Swiss Institute of Bioinformatics

Generalized Additive Models

Rachel Marcone

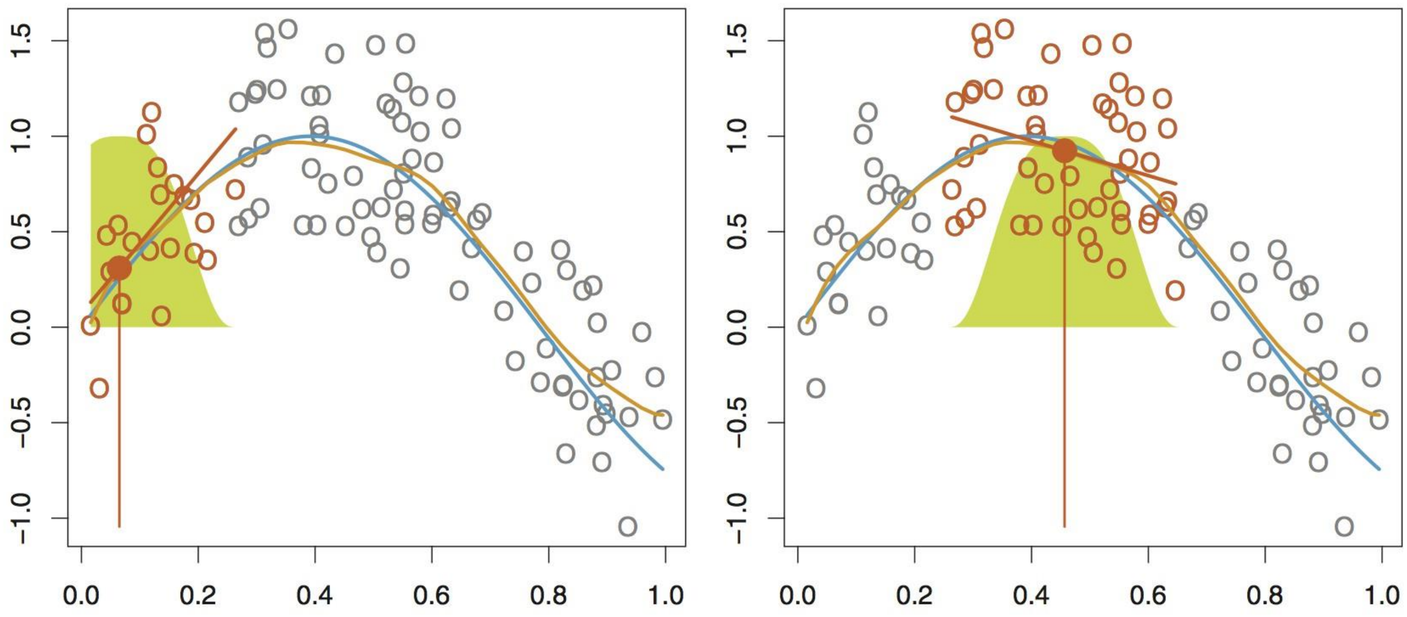
Translation Data Science group, SIB, LAUSANNE

Aug, 2022 - Lausanne

Local regression

Local Regression

- Local regression is a different approach for fitting flexible nonlinear functions, which involves computing the fit at a target point x_0 using only the nearby training observations



Local Regression

- In order to perform local regression, there are a number of choices to be made, such as:
  - how to define the weighting function
  - whether to fit a constant, linear, quadratic regression, etc.
  - The most important choice is the “span”. The span plays a role like that of the tuning parameter λ in smoothing splines: it controls the flexibility of the non-linear fit
    - The smaller the value of s, the more local and wiggly will be our fit; alternatively, a very large value of s will lead to a global fit to the data using all of the training observations
- In R via stats::loess or its ancestor stats::lowess
  - Multivariate (up to 4 predictors)
  - By default fits polynomial degree 2
  - Using tricubic weighting $w(z) = \begin{cases} (1-|z|^3)^3 & if|z| < 1 \\ 0 & otherwise \end{cases}$

Generalized Additive Models (GAMs)

- So far, we have seen a number of approaches for flexibly predicting a response Y on the basis of a single predictor X
- Here we explore the problem of predicting Y on the basis of several predictor X1, X2, …, Xp
- GAMs provide a general framework for extending a standard linear model by allowing smooth functions of each of the variables, while maintaining additivity
  - The response can be either quantitative or qualitative

Generalized Additive Models (GAMs)

- A natural way to extend the multivariable linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

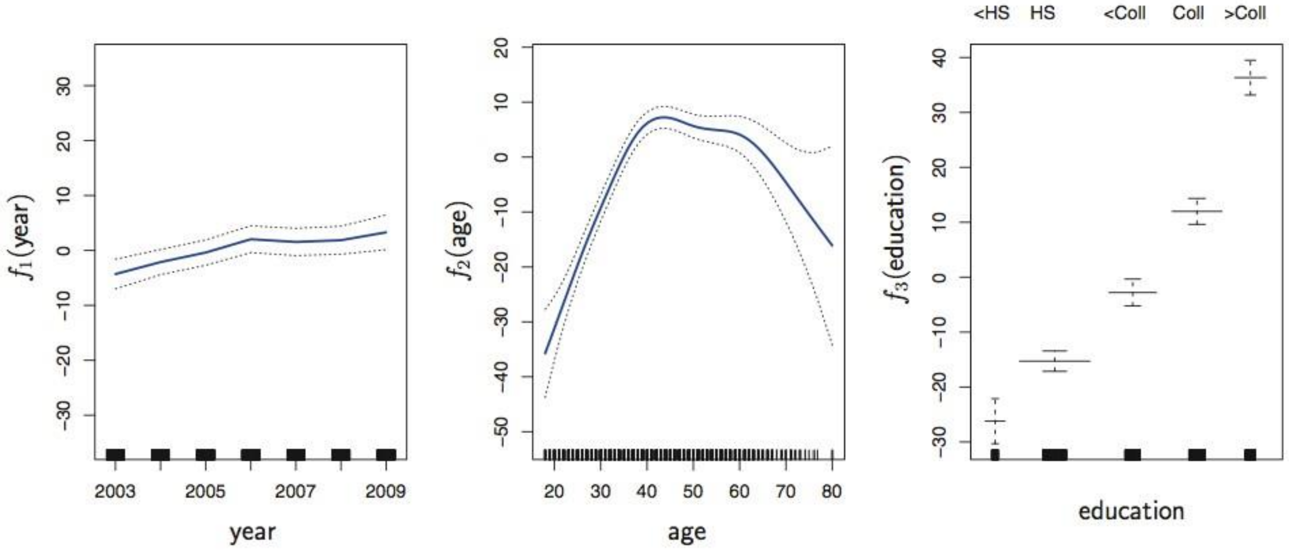
in order to allow for smooth relationships between each feature and the response is to replace each linear component with a smooth function:

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \varepsilon_i$$

- This is an example of a GAM
- It is called additive model because we calculate a separate f_j for each X_j, and then add together all of their contributions
- The beauty of GAMs is that we can use various smoothing methods as building blocks for fitting an additive model
  - Spline regression
  - Smoothing splines
  - Local regression (loess)

$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$

smoothing
splines



Fitting GAMs

$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$

- To fit a GAM using smoothing splines and local regression

$$\texttt{gam}(\texttt{wage} \sim \texttt{s}(\texttt{year}, \texttt{df} = 5) + \texttt{lo}(\texttt{age}, \texttt{span} = .5) + \texttt{education})$$

- Coefficients not that interesting; fitted functions are

- Can mix terms (linear or nonlinear) and use anova(…) to compare models

Fitting GAMs

- The gam package in R uses an approach known as ***backfitting***
  - Involves repeated updating of the fit for each predictor while holding others fixed
  - Each time we update a function, we simply apply the fitting method for that variable to a partial residual

- A partial residual for X_3 for example, has the form $r_i = y_i - f_1(x_{i1}) – f2(x_{i2})$ and therefore If we know f_1 and f_2 then we can fit f3 by treating this resudal as a response in a smooth regression on X_3

- The mgcv package in R uses mixed modeling framework for smoothing

Fitting GAMs

- GAMs allow us to fit a smooth f_j to each X_j, so that we can automatically model non-linear relationships that standard linear regression will miss
  - This means we do not need to manually try many different transformations on each variable independently
- The smooth fits can potentially make more accurate predictions for the response Y
- Because the model is additive, we can still examine the effect of each Xj on Y individually while holding all of the other variables fixed

Smoothing Exercise: The “wage” data

- Mid-Atlantic Wage Data
  - Wage and other data for a group of 3000 workers in the Mid-Atlantic region

References

- Semiparametric Regression; by David Ruppert, M.P. Wand, and R.J. Carroll; Cambridge University Press
- Generalized Additive Models; by T.J. Hastie and R.J. Tibshirani; Chapman & Hall/CRC
- Generalized Additive Models - An Introduction with R (2nd Edition); by Simon N. Wood; CRC Press