DSCI445 - Homework 3

Your Name

Be sure to `set.seed(445)` at the beginning of your homework.

```
#reproducibility
set.seed(445)
```

1. Explain how k-fold cross-validation is implements.
2. What are the advantages and disadvantages of k-fold cross-validation relative to
    a. The validation set approach?
    b. LOOCV?
3. In Ch. 4, we used logistic regression to predict the probability of `default` using `income` and `balance` on the `Default` data set. We will now estimate the test error of this logistic regression model using the validation set approach.
    a. Fit a logistic regression model that uses `income` and `balance` to predict `default`.
    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 training and validation sets.
        ii. Fit a multiple logistic regression model using only the training observations.
        iii. Obtain a prediction of `default` status for each individual in the validation set.
        iv. Compute the validation set error.
    c. Repeat the process in b. using 3 different splits of the observations into training and validation sets. Comment on the results obtained.
    d. Now consider logistic regression model that predicts the probability of default using `income`, `balance`, and a dummy variable for `student`. Estimate the test error for this model using the validation set approach. Comment on whether or not including `student` leads to a reduction in the test error rate.
4. The `vfold_cv()` function can be used to compute the LOOCV test error rate estimate. Alternatively, one could compute those quantitities using a `for` loop. You will take the second approach in the following problem.
    a. Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2` using the `Weekly` data set.
    b. Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2` using the `Weekly` data set *using all but the first observation.*
    c. Use the model from b. to predict the direction of the first observation. Was this observation correctly classified?
    d. Write a for loop from $i = 1$ to $i = n$ where n is the number of observations in the data set that performs each of the following steps:
        i. Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2` using the `Weekly` data set *using all but the ith observation.*

ii. Predict the direction of the ith observation.

iii. Determine whether or not an error was made in predicting the direction for the ith observation. If an error was made, indicate this as a 1 and if not, a 0.

e. Take the average of the n numbers obtained in d. iii. to obtain the LOOCV estimate for the test error. Comment on the results.

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

   a. Generate a simulated data set as follows

```
x <- rnorm(100)
y <- x - 2*x^2 + rnorm(100)

df <- data.frame(x = x, y = y)
```

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

   b. Create a scatterplot of X against Y. Commend on what you see.

   c. Compute the LOOCV errors that result from fitting the following four models using least squares:

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

   d. Repeat c. using a random seed of 200 (`set.seed(200)`) and report your results. Are your results the dame as what you got in c.? Why or why not?

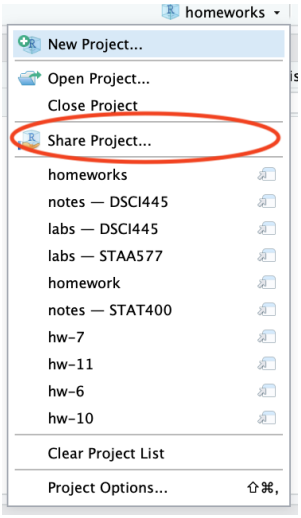
   e. Which of the models in c. had the smallest LOOCV error? Is this what you expected? Explain your answer.

   f. Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in c. using least suares. Do these results agree with your conclusions drawn based on the cross-validation results?

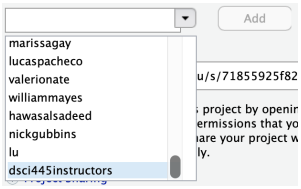
Turn in in a pdf of your homework to canvas using the provided Rmd file as a template. Your Rmd file on the server will also be used in grading, so be sure they are identical.

Be sure to share your server project with the instructor and grader. You only need to do this once per semester.

1. Open your `homeworks` project on liberator.stat.colostate.edu
2. Click the drop down on the project (top right side) > Share Project...



3. Click the drop down and add "dsci445instructors" to your project.



This is how you **receive points** for reproducibility on your homework!