Assignment3

Priyarani Patil

2023-11-20

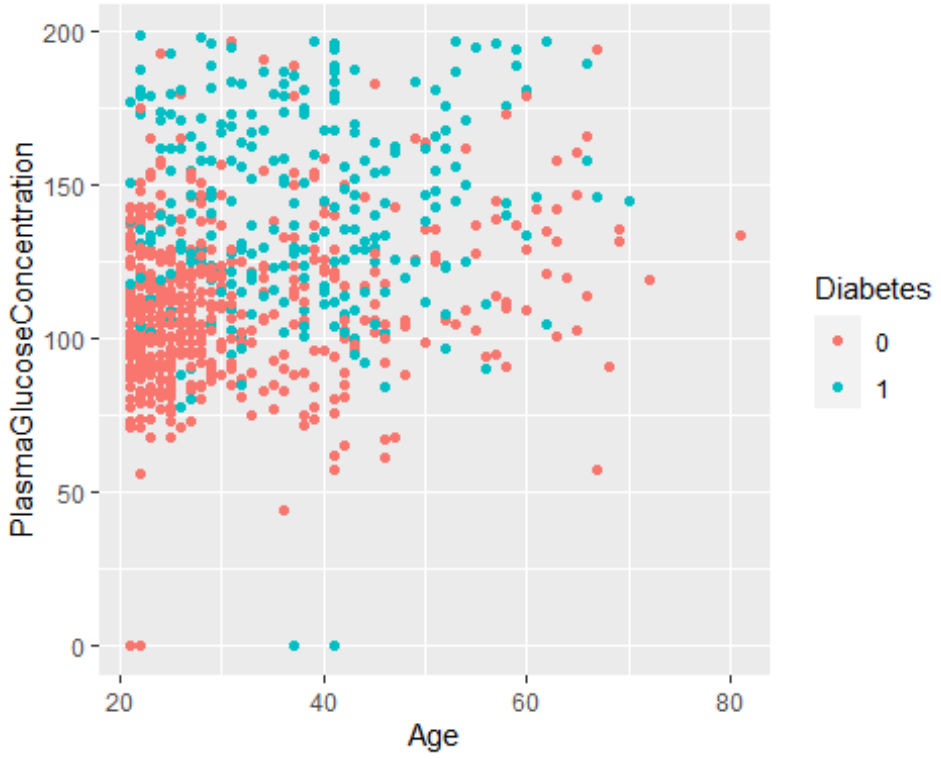
R Markdown

Assignment 3. Logistic regression and basis function expansion
1. Make a scatterplot showing a Plasma glucose concentration on Age where observations are colored by Diabetes levels. Do you think that Diabetes is easy to classify by a standard logistic regression model that uses these two variables as features? Motivate your answer.

```
setwd("C:/Users/priya/OneDrive/Desktop/Machine learning/Lab Assignments")

library("ggplot2")
#datain the form of dataframe
data <- read.csv("pima-indians-diabetes.csv")
#Modifying column names
colnames(data)[c(2,8,9)]<-c('PlasmaGlucoseConcentration','Age','Diabetes')

data$Diabetes <- as.factor(data$Diabetes)
plot1<- ggplot(data=data,
aes(x=Age,y=PlasmaGlucoseConcentration))+geom_point(aes(col=Diabetes))
print(plot1)
```



The logistic regression model can effectively classify diabetes using these two variables. This conclusion is drawn from the plot, which indicates that individuals with an age less than 35 years and a Plasma Glucose Concentration below 150 are less tendency to having diabetes.

2.Train a logistic regression model with □□ =Diabetes as target □□1 =Plasma glucose concentration and □□2 =Age as features and make a prediction for all observations by using □□ = 0.5 as the classification threshold. Report the probabilistic equation of the estimated model (i.e., how the target depends on the features and the estimated model parameters probabilistically). Compute also the training misclassification error and make a scatter plot of the same kind as in step 1 but showing the predicted values of Diabetes as a color instead. Comment on the quality of the classification by using these results.

```
#training a Logistic Regression model
trained_model <- glm(Diabetes ~ PlasmaGlucoseConcentration + Age, data =
data, family= "binomial")
#prediction of all observations
prediction<- predict(trained_model, data, type = "response")
prediction1<- ifelse(prediction>0.5,1,0)

trained_model$coefficients

##                (Intercept) PlasmaGlucoseConcentration
##                -5.89785793                 0.03558250
```

```
##                         Age
##                  0.02450157
```

$$p(\text{Diabetes} = 1) = \frac{1}{1 + e^{-5.89785793+0.03558250\times\text{PlasmaGlucoseConcentration}+0.02450157\times\text{Age}}}$$

Misclassification error

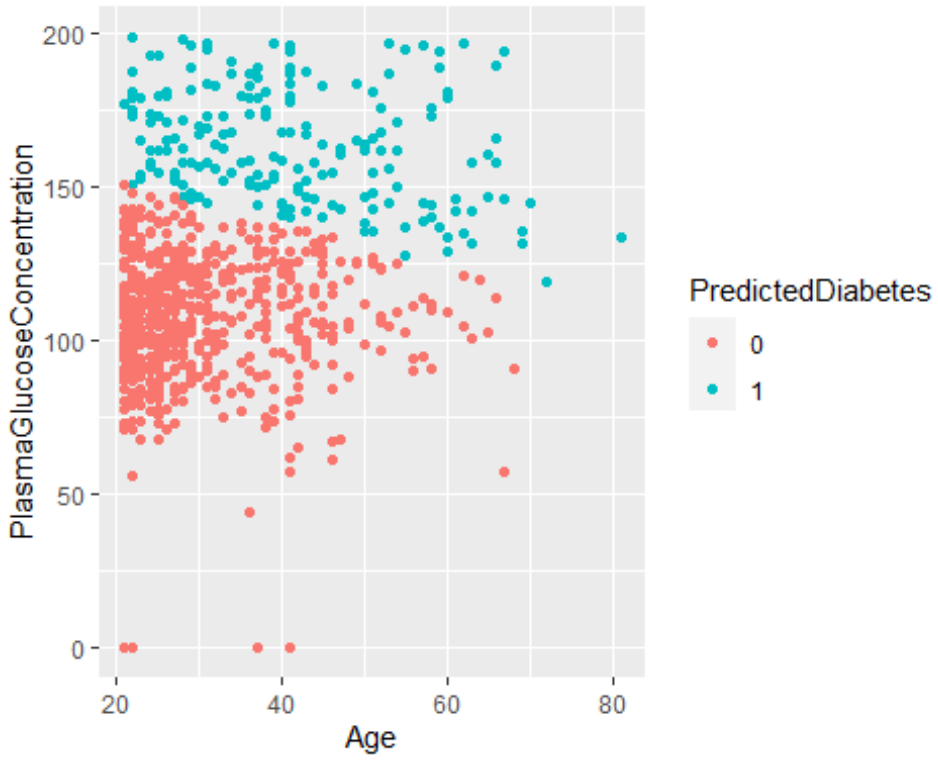
```
#Model accuracy using confusion matrix
table(data$Diabetes,prediction1)

##    prediction1
##       0   1
##   0 436  64
##   1 140 127

#A function to calculate misclassification error
missclass<- function(pred,actual){
 n=length(pred)
 return(1-sum(diag(table(pred, actual)))/n)
}
#Computing misclassification error for r= 0.5
missclass(data$Diabetes,prediction1)

## [1] 0.2659713

#scatter plot with predicted values of diabetes
data$PredictedDiabetes <- as.factor(prediction1)
plot2<- ggplot(data=data,
aes(x=Age,y=PlasmaGlucoseConcentration))+geom_point(aes(col=PredictedDiabetes
))
print(plot2)
```



Analyzing the misclassification error reveals that approximately 26% of observations are inaccurately classified. The graphical representation indicates accurate classification for individuals with a Plasma Glucose Concentration below 150 and an age below 35 years.However, individuals aged above 35 years tend to be predominantly misclassified.

3.Use the model estimated in step 2 to a) report the equation of the decision boundary between the two classes b) add a curve showing this boundary to the scatter plot in step 2. Comment whether the decision boundary seems to catch the data distribution well.

```
#plotting the boundaries
Ag1<-15:85 #considering Age as Ag1
Ag2<-(5.89785793/0.03558250)-(0.02450157/0.03558250)*Ag1 # #taking Plasma
Glucose as Ag2
boundary_df<-data.frame(Ag1,Ag2 )
plot3<-plot2 + geom_line(data = boundary_df, aes(x=Ag1,y=Ag2))
print(plot3)
```

The decision boundary inadequately captures the distribution of the data, as numerous points indicating diabetes are scattered on both sides of the line.

4.Make same kind of plots as in step 2 but use thresholds □□ = 0.2 and □□ = 0.8. By using these plots, comment on what happens with the prediction when □□ value changes.

```
#prediction with r=0.2
prediction2<- ifelse(prediction>0.2,1,0)
data$PredictedDiabetes2 <- as.factor(prediction2)


#Computing model accuracy using confusion matrix
table(data$Diabetes,prediction2)

##    prediction2
##       0   1
##   0 238 262
##   1  25 242

#Computing misclassification error
missclass(data$Diabetes,prediction2)

## [1] 0.3741851

#scatter plot of classification with r=0.2
plot4<- ggplot(data=data,
```

```
aes(x=Age,y=PlasmaGlucoseConcentration))+geom_point(aes(col=PredictedDiabetes
2))
plot4<-plot4 + geom_line(data = boundary_df, aes(x=Ag1,y=Ag2))
print(plot4)
```


```
#prediction with r=0.8
prediction3<- ifelse(prediction>0.8,1,0)
data$PredictedDiabetes3 <- as.factor(prediction3)

#Computing model accuracy using confusion matrix
table(data$Diabetes,prediction3)

##    prediction3
##       0   1
##   0 490  10
##   1 231  36

#Computing misclassification error
missclass(data$Diabetes,prediction3)

## [1] 0.3142112

#scatter plot of classification with r=0.8
plot5<- ggplot(data=data,
aes(x=Age,y=PlasmaGlucoseConcentration))+geom_point(aes(col=PredictedDiabetes
3))
plot5<-plot5 + geom_line(data = boundary_df, aes(x=Ag1,y=Ag2))
print(plot5)
```

The models with correlation coefficients of 0.2 and 0.8 have misclassified a substantial number of observations when contrasted with the model having a correlation coefficient of 0.5. Specifically, the model with a correlation coefficient of 0.2 has erroneously labeled numerous non-diabetic individuals as diabetic,while the model with a correlation coefficient of 0.8 has misclassified many diabetic individuals as non-diabetic.

5.Perform a basis function expansion trick by computing new features □□1 = □□1 4, □□2 = □□1 3□□2, □□3 = □□1 2□□2 2, □□4 = □□1□□2 3, □□5 = □□2 4 , adding them to the data set and then computing a logistic regression model with □□ as target and □□1, □□2, □□1, … , □□5 as features. Create a scatterplot of the same kind as in step 2 for this model and compute the training misclassification rate. What can you say about the quality of this model compared to the previous logistic regression model? How have the basis expansion trick affected the shape of the decision boundary and the prediction accuracy?

```
#computing and adding new features to the data set
data$z1<-data$PlasmaGlucoseConcentration^4
data$z2<-(data$PlasmaGlucoseConcentration^3)*(data$Age)
data$z3<-(data$PlasmaGlucoseConcentration^2)*(data$Age^2)
data$z4<-(data$PlasmaGlucoseConcentration)*(data$Age^3)
data$z5<-data$Age^4
new_train_model <- glm(Diabetes ~
PlasmaGlucoseConcentration+Age+z1+z2+z3+z4+z5, data = data, family=
"binomial")
new_prediction<- predict(new_train_model, data, type = "response")
new_prediction<- ifelse(new_prediction>0.5,1,0)
```

```r
#accuracy using confusion matrix
table(data$Diabetes,new_prediction)

##    new_prediction
##       0   1
##   0 433  67
##   1 122 145

#misclassification error
missclass(data$Diabetes,new_prediction)

## [1] 0.2464146

#scatter plot of classification with new features
data$NewPredictedDiabetes <- as.factor(new_prediction)
plot6<- ggplot(data=data,
aes(x=Age,y=PlasmaGlucoseConcentration))+geom_point(aes(col=NewPredictedDiabe
tes))
print(plot6)
```

In comparison to the preceding model, the misclassification error in this model is lower, indicating its superiority. Notably,the decision boundary in this model exhibits curvature, taking on a parabolic shape,whereas the decision boundary.