STAT 452/652: Statistical Learning and Prediction

Owen G. Ward
owen_ward@sfu.ca

Fall 2023

These notes are largely based on previous iterations of this course, taught by Prof. Tom Loughin and Prof. Haolun Shi.

10 Generalized Additive Models

(Reading: ISLR 7.7)

Goals of this section

- Splines gave us the ability to fit smooth, flexible curves to data without specifying variable shapes
- Unfortunately, they don't work well when $p > 1$.
- We need flexible methods that CAN scale up to multiple dimensions
- Here is the first one.

10.1 Generalized Additive Model

Return to the situation where we have p explanatory variables within X

- The linear model $f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ is limited to fitting flat surfaces
  - Flat/linear in each X_j direction.
- **Instead of linear $\beta_j X_j$, use arbitrary functions of each X_j, $f_j(X_j)$:**

$$f(X) = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p)$$

  - Each variable has its own function
  - Typically fit a smoothing spline in each dimension
    - Adding up p separate univariate splines, rather than 1 multivariate spline
    - *NO interactions, unless we create them as engineered features*
  - Can leave some dimensions as linear $\beta_j X_j$ if confident of linearity (or can *specify* some other function with parameters if desired)

- Great for adaptively modeling "marginal" shapes with flexible curves.
  - "Marginal" means, e.g., looking at the 3D cube straight through one of the sides
  - NOT great if variables interact.
- Resulting surface is a little restricted due to lack of interactions
  - Overall performance is bias-variance tradeoff
  - Depends on how much of variability in $g(\mathbb{X})$ comes from
    - curvature *within* the variables
    - interactivity *between* variables
  - Creates surface that can have any shape in each X_j-direction
  - Result is a Generalized Additive Model (GAM)
    - Hastie, T. J.; Tibshirani, R. J. (1990). *Generalized Additive Models.*
- Can include additional terms in the sum to account for interactions, but must specify them
  - Two-factor $f_{jk}(X_jX_k)$
    - Single function in the direction of X_jX_k (has the same value for any combination of X_j and X_k for which $X_j * X_k = c$ for a given value c)
    - One for each j, k combination
    - Don't have to use all pairs (too many dimensions)
    - Which ones to use??? BV Tradeoff!!!
    - Rarely done
  - Higher order possible, but like with splines, not often used.
- Fitted model is at least a little bit interpretable, because we can see marginal effect of each X_j through its spline.

Fitting GAMs via backfitting

- Tricky computational problem to compute p separate *nonparametric* (spline) functions simultaneously.
  - We don't want just the best spline in each dimension
  - We want the best *combination* of splines
  - Not the same thing when explanatory variables have any correlation
- Fitting a certain function for $f_1(X_1)$ first changes how $f_2(X_2)$ needs to adapt to provide optimal prediction
  - And vice versa
  - Where do you start???
  - Normally have estimating equations that can be solved analytically or numerically to find estimates for specific parameters.

  - Here, functions fit to one dimension may change depending on those fit in other dimensions.
    * Don't even know what parameters you want to estimate!

- Backfitting algorithm
  - Cycle through dimensions repeatedly
  - *Fit each dimensions spline to residuals from other dimensions*
  - Repeat until some convergence criterion is met (functions no longer change [much])

- If all functions are linear—and you are just doing multiple linear regression—this algorithm actually returns the LSE's for each β_j!

Example: GAM on the Prostate Data (`Sec10_GAM_Prostate.R`)

The `gam()` function in several different packages will fit generalized additive models. We will use the one from `mgcv` because it has the added feature that it automatically selects the "best" spline in each dimension, so we don't have to worry too much about that. (Of course, we have seen cases where some "optimal" spline is silly, so there is risk in giving up control of the splines.) Models are fit using the formula `y ~ s(x1) + s(x2) + ...` , which fits a smoothing spline in each dimension. Indicators, variables with few levels, and other variables that have known linear trends can be included in the model without the "`s()`". There is the potential to add two-term interactions with `s(x1,x2)`.

We fit GAMs using all explanatory variables first, then refit using just `lcavol` and `pgg45` so that we can see the surface. We don't need to worry about the tests it produces to know whether they are exactly reliable, but they do at least give a sense of relative importance of the variables. You also can see how much nonlinearity is estimated within each variable by looking at the `edf` column. See the results when you run the code.

What to take away from this

- GAMs can offer a way to extend splines to multi-dimensional X
- They are useful, up to a point
  - Cross-product terms must be chosen individually
  - Otherwise, not good at modeling interactions
- Viable candidates for curved, multidimensional surfaces with little interaction.

Problem Set 12: Generalized Additive Models

Application

Refer to the Air Quality data described previously, and the analyses we have done with `Ozone` as the response variable, and the five explanatory variables (including the two engineered features).

1. Use GAM to model the relationship between `Ozone` and all five explanatories:
   (a) Print out and look at the summary from the `gam` object.

i. **Show the summary**

ii. **According to the summary, are there any variables that seem unimportant? If so which ones?**

iii. Which variables seem to have the most nonlinear influence on `Ozonw`, according to their degrees of freedom?

(b) Plot the marginal splines for each variable, making sure that the plot is large enough for you to see the patterns and the error bounds

i. **Present these plots.**

ii. For the two most nonlinear patterns identified in part (a), comment on the shape of the patterns. **Does the nonlinearity suggest a clear non-monotone relationship, or mostly just vary the rates of increasing and decreasing trends?**

2. Add GAM on all variables to the 10-fold CV comparison that has been used for LASSO, Ridge, and other methods. Use the same folds for GAM that were used for the other methods.

(a) **Report the separate MSPEs from each fold,** $MSPE_v$, $v = 1, \ldots, 10$ **and the MSPE for the full data.**

(b) **Starting with boxplots the plots made earlier for least squares, hybrid stepwise, ridge, and LASSO, ADD a boxplot of the 10 CV error estimates for GAM as the last box on the right. Comment on how GAM compares to other methods**

(c) **Repeat this using relative MSPE.**

(d) Using the knowledge gained from the analysis you did in Question 1, give a 1-sentence explanation for why GAM performs as it does. (If it is better than other methods, why? If it is no better than other methods, why?)