# Problem Set #4

Kevin McAlister

February 9th, 2022

This is the fourth problem set for QTM 385 - Intro to Statistical Learning. This homework will cover a derivation related to splines and two applied exercises related to smoothing splines, GAMs, and simple regression trees.

Please use the intro to RMarkdown posted in the Intro module and my .Rmd file as a guide for writing up your answers. You can use any language you want, but I think that a number of the computational problems are easier in R. Please post any questions about the content of this problem set or RMarkdown questions to the corresponding discussion board.

Your final deliverable should be two files - a .Rmd/.ipynb file and either a rendered HTML file or a PDF. Students can complete this assignment in groups of up to 3. Please identify your collaborators at the top of your document. All students should turn in a copy of the solutions, but your solutions can be identical to those of your collaborators.

This assignment is due by February 18th, 2022 at 11:59 PM EST.

---

## Problem 1 (20 pts)

Regression splines are a broadly applicable method for regression analysis because they can be represented and estimated as an augmented OLS problem.

A generic cubic regression spline with $K$ knots can be represented as a linear model with $K + 4$ basis expansion terms:

$$h_1(x) = 1 \; ; \; h_{2 \text{ to } 4}(x) = \{x, x^2, x^3\}$$

$$h_{5 \text{ to K + 4}} = \{(x - \xi_1)^3_+, (x - \xi_2)^3_+, ..., (x - \xi_K)^3_+\}$$

$$y_i = \alpha + \sum_{k=2}^{K+4} \beta_k h_k(x_i) + \epsilon_i$$

Cubic regression splines fit a function to the data that is continuous with respect to $x$ and continuous in its first two derivatives.

For an arbitrary collection of $K \leq N$ knots, prove that the cubic regression spline provides a function that is continuous in the first and second derivative at the knots.

Notes:

1. You can assume that the function is class $C^2$ continuous away from the knots. This is true and provable, but you can just take that for granted without further explanation.
2. With an appropriate argument about the relationship between continuity in the 1st and 2nd derivatives, you don't need to show continuity on the first derivatives.

## Problem 2 (80 pts)

The data sets `college_train.csv` (600 observations) and `college_test.csv` (177 observations) include information about different colleges in the U.S. We're going to use this data set to try to build a model that predicts the logarithm of out of state tuition for a college using a variety of predictors related to college quality. Note that this data was collected back in 1995 - a magical time in U.S. history where "Run Around" by Blues Traveler was playing on the radio, Bill Clinton had a plan to actually balance the U.S. budget, and young Dr. McAlister learned to tie his shoes in Ms. Lamb's first grade class. It is also notable that college used to be affordable! Don't be too downtrodden when you look at this data set and see Emory's out of state tuition back then…

As always, the test set is intended to be used only for quantifying expected prediction error after choosing some trained models. A description of the variables in the data set can be found here. I've added an additional predictor called `AcceptRate` which is the acceptance rate of the school in 1995.

Note: I would recommend just recoding `Outstate = log(Outstate)` right at the beginning.

### Part 1 (20 pts.)

Let's start by looking at a single predictor - the student/faculty ratio `S.F.Ratio`. Plot log out of state tuition against the student/faculty ratio. Does this look linear?

Using a measure of expected prediction error appropriate for a standard linear model, find the order of global polynomial that minimizes EPE. Be sure to note your choice and why you made it. Using this value, train your model on the full training set and plot the prediction curve implied by the polynomial model on your graph.

Next, estimate a cubic regression spline, a cubic natural spline, and a smoothing spline using the entire training data. For the regression spline and natural spline, you need to choose the number of knots or degrees of freedom. I would recommend setting these to 5 to start and playing with it until you get something that looks right. For the smoothing spline, you should choose the final form using a built-in cross-validation method (most likely GCV). Add the prediction curve to your plot. How do the drawn curves differ between methods?

Finally, use your polynomial model and splines to create predictions for the test set and quantify the mean squared error for the test set. Which model performs best? Worst? Provide some rationale for this outcome.

### Part 2 (15 pts.)

Now, let's consider the multivariate case with all of the predictors. Let's improve on the standard linear model by using LASSO to do some variable selection and shrinkage. Fit the LASSO model to the training data and use K-fold cross validation to select a value of $\lambda$. Be sure to explain why you made the choice that you did. How many variables are used in the "optimal" model? Be sure to record the optimal $\lambda$ for later use.

### Part 3 (15 pts.)

Now, let's see if we can improve on the standard LASSO regression model using a GAM. There are three approaches you can take here:

1. Use all of the predictors.
2. Only use the predictors selected by LASSO under your optimal model.
3. Try to fit the GAM with an even smaller model using a subset of predictors from a higher sparsity point on the LASSO path.

Any of these approaches are fine for this problem. However, your model should include `S.F.Ratio` and `Private` (I'm pretty sure that both of these will pop up early in the LASSO path). Subset selection for GAMs is something that people are working on! There are some ways to build-in LASSO style penalties, but it's difficult to determine what a zero even is - since these are functions instead of linear combinations, what would we even shrink?

Try different combinations of linear terms, spline terms, and think-plate/tensor spline terms to try to minimize the GCV associated with your GAM. Most implementations will return this as part of the model object.

Create a plot that shows the function for each predictor. Do the marginal relationships make sense? For any 2-predictor spline terms, do they capture anything that wouldn't be captured by a linear model? Look at your function for `S.F.Ratio`. Does it look like the smoothing spline you uncovered in part 1? What does this say about the plausibility of the additivity assumption for this specific variable?

Be sure to note your "optimal" model and why you chose that one. Your search doesn't need to be exhaustive (that's impossible), just a reasonable attempt to get a good predictive model!

Compare the GCV for your chosen GAM to the K-fold CV estimate for LASSO. Does capturing nonlinearities within the data improve prediction accuracy?

### Part 4 (15 pts.)

Finally, let's use the training data to build a single regression tree. Using a CART implementation, grow a regression tree over all predictors. If you're interested, try changing some of the stopping criteria for the regression tree growth procedure and see how deep the tree goes.

Create a graphical representation of this tree. What variables are selected by the CART procedure? Does this line up with the LASSO choices?

Using cross-validation, find an "optimal" tree size with respect to the cost-complexity criterion. Use your choice to create a pruned tree and plot it. How does the pruned tree differ from the full tree?

Be sure to save the model objects for the full and pruned trees. These will be used to create predictions for the test set.

Note that we don't yet have a measure of expected prediction accuracy for trees! For now, don't worry about that.

### Part 5 (15 pts)

Finally, let's compare the three models to see which one performs best on the out of sample data. Create predictions using your LASSO model, GAM, full tree, and pruned tree for the out of sample test data. Use these predictions to compute the out of sample MSE for each method. Which performs best? Worst?

In a few sentences, discuss when you think LASSO will work better than GAMs and vice versa. Given our toolset from this week, trees need some work. And we'll do that next week!