Assignment #4: Cross Validation

Problem 1

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.

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

(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.
ii. Fit a logistic regression model using only the training data set.
iii. Obtain a prediction of default status for each individual in the validation set using a threshold of 0.5.
iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

(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.

(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?

Problem 2

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

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

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

i. $Y = \beta_0 + \beta_1 X + \epsilon$
ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$
iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$

Compute the LOOCV errors that result from fitting these models.

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

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

(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?

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

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

Deadline: October 24, Tue @11:59pm