Data mining assignment1

Rajiv Devidi Reddy

2022-10-29

PART A

1.What is the main purpose of regularization when training predictive models?

When training a data model performs exceptionally well but sometimes it fails during the testing, this kind of situation is called overfitting, which is dealt with regularization. Regularization is a technique used to reduce errors by fitting a function properly on the given training set and it avoids the overfitting. It is a technique which regularizes or shrinks the coefficients towards zero, to avoid overfitting.

1. What is the role of a loss function in a predictive model? And name two common loss functions for regression models and two common loss functions for classification models.

Loss function is a method of measuring models algorithm such that how well your model algorithm suits for the featured data set, and it also measures how good your model is in terms of predicting the expected outcome. Loss function also determines the distance between expected outcomes and output of algorithm. Two common loss functions of classification model 1.Binary Cross-Entropy loss/Log loss 2.Hinge Loss Two common loss functions of regression model 1. Mean square Error / Quadratic loss / L2 Loss 2. Mean Absolute Error / L2 Loss

3.Consider the following scenario. You are building a classification model with many hyper parameters on a relatively small dataset. You will see that the training error is extremely small. Can you fully trust this model?

The provided scenario is the model with numerous hyperparameters was created using a small dataset. There is a good chance that the model is going to be overfitted. A model is only regarded to be a good prediction model if it performs well on both train and unseen data. Because the given model is prone to overfitting and has a high variance, it performs well on the training set but is likely to perform poorly on unseen data due to overfitting. Hence the model cannot be trusted because it may be overfit.

4.What is the role of the lambda parameter in regularized linear models such as Lasso or Ridge regression models?

Lambda parameter, which is also termed as Regularization rate, it controls the amount of regularization applied to the model. Increasing the lambda reduces the over fitting. Increasing the lambda shrinks the coefficient values for both lasso and regression models. Higher lambda means that it gives heavier regularization.

PART B

#Loading required packages  
  
library(ISLR)  
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

library(glmnet)

## Loading required package: Matrix

## Loaded glmnet 4.1-4

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

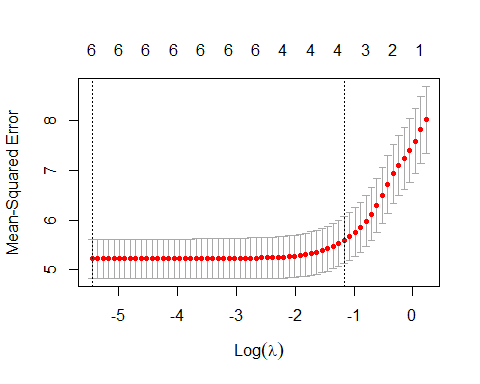
attach(Carseats)  
  
summary(Carseats)

## Sales CompPrice Income Advertising   
## Min. : 0.000 Min. : 77 Min. : 21.00 Min. : 0.000   
## 1st Qu.: 5.390 1st Qu.:115 1st Qu.: 42.75 1st Qu.: 0.000   
## Median : 7.490 Median :125 Median : 69.00 Median : 5.000   
## Mean : 7.496 Mean :125 Mean : 68.66 Mean : 6.635   
## 3rd Qu.: 9.320 3rd Qu.:135 3rd Qu.: 91.00 3rd Qu.:12.000   
## Max. :16.270 Max. :175 Max. :120.00 Max. :29.000   
## Population Price ShelveLoc Age Education   
## Min. : 10.0 Min. : 24.0 Bad : 96 Min. :25.00 Min. :10.0   
## 1st Qu.:139.0 1st Qu.:100.0 Good : 85 1st Qu.:39.75 1st Qu.:12.0   
## Median :272.0 Median :117.0 Medium:219 Median :54.50 Median :14.0   
## Mean :264.8 Mean :115.8 Mean :53.32 Mean :13.9   
## 3rd Qu.:398.5 3rd Qu.:131.0 3rd Qu.:66.00 3rd Qu.:16.0   
## Max. :509.0 Max. :191.0 Max. :80.00 Max. :18.0   
## Urban US   
## No :118 No :142   
## Yes:282 Yes:258   
##   
##   
##   
##

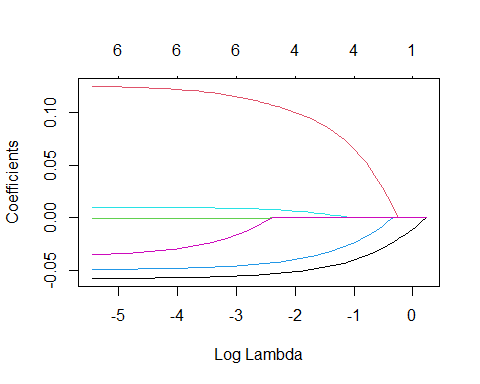
Carseats\_Filtered <- Carseats %>% select("Sales", "Price", "Advertising","Population","Age","Income","Education")   
  
x <- data.matrix(Carseats %>% select("Price", "Advertising","Population","Age","Income","Education") )  
  
y <- data.matrix (Carseats %>% select("Sales"))  
  
#Q1.Build a Lasso regression model to predict Sales based on all other attributes ("Price","Advertising", "Population", "Age", "Income" and "Education"). What is the best value of lambda for such a lasso model?   
  
  
  
library(glmnet)  
  
#performing k-fold to find optimal lambda  
cv\_fit <- cv.glmnet(x, y, alpha = 1)  
  
#find optimal lambda value that minimizes test  
  
best\_lambda <- cv\_fit$lambda.min  
best\_lambda

## [1] 0.004305309

plot(cv\_fit)



lr <- glmnet(x, y , alpha = 1 )  
plot(lr, xvar ="lambda")



#best lambda value for such lasso model is .004305309  
  
#Q2. What is the coefficient for the price (normalized) attribute in the best model (i.e. model with the optimal lambda)?   
  
#finding the coefficients of the best model  
  
b\_model = glmnet(x,y,alpha = 1, lambda = best\_lambda)  
coef(b\_model)

## 7 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) 15.8946432052  
## Price -0.0571806037  
## Advertising 0.1245169280  
## Population -0.0008863179  
## Age -0.0486751305  
## Income 0.0103378849  
## Education -0.0347353427

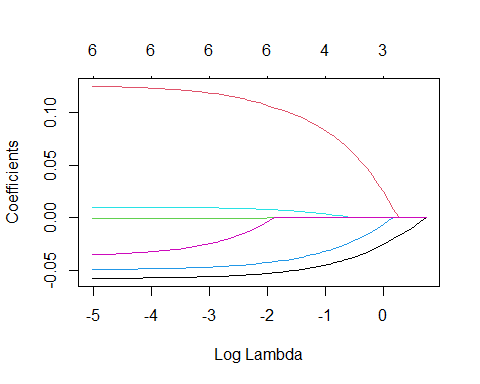
#The best coefficient for the price normalized attribute in the best model with the optimal lambda is -0.0571806037  
  
#Q3. How many attributes remain in the model if lambda is set to 0.01? How that number changes if lambda is increased to 0.1?  
  
#setting lambda at 0.01  
  
change0.01 = glmnet(x,y,alpha = 1, lambda = 0.01)  
coef(change0.01)

## 7 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) 15.8120560645  
## Price -0.0569054918  
## Advertising 0.1233407471  
## Population -0.0008269765  
## Age -0.0482649383  
## Income 0.0101795876  
## Education -0.0324465388

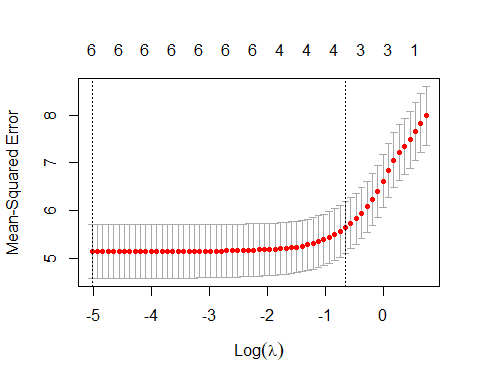
#setting lambda at 0.1  
  
change0.1 = glmnet(x,y,alpha = 1, lambda = 0.1)  
coef(change0.1)

## 7 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) 14.59030452  
## Price -0.05257390  
## Advertising 0.10536612  
## Population .   
## Age -0.04182286  
## Income 0.00764389  
## Education .

#Q4.Build an elastic-net model with alpha set to 0.6. What is the best value of lambda for such a model?   
  
  
e.netmodel = glmnet(x,y, alpha = 0.6)  
plot(e.netmodel, xvar = "lambda")



plot(cv.glmnet(x,y, alpha= 0.6))



#performing k-fold validation to find optimal lambda  
  
elasticmodel = cv.glmnet(x,y, alpha= 0.6)  
  
#finding optional lambda value that minimizes test  
  
bestlambdaela = elasticmodel$lambda.min  
bestlambdaela

## [1] 0.01253942

#the best value for such elastic-net model is .00653