MATH 4322 Homework 4

Saim Ali

Spring 2023

Instructions

1. Due date: March 21, 11:59 PM
2. Answer the questions fully for full credit.
3. Scan or Type your answers and submit only one file. (If you submit several files only the recent one uploaded will be graded).
4. Preferably save your file as PDF before uploading.
5. Submit in Canvas.
6. These questions are from *An Introduction to Statistical Learning with Applications* in R by James, et. al., chapter 5.
7. The information in the gray boxes are R code that you can use to answer the questions.

Problem 1

We will review k-fold cross-validation.

(a) Explain how k-fold cross-validation is implemented.
In k-fold cross validation you must randomly divide the set of observations into k folds or groups of equal size. Of the k folds the first serves as a validation set. The method is then fit on the remaining k folds/groups. The MSE is computed by the folds other than the first one, repeating the steps k times where each time has different grouped observations where each observation then has a different validation set.

(b) What are the advantages and disadvantages of k-fold cross-validation relative to:

i. The validation set approach?
The validatoin estimate of the test error rate can vary highly, depending on the observations which are included in each set. Since only a subset of observations are used to fit the model, the validation set error can overestimate the test error rate for the model.

ii. LOOVC?
Each time we split based on 1 observation makes LOOCV very computationally expensive. One advantage is: Less bias. You keep fitting the statistical learning method using training data that contains n-1 observations, therefore, almost all the data being used LOOCV creates a less variable MSE.

Problem 2

Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X. Carefully describe how we might estimate the standard deviation of our prediction.

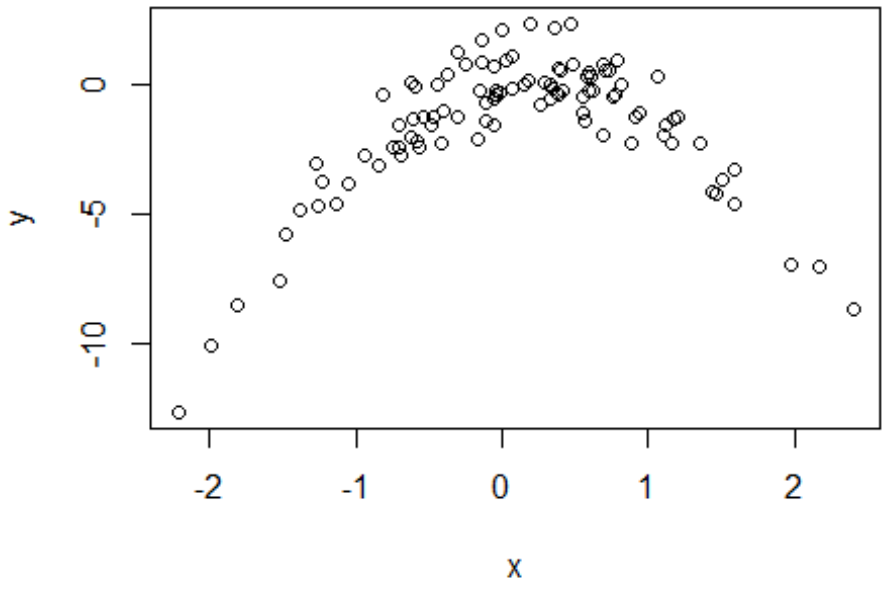
You can estimate it by using the bootstrap method discussed in class. Instead of getting new independent data sets from the population and fitting the model, we can get random samples from the original data set using the bootstrap method. In this case, we use sampling with replacement B times to then find the corresponding estimates and SD of those B estimates.

Problem 3

We will perform cross-validation on a simulated data set.

(a) Generate a simulated data set as follows:

```
set.seed(1)
x=rnorm(100)
y=x-2*x^2+ rnorm(100)
plot(x, y)
```



In this data set, what is n and what is p? Write out the model used to generate the data in equation form.

n = 100; p = 2

(b) Create a scatterplot of X against Y . Comment on what you find.
We see a curved relationship forming in the scatterplot between x and y.

(c) Set a random seed, and then compute the LOOCV errors that result from fitting the following four models using least squares:

i. $Y = \beta_0 + \beta_1 X + \epsilon$
7.288162

ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
0.9374236

iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$
0.9566218

iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$
0.9539049

Note: you might find it helpful to use the `data.frame()` function to create a single data set containing both X and Y.

(d) Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?
The results for this new random seed are the same as the results from part (c) because LOOCV looks at n folds of a single observation from x and y.

(e) Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.
Of the models in part (c) the smallest LOOCV error was in model 2. This makes sense since the plotted x and y looks quadratic lining up with this observation.

(f) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?
Our results from part (e) are futher backed up when looking at the summary as we see that the quadratic and linear variables are significants while the 4th degree and cubic variables are not significant.

Problem 4

We will use a logistic regression to predict the probability of `default` using income and `balance` on the `Default` data set in the ISLR package. We will now 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 uses income and balance to predict default.
model = glm(default ~ income + balance, data = Default, family = "binomial")

(b) Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:

i. Split the sample set into a training set and a validation set. **trainset = sample(dim(Default)[1], dim(Default)[1] / 2)**
ii. Fit a multiple logistic regression model using only the training observations.
glm(default ~ income + balance, data = Default, family = "binomial", subset = trainset)
iii. Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the default category if the posterior probability is greater than 0.5.
glm.pred=rep("N",5000), glm.pred[glm.probs>0.5] = "Y"
iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.
0.0248

(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.
0.0248, 0.0272, 0.0244; I can see from my three different splits that the set error rate can change depending on which data from the dataset is included in the training set/test set we use.

Problem 5

We continue to consider the use of a logistic regression model to predict the probability of `default` using income and `balance` on the `Default` data set. In particular, we will now compute estimates for the standard errors of the income and balance logistic regression coefficients in two different ways: (1) using the bootstrap, and (2) using the standard formula for computing the standard errors in the *glm()* function. Do not forget to set a random seed before beginning your analysis.

(a) Using the *summary()* and *glm()* functions, determine the estimated standard errors for the coefficients associated with income and balance in a multiple logistic regression model that uses both predictors.
0.4347564, 0.0000050, 0.0002274

(b) Write a function, *boot.fn()*, that takes as input the Default data set as well as an index of the observations, and that outputs the coefficient estimates for income and balance in the multiple logistic regression model.

```
boot.fn <- function(data, index)
  {fit <- glm(default ~ income + balance, data = data, family =
"binomial", subset = index)
    return (coef(fit))}
```

(c) Use the *boot()* function together with your *boot.fn()* function to estimate the standard errors of the logistic regression coefficients for income and balance.
0.4239, .0000045, .0002268

(d) Comment on the estimated standard errors obtained using the *glm()* function and using your bootstrap function.
Using the two methods I found that the standard errors seem to be pretty close to each other.

Problem 6

We will now consider the `Boston` housing data set, from the `MASS` library.

(a) Based on this data set, provide an estimate for the population mean of *medv*. Call this estimate $\hat{\mu}$.
22.53281

(b) Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result. *Hint*: We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.
0.4088611

(c) Now estimate the standard error of $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (b)?
0.4119, this is an increase from the calculated in part (b)

(d) Based on your bootstrap estimate from (c), provide a 95% confidence interval for the mean of *medv*. Compare it to the results obtained using `t.test(Boston$medv)`.
Hint: You can approximate a 95% confidence interval using the formula

$[\hat{\mu} - 2 \times \text{SE}(\hat{\mu}), \hat{\mu} + 2 \times \text{SE}(\hat{\mu})]$.
We get a 95% confidence interval: (21.70901, 23.35661) which is also close to the condfidence intervasl presented in the t-test.

(e) Based on this data set, provide an estimate, $\hat{\mu}_{\text{med}}$, for the median value of *medv* in the population.
21.2

(f) We now would like to estimate the standard error of $\hat{\mu}_{\text{med}}$. Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings.
After using the bootstrap we find that the estimate are the same from previous part (21.2). Also the boostrapped standard error for the median is .3920 (small).

(g) Based on this data set, provide an estimate for the tenth percentile of medv in Boston suburbs. Call this quantity $\hat{\mu}_{0.1}$. (You can use the quantile() function.)
10% - 12.75

(h) Use the bootstrap to estimate the standard error of $\hat{\mu}_{0.1}$. Comment on your findings.
We get the same value from part (g) (12.75), with standard error of .5099 (small) when comparing to the 10th percentile.