DSCI445 - Homework 5

Your Name

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

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

Non-linear Models

1. We know that a cubic regression spline with one knot at ξ can be obtained using a basis of the form $x, x^2, x^3, (x-\xi)^3_+$ where $(x-\xi)^3_+ = (x-\xi)^3$ if $x > \xi$ and 0 otherwise. We will now show that a function of the form
$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)^3_+$$
is a cubic regression spline, regardless of the values of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

   a) Find a cubic polynomial
   $$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$$
   such that $f(x) = f_1(x)$ for all $x \leq \xi$. Express a_1, b_1, c_1, d_1 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

   b) Find a cubic polynomial
   $$f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3$$
   such that $f(x) = f_2(x)$ for all $x > \xi$. Express a_2, b_2, c_2, d_2 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. We have now established that $f(x)$ is a piecewise polynomial.

   c) Show that $f_1(\xi) = f_2(\xi)$. That is, $f(x)$ is continuous at ξ.

   d) Show that $f_1'(\xi) = f_2'(\xi)$. That is, $f'(x)$ is continuous at ξ.

   e) Show that $f_1''(\xi) = f_2''(\xi)$. That is, $f''(x)$ is continuous at ξ.

   Hint: Parts (d) and (e) require knowledge of single-variable calculus. As a reminder, given a cubic polynomial
   $$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$$
   the first derivative takes the form
   $$f_1'(x) = b_1 + 2c_1 x + 3d_1 x^2$$
   and the second derivative takes the form
   $$f_1''(x) = 2c_1 + 6d_1 x$$

2. GAMs are generally fit using a *back-fitting* approach. The idea behind backfitting is quite simple and we will use a linear regression example to explore it.

   Suppose we would like to perform multiple linear regression, but we do not have software to do so. Instead, we only have software to perform simple linear regression. We could take the following iterative approach:

```
 i. Hold all but one coefficient estimate fixed at it's current value.
 ii. Update only the one coefficient using simple linear regression.
 iii. Move through and update all coefficients using steps i.-ii.
```

Repeat the above approach until we have reached *convergence* – that is, until the coefficient esttimates stop changing. We will try this on a toy example.

a. Generate a response Y and two predictors X_1 and X_2 using the following model:

$$Y = 1 + 3X_1 - 4X_4 + \epsilon, \epsilon \sim N(0, 0.5)$$

where $X_1, X_2 \sim N(0, 1)$.

b. Initialize $\hat{\beta}_1$ to take value $= 10$.

c. Keeping $\hat{\beta}_1$ fixed, fit the model

$$Y - \hat{\beta}_1 X_1 = \beta_0 + \beta_2 X_2 + \epsilon$$

Set $\hat{\beta}_2 =$ the resulting coefficient from your fit.

d. Keeping $\hat{\beta}_2$ fixed, fit the model

$$Y - \hat{\beta}_2 X_2 = \beta_0 + \beta_1 X_1 + \epsilon$$

Set $\hat{\beta}_1 =$ the resulting coefficient from your fit.

e. Write a for loop to repeat c. and d. 1000 times. Make a line plot of the estimates of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, at each iteration of the for loop with $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ each in a different color.

f. Compare your answer to e. to the results of performing multiple linear regression to predict Y using X_1 and X_2 To do this, overlay a horizontal line for each coefficient value on your plot from e. (You can use `ggplot::geom_hline()` to add the horizontal lines).

g. On this data set, how many backfitting iterations were required to obtain a "good" approximation to the multiple regression coefficients.

h. Choose a different starting value for β_1 and repeat steps b.-g. Compare your results.

Tree-based Models

4. This problem involves the *OJ* data set in the *ISLR* package.

   a. Create a training set containing a random sample of 800 observations and a test set containing the remaining observations.

   b. Fit a tree to the training data with `Purchase` as the response and the other variables as predictors. Describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?

   c. Create a plot of the tree and interpret.

   d. Predict the response on the test data and produce a confusion matrix comparing the test labels to the predicted labels. What is the test error rate?

   e. Apply the cross validation to the training set in order to determine the optimal tree size.

   f. Produce a plot with complexity on the x-axis and CV error rate (or CV accuracy) on the y-axis. Which tree size corresponds to the lowest CV classification error rate?

   g. Produce a pruned tree corresponding to the optimal tree size. If CV doesn't lead to the selection of a pruned tree, then create a pruned tree with five terminal nodes.

h. Compare the training error rates between the pruned and unpruned tree.

i. Compare the test error rates between the pruned and unpruned tree.

5. We will use boosting, bagging, and random forests to predict `Salary` in the `Hitters` data set.

   a. Remove the observations for which the salary information is unknown and then log-transform the salaries.

   b. Create a training set consisting of the first 200 observations and a test set consisting of the remaining observations.

   c. Perform boosting on the training set with $1,000$ trees for a range of values of the shrinkage parameter λ. Produce a plot with different shrinkage values on the x-axis and the corresponing training MSE on the y-axis.

   d. Produce a plot with different shrinkage values on the x-axis and the corresponing test MSE on the y-axis.

   e. Compare the test MSE of boosting to the test MSE that results from two other regression approaches (Something from Ch. 3, 6, or 7)

   f. Which variables appear to be the most important predictors in the boosted model?

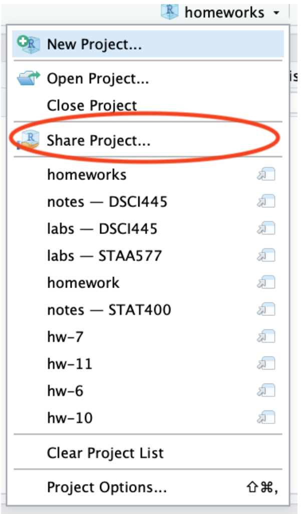
   g. Now apply bagging to the training data set. What is the test MSE for this approach?

   h. Now apply random forest to the training data set. What is the test MSE for this approach?

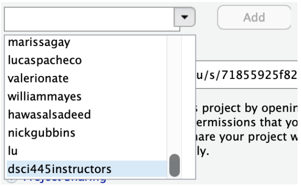
Turn in in a pdf of your homework to canvas using the provided Rmd file as a template. Your Rmd file on the server will also be used in grading, so be sure they are identical.

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.



u/s/71855925f82

; project by openin
ermissions that yo
nare your project w
ly.

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