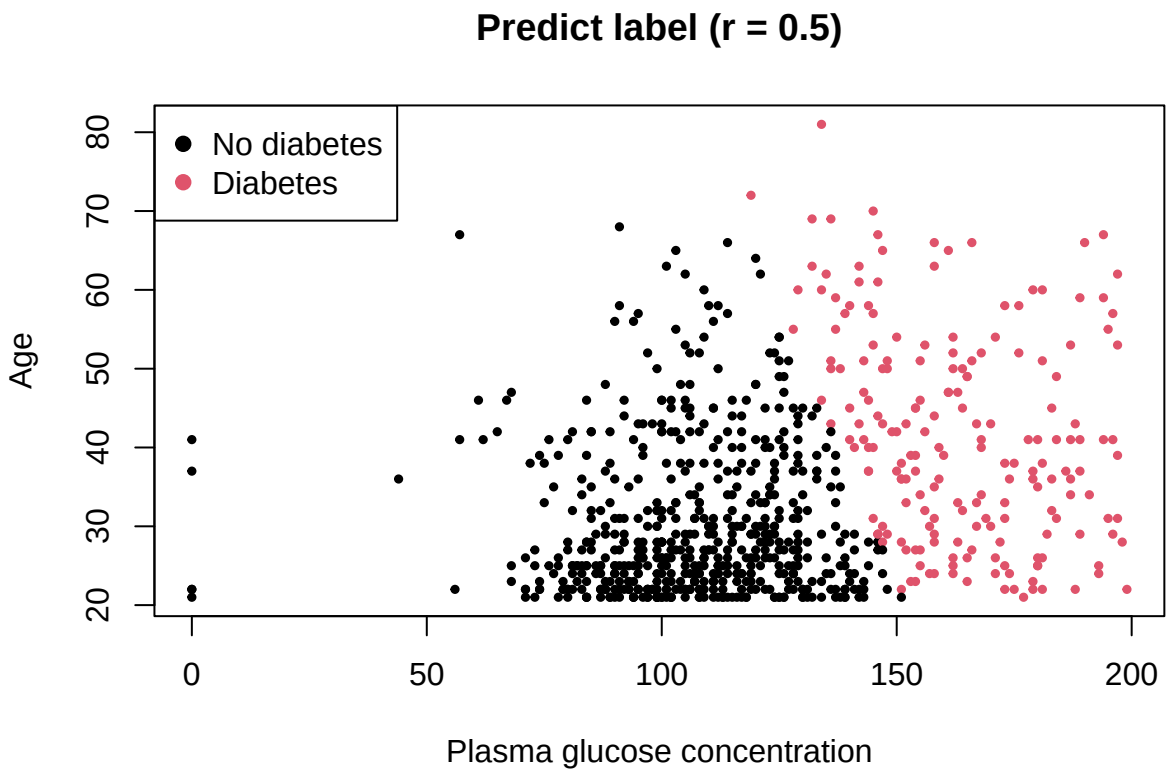
```
Prob=predict(m1, type="response")
pred=ifelse(Prob>0.5, 1, 0)

true <- data1$V9
con_table <- table(true, pred)
con_table
```

```
##     pred
## true   0   1
##    0 436  64
##    1 138 130
```

```
train_error <- 1-sum(diag(con_table))/sum(con_table)
cat("The training misclassification error is:",train_error,"\n")
```

```
## The training misclassification error is: 0.2630208
```

```
plot(data1$V2, data1$V8, col = pred + 1, pch = 19, cex = 0.5, xlab = "Plasma glucose concentration", yla
legend("topleft", legend = c("No diabetes", "Diabetes"), col = c(1, 2), pch = 19)
```



$$p(y = 1 \mid \mathbf{x}) = \frac{1}{1 + exp(5.912 - 0.036x_1 - 0.025x_2)}$$

The training misclassification error means that more than a quarter of cases are misclassified. And the confusion table shows that more than half of the people with diabetes are not diagnosed.

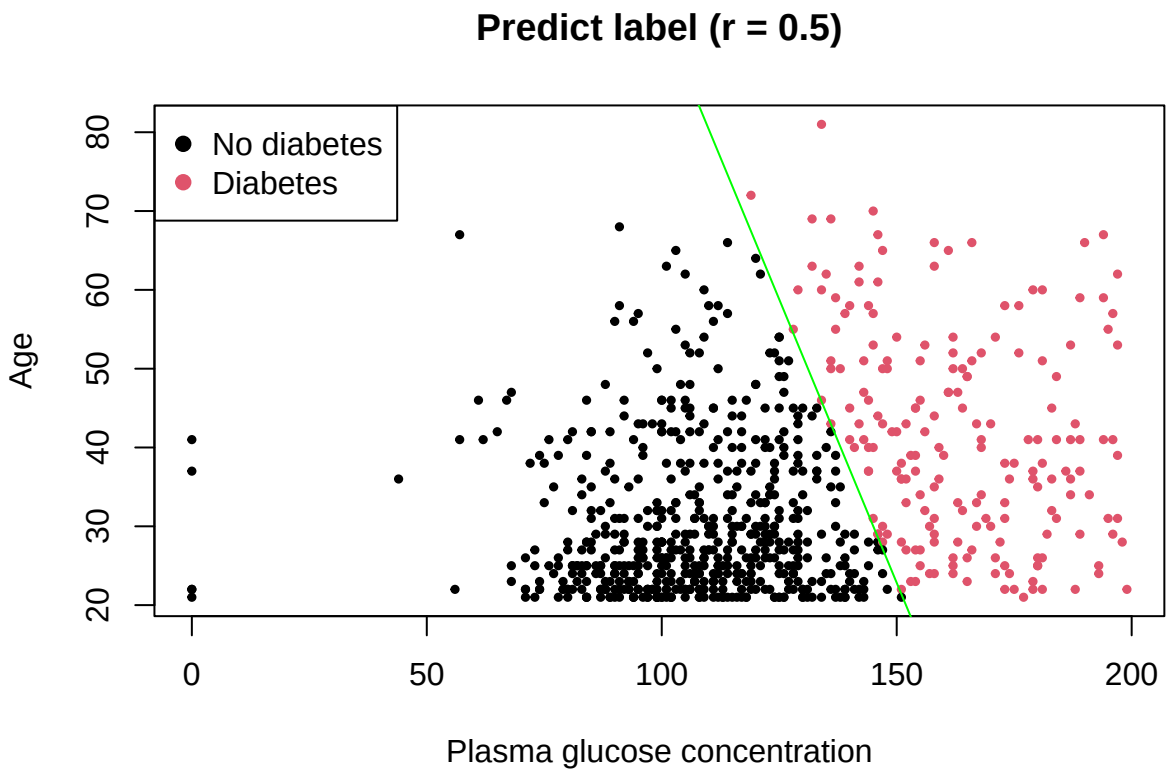
3

Set the above probability equation equals 0.5 to get the equation of the decision boundary:

$$5.912 - 0.036x_1 - 0.025x_2 = 0$$

.

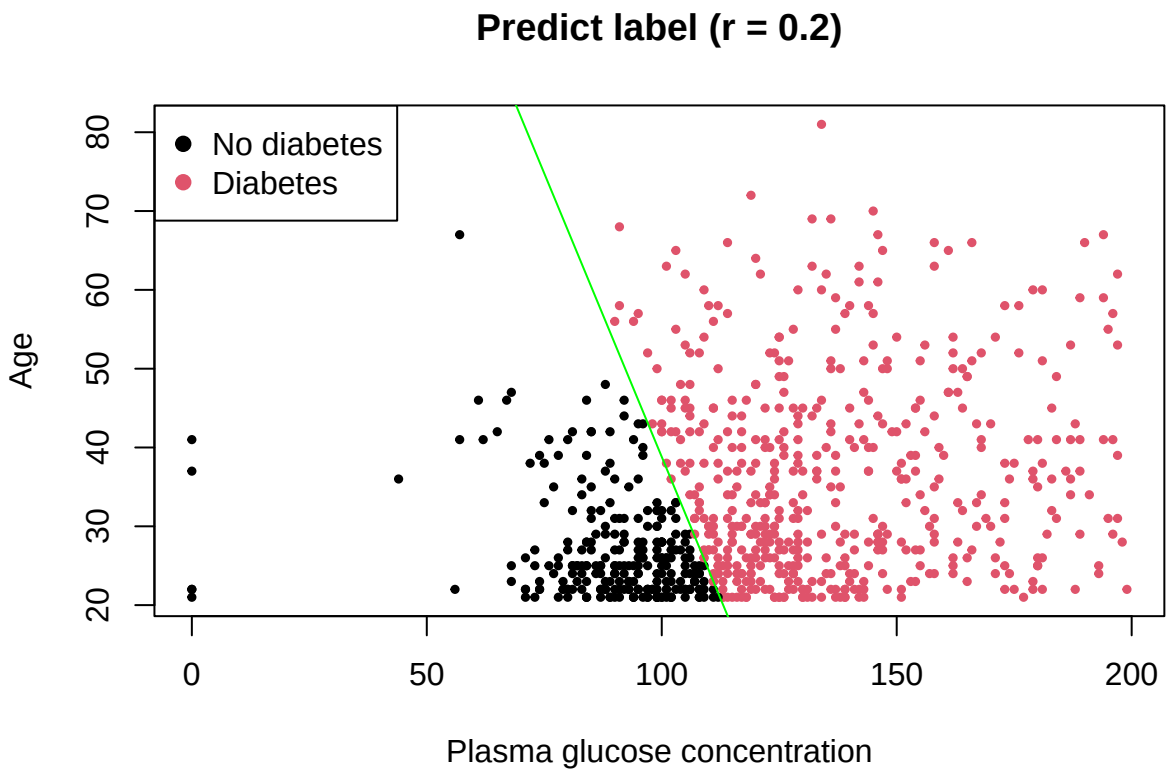
```
plot(data1$V2, data1$V8, col = pred + 1, pch = 19, cex = 0.5, xlab = "Plasma glucose concentration", yla
legend("topleft", legend = c("No diabetes", "Diabetes"), col = c(1, 2), pch = 19)
abline(a = 5.91244906/0.02477835, b = -0.03564404/0.02477835, col = "green")
```

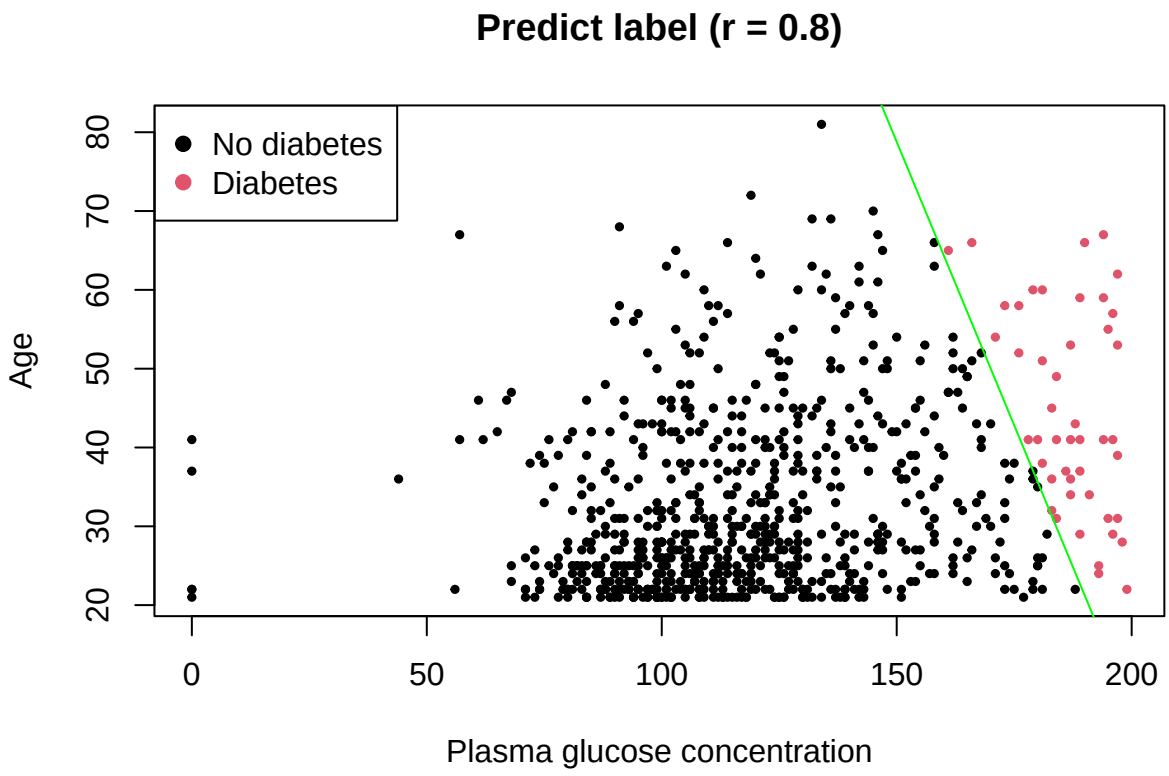


Catch well.

4

```
threshold <- c(0.2,0.8)
store_list <- list()

for (i in 1:length(threshold)){
  store_list$r[i] <- threshold[i]
  pred=ifelse(Prob>threshold[i], 1, 0)
  con_table <- table(true, pred)
  #store_list$confusion_table[i] <- con_table
  store_list$train_error[i] <- 1-sum(diag(con_table))/sum(con_table)

  plot(data1$V2, data1$V8, col = pred + 1, pch = 19, cex = 0.5, xlab = "Plasma glucose concentration", y
  legend("topleft", legend = c("No diabetes", "Diabetes"), col = c(1, 2), pch = 19)
  abline(a = -(log((1-threshold[i])/threshold[i])-5.91244906)/0.02477835, b = -0.03564404/0.02477835, co
}
```


Predict label (r = 0.2)
No diabetes
Diabetes
Age
20
30
40
50
60
70
80
0
50
100
150
200
Plasma glucose concentration




```
store_df <- as.data.frame(store_list)
print(store_df)
```

```
##     r train_error
## 1 0.2   0.3723958
## 2 0.8   0.3151042
```

The decision boundary moves right as r value increases, making less people can be predicted as diabetes.

5

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
data1 <- data1%>%mutate(Z1=V2^4, Z2=V2^3*V8, Z3=V2^2*V8^2, Z4=V2*V8^3, Z5=V8^4)

m1=glm(V9~., data1, family = "binomial")
m1$coefficients
```

```
##   (Intercept)            V2            V8            Z1            Z2
## -9.309821e+00  3.793014e-02  1.456805e-01  1.278015e-08 -1.779600e-07
##            Z3            Z4            Z5
##  8.515150e-07 -1.698011e-06  8.126623e-07
```

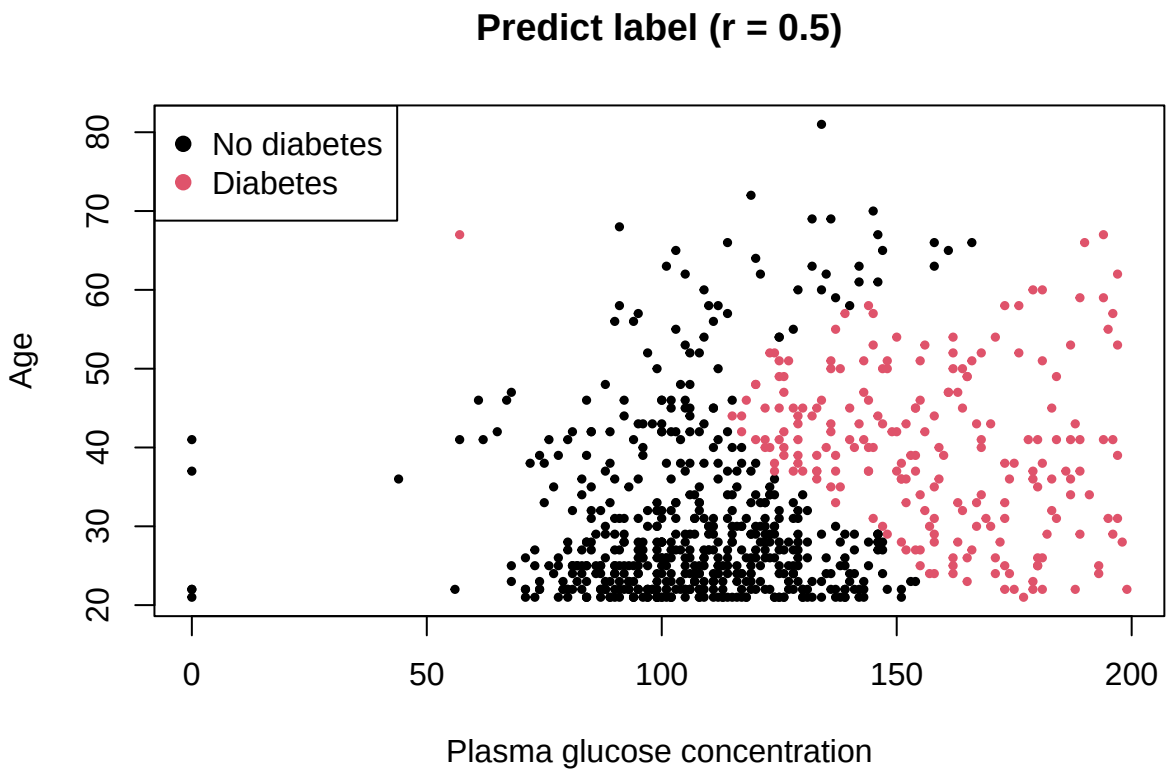
```r
Prob=predict(m1, type="response")
pred=ifelse(Prob>0.5, 1, 0)

true <- data1$V9
con_table <- table(true, pred)
con_table
```

```
##     pred
## true   0   1
##    0 433  67
##    1 121 147
```

```r
train_error <- 1-sum(diag(con_table))/sum(con_table)
cat("The training misclassification error is:",train_error,"\n")
```

```
## The training misclassification error is: 0.2447917
```

```r
plot(data1$V2, data1$V8, col = pred + 1, pch = 19, cex = 0.5, xlab = "Plasma glucose concentration", yla
legend("topleft", legend = c("No diabetes", "Diabetes"), col = c(1, 2), pch = 19)
```



The basis expansion trick makes the desicion boundary change from straight line to curved line. But the training misclassification rate does not improve much, still remaining closed to a quarter. From the confusion table, it can be seen that there are 45.1% people with diabetes who are not diagnosed. Therefore, this model still does a poor classification work.