Generalized Additive Models

Alberto Leiro Thao Nhan
Advisor: Achraf Cohen, PhD

Summer 2022

UNIVERSITY of WEST FLORIDA

Outline

1. Introduction
2. Methods
3. Illustrative Example
4. Limitations
5. Conclusion

Introduction

- Linear regression is an extremely useful tool to perform such kinds of predictive analysis. However, in real world problems and datasets, there are times that the relationships are nonlinear. It would be insufficient to use linear regression to model the data.
- So, the question is: "How can we properly perform a regression model on a nonlinear relationship?" One option is to use a Generalized Additive Model (GAM).
- The purpose of this report is to explain how GAMs work in data science by giving its definition and methodology.

What are GAMs?

- GAMs were invented by Trevor Hastie and Robert Tibshirani in 1986.[Hastie and Tibshirani, 1986] GAMs are used when instead of a linear response variable depending on a linear function, it depends on smooth functions called splines, which are polynomial functions that cover a small range[Shafi, 2021].
- GAMs are also very easily analyzed through programming languages such as R. The main aspect of GAMs is seeing how different covariates or predictor variables can be inferred with a more fluid function that would allow for nonlinear associations[Wood, 2017].

Applications in Previous Study

- A case study done in Australia used GAMs to make water quality assessments[Richards et al., 2013]. In another field, GAMs were used to study and analyze time series of air pollution and health[Dominici, 2002].
- Study in 2017 of scholar Xi Gong, Aaron Kaulfus, Udaysankar Nair, and Daniel A. Jaffe used Generalize Additive Models to analyze the Quantifying O_3 impacts in urban area due to wildfires. GAMs were able to give helpful information about the daily average O_3 in this case study. The results of the GAMs method was directly applicable to the EPA guidance on excluding data due to an uncontrollable source [Gong et al., 2017].

Generalize Linear Models (GLMs)

- GAMs are an extension of Generalized Linear Models(GLMs), which in turn are extensions of the linear regression model [Shafi, 2021]. GLMs are shown as[Shafi, 2021]:

$$g(E(Y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{1}$$

  The GLM consists of the following parts: g is the link function, $x_1, x_2, ..., x_n$ are the predictor variables, $\beta_1, \beta_2, ..., \beta_n$ are the weight of each function, and $E(Y)$ is the probability distribution from the exponential family that defines it[Shafi, 2021].

Generalize Additive Models (GAMs)

- Instead of the variable being paired with a weight β, it is paired with a spline(smooth nonparametric functions) s, which could be linear if needed[Shafi, 2021]. GAMs do not assume the relationship to be linear. Instead, GAMs use a non-linear combination of variables. We can write the GAM structure as

$$g(E(Y)) = \beta_0 + s_0 + s_1 x_1 + s_2 x_2 + ... + s_n x_n \qquad (2)$$

where Y is the dependent variable (i.e., what we are trying to predict), $E(Y)$ denotes the expected value, and $g(Y)$ denotes the link function that links the expected value to the predictor variables x_1,. . . ,x_p. The term x_1,. . . ,x_p denote smooth, non-parametric functions [Larsen, 2015]. The spline function equation s can be determined by:

$$s(x) = \sum_{k=1}^{k} \beta_k b_k(x) \qquad (3)$$

Where k is the weight and $b_k(x)$ is a function [Shafi, 2021]

Example: Atmospheric Carbon Dioxide at the South Pole

The example that we will be using is the data set *co2s* which contains data on the amount of Carbon Dioxide concentrations per million on the South Pole over months[Wood, 2006]. The following are the variables that are being used:

- **co2**: atmospheric CO2 concentration in parts per million
- **time**: cumulative number of months since January 1957
- **month**: month of the year

We will be using *co2* as our dependent variable and the others as independent predictor variables.

Example: Atmospheric Carbon Dioxide at the South Pole

- First, we will show a Linear regression Analysis in order to showcase the difference between it and the GAMs.
- Next, we will plot the residual results from the Linear Regression Model.
- Finally, we will perform the same analysis using GAMs and compared results to the linear regression model.

R and R Packages

- Two R packages were used in order to perform the analysis in our examples. Data sets used as examples that can properly showcase and utilize GAMs were retrieved from the *gamair* package [Wood, 2006]. The GAM functions and graphs created will be derived from the *mgcv* package [Wood, 2017].

Diving into R-Studio

- Illustrative Example with R by Alberto Leiro

Summary of Results

- The linear regression model shows that both the overall time and the months of the year are significant predictors of CO2 levels at the south pole. This makes sense since the predictor variables can be defined as linear functions.
- The results are interesting since not only does it show similar results to the regular linear regression, stating that the two independent variables are also significant, but it shows a higher r-squared result of 1 with 100% of the deviance explained!

Example: Chicago Air Pollution

The example that we will be using is the data set *chicago* which contains data on the daily air pollution and death rates for *chicago*[Wood, 2006]. The following are the variables that will be used:

- *deaths*: total amount of deaths per day
- *pm10*: median particles in 2.5-10 mg per cubic meter
- *pm25*: median particles less than 2.5 mg per cubic meter (considered more dangerous)
- *o3*: ozone in parts per billion
- *so2*: Median Sulfur dioxide measurement
- *time*: time in days
- *temp*: temperature in Fahrenheit

We will be using *deaths* as the dependent variable and the others as predictor variables.

Summary of Results

- The linear regression model shows a flat line which is not helpful in our analysis and does not tell us much about our predictor variables and their relationship with the dependent variable.
- When compared to the linear regression model, GAMs has a higher r-squared value(though still a bit low) and provides more significant results for the predictor variables. It states that pm10, pm25, time, and temperature provide significant results.

Limitations

- GAMs depend on assumptions on the data generating process[Molnar, 2022]
- If the data assumptions are not met, any review of the data's weight could not be deemed valid[Molnar, 2022].
- Additionally, since GAMs are all about dealing with complex relationships by using smooth functions to fit each part of the data, sometimes it would result in over-fitting. To avoid over-fitting when using GAMs in analysis, it is best to practice comparing results between GLMs and GAMs.
- Finally, if there are any link functions identified in the GLM apart from the identity function, then it leads to complications with the GAM[Molnar, 2022].

Conclusion

- GAMs are a very useful tool used in statistics.
- The examples have shown that it can accommodate few or multiple predictor variables and swiftly provide a visual representation of the relationship between the independent predictor variables and the dependent variable.
- The benefit of the flexibility that GAMs allows for can be applied to multiple fields and can help us in understanding how predictor variables interact in our regression analysis
- Consider putting GAMs in your toolbox so it can enhance our analysis ability on any type of data which will result in a better understanding of the data and its variables.

References I

Dominici, F. (2002).
On the use of generalized additive models in time-series studies of air pollution and health.
American Journal of Epidemiology, 156(3):193–203.

Gong, X., Kaulfus, A., Nair, U., and Jaffe, D. A. (2017).
Quantifying o3 impacts in urban areas due to wildfires using a generalized additive model.
Environmental science & technology, 51(22):13216–13223.

Hastie, T. and Tibshirani, R. (1986).
Generalized additive models.
Generalized Additive Models, 1(3):297–318.

Larsen, K. (2015).
Gam: the predictive modeling silver bullet.
Multithreaded. Stitch Fix, 30:1–27.

References II

Molnar, C. (2022).
Interpretable machine learning: A guide for making Black Box models explainable.
Christoph Molnar.

Richards, R., Hughes, L., Gee, D., and Tomlinson, R. (2013).
Using generalized additive models for water quality assessments: A case study example from australia.
Journal of Coastal Research, 65:111–116.

Shafi, A. (2021).
What is a generalized additive model?
Towards Data Science.

Wood, S. (2017).
Generalized Additive Models: An Introduction with R.
Chapman and Hall/CRC, 2 edition.

References III

Wood, S. N. (2006).
Generalized Additive Models: An Introduction with R.
Chapman Hall/CRC, Boca Raton, Florida, first edition.
ISBN 1-58488-474-6.

Questions?