**Assignment #4: Cross Validation**

**Submit through link: eCampus -> Assignments->Assignment 4 Submission**

**Deadline: October 27 (Saturday) 5:00 pm**

**The filename should have this format: LastName-FirstName-hw03.doc**

**Problem 1 (4pt)**

This question should be answered using the Default data set. In Chapter 4 on classification, we used logistic regression to predict the probability of default using income and balance. Now we will estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

1. Fit a logistic regression model that predicts default using income and balance.

Answer:

library(ISLR)

attach(Default)

glm.fit<-glm(default~income+ balance, data= Default,family = binomial)

(b) Using the validation set approach, estimate the test error of this model. You need to perform the following steps:

i. Split the sample set into a training set and a validation set.

Answer:

set.seed(1)

train = sample(dim(Default)[1], 7000)

ii. Fit a logistic regression model using only the training data set.

Answer :

glm.fit<-glm(default~income+ balance, data=Default,family =binomial, subset = train)

iii. Obtain a prediction of default status for each individual in the validation set using a threshold of 0.5.

Answer:

glm.prob<- predict(glm.fit, newdata = Default[-train,], type ="response" )

glm.pred<- rep("No",3000)

glm.pred[glm.prob > 0.5]<- "Yes"

glm.pred<-as.factor(glm.pred)

iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

Answer: table(glm.pred, default[-train])

|  |  |  |
| --- | --- | --- |
|  | Actual No | Actual yes |
| Predicted No | 2889 | 72 |
| Predicted Yes | 12 | 27 |

mean(glm.pred != default[-train])

= 0.028

(c) Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.

**Answer:** using different random seed and split with train subset we can get different training sets.

set.seed(2)

train = sample(dim(Default)[1], 6000)

set.seed(3)

train = sample(dim(Default)[1], 6500)

set.seed(4)

train = sample(dim(Default)[1], 5500)

using the below commands we will get the validation set error rates

glm.fit<-glm(default~income+ balance, data=Default,family =binomial, subset = train)

glm.prob<- predict(glm.fit, newdata = Default[-train,], type ="response" )

glm.pred<- rep("No",3000)

glm.pred[glm.prob > 0.5]<- "Yes"

glm.pred<-as.factor(glm.pred)

mean(glm.pred != default[-train])

validation error rates for the above three sets are 0.02825, 0.02314 and 0.0271 respectively

so that means we will have a different error rate for the same model depending upon the split of the data in training and validation sets.

(d) Consider another logistic regression model that predicts default using income, balance and student (qualitative). Estimate the test error for this model using the validation set approach. Does including the qualitative variable student lead to a reduction of test error rate?

**Answer:**

glm.fit<-glm(default ~ income+ balance + student , data= Default,family = binomial, subset = train)

glm.prob<- predict(glm.fit, newdata = Default[-train,], type ="response" )

glm.pred<- rep("No",3000)

glm.pred[glm.prob > 0.5]<- "Yes"

glm.pred<-as.factor(glm.pred)

mean(glm.pred != default[-train])

= 0.0276

The introduction of the predictor “student” doesn’t seem to much in the improvement of error rate.

There is no significant improvement after trying many different validation test sets.

**Problem 2 (7pt)**

This question requires performing cross validation on a simulated data set.

(a) Generate a simulated data set as follows:

set.seed(1)

x=rnorm(200)

y=x-2\*x^2+rnorm(200)

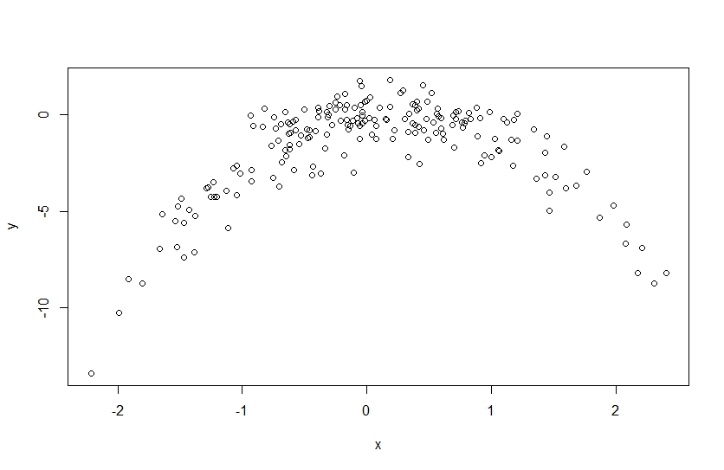
In this data set, what is and what is ? Write out the model used to generate the data in equation form (i.e., the true model of the data).

**Answer:** n=200 (as there are 200 data points)

P= 2 ( as there are two predictors x and x^2)

(b) Create a scatter plot of vs . Comment on what you find.

**Answer**: plot(x,y)



X and Y seem to be following a not linear relationship which seem to be quadratic**.**

(c) Consider the following four models for the data set:

i.

ii.

iii.

iv.

Compute the LOOCV errors that result from fitting these models.

**Answer:** data<-data.frame(x,y)

model1<- glm(y~x, data=data)

model2<- glm(y~x+ I(x^2),data=data)

model3<- glm(y~x+ I(x^2)+ I(x^3),data=data)

model4<- glm(y~x+ I(x^2)+ I(x^3) + I(x^4),data=data)

cv.glm(glmfit = model1, data=data)$delta

= 6.038

cv.glm(glmfit = model2, data=data)$delta

= 1.041

cv.glm(glmfit = model3, data=data)$delta

= 1.039

cv.glm(glmfit = model4, data=data)$delta

= 1.028

(d) Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?

**Answer:** if we use set.seed(2) we get the following cv errors

cv.glm(glmfit = model1, data=data)$delta

= 7.101

cv.glm(glmfit = model2, data=data)$delta

= 0.962

cv.glm(glmfit = model3, data=data)$delta

= 0.963

cv.glm(glmfit = model4, data=data)$delta

= 0.976

The CV errors are different if we use different random seed as the new training and test set is different from the previous one and the CV error always changes if we use a different training and validation sets because a model always have slightly different parameters when fitted on a new data set also the validation set square error tend to be different as new data points are there in it over which the error is calculated so naturally they will be slightly different for f=different data sets.

(e) Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.

**Answer:** Mostly model 4 (Biquadratic model) has lower CV error.The CV error in model 2, model 3, and model 4 are very comparable. Using different seeds sometimes result in lower CV error for one or the other model. Although model 1 which is a linear model is having a very large CV error as compared to the other 3 models.

yes it was expected since the original data is simulated such that Y is quadratic function of X so when we use quadratic model in our model we should get the best fit and hence lower CV error.

(f) Now we use 5-fold CV for the model selection. Compute the CV errors that result from fitting the four models. Which model has the smallest CV error? Are the results consistent with LOOCV?

**Answer:**

cv.glm(glmfit = model1, data=data,K = 5)$delta

=5.92

cv.glm(glmfit = model2, data=data, K=5)$delta

=1.06

cv.glm(glmfit = model3, data=data,K=5)$delta

= 1.04

cv.glm(glmfit = model4, data=data,K=5)$delta

=1.03

The 5 fold CV errors are different for each model from the LOOCV. But the trend is very similar.

CV error is quite large for the linear model and then very small and comparable for the models of higher complexity.

(g) Repeat (f) using 10-fold CV. Are the results the same as 5-fold CV?

Answer:

cv.glm(glmfit = model1, data=data,K = 10)$delta

=6.10

cv.glm(glmfit = model2, data=data, K=10)$delta

=1.05

cv.glm(glmfit = model3, data=data,K=10)$delta

=1.04

cv.glm(glmfit = model4, data=data,K=10)$delta

=1.02

The CV error for 10 fold CV error are different from 5 fold CV error but again the trend is consistently similar having large CV error for linear model and quite low for higher order models.