Generalized Additive Model

Hwang Charm Lee

July 3, 2017

Introduction to Generalized Additive Models

What is GAM?

- An additive model is of the form

$$E(y|x) = \alpha + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

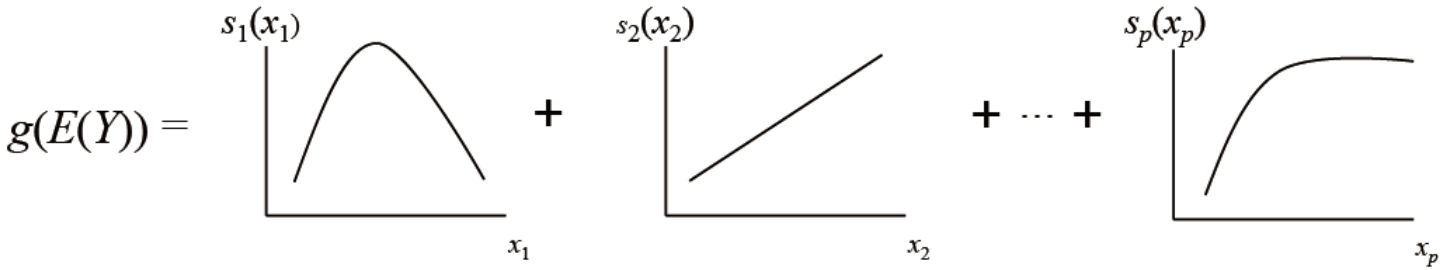
- By introducing a distribution and link function into linear regression, we have generalized linear models (GLMs)

- By introducing a distribution and link function into additive models, we have generalized additive models (GAMs)

$$g(E(y|x)) = \alpha + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

What is GAM?

- The terms $f_1(x_1), \cdots f_p(x_p)$ denote smooth, *nonparametric* fuctions, which mean that the shape of predictor functions are fully determinded by the data as opposed to *parametric* functions that are defined by a small set of parameters.



- GAMs can also contain parametric terms as well as 2-dimensional smoothers(e.g. Tensor product smooth, thin plate). Moreover, like generalized linear models(GLM), GAM supports mulitple link functions.

Why use GAM?

- Interpretability
  - The (transformed) expected value of Y increases linearly as x_2 increases, holding everything else constant. Or, the (transformed) expected value of Y increases with x_p until x_p hits a certain point,
- Flexibility and Automation
  - Predictor functions are automatically derived during model estimation.
  - We don't have to know up front what type of functions we will need. This will help us find patterns we may have missed with a parametric model.

Why use GAM?

- Regularization
  - By controlling the wiggliness of the predictor functions, we can directly tackle the bias/variance tradeoff.
- Curse of dimensionality
  - This additive structure greatly alleviates the curse of dimensionality
    - ★ From a local regression perspective, it is much easier to find points in a one-dimensional neighborhood
  - As we will see, additive models are also easy to fit computationally.

Model Estimation

Smoothers

- Smoothers are the cornerstones of GAM. At a high level, there are three classes of smoothers used for GAM
  - Local regression (LOESS)
  - Smoothing splines
  - Regression splines (B-splines, P-spline, thin plate splines)

LOESS (LOcal regrESSion)

- LOESS belongs to the class of nearest neighborhood-based smoothers.
- LOESS produces a smoother curve than the running mean by fitting a weighted regression within each nearest-neighbor window, where the weights are based on a kernel that suppresses data points far away from the target data point.

  - Determine smoothness using the span parameter.
  - Calculate $d_i = \frac{(x_i - x)}{h}$ where h is the width of the neighborhood. Create weights using the *tri-cube* function $w_i = (1 - d_i^3)^3$, if x_i is inside the neighborhood, and 0 elsewhere.
  - Fit a weighted regression with Y as the dependent variable using the weights. The fitted value at target data point x is the smoothed value.

Smoothing Splines

- We estimate the smooth function by minimizing the penalized sum of squares.

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx$$

- Tradeoff between model fit and smoothness is controlled by the smoothing parameter, λ.

- It turns out that the function that minimizes the penalized sum of squares is a natural cubic spline with knots at every data point, which is also known as a smoothing spline.

Multidimensional Splines(Thin plate splines)

- The multidimensional analog of smoothing splines are called thin plate splines.
- For d-dimensions, Thin-plate spline smoothing estimates f by finding the function $\hat{f}$ minimizing.

$$\min_f \sum_{i=1}^{N} \{y_i - f(\mathbf{x}_i)\}^2 + \lambda J_{md}(f)$$

- $J_{md}(f)$ is is a penalty functional measuring the 'wiggliness' of f
- λ is a smoothing parameter, controlling the tradeoff between data fitting and smoothness of f.

Multidimensional Splines(Thin plate splines)

- The wiggliness penalty is defined as

$$J_{md} = \int \cdots \int_{\mathcal{R}^d} \sum_{v_1+\cdots+v_d=m} \frac{m!}{v_1! \cdots v_d!} \left(\frac{\partial^m f}{\partial x_1^{v_1} \cdots \partial x_d^{v_d}} \right)^2 dx_1 \cdots dx_d$$

- In the case of a smooth of two predictors with wiggliness measured using second derivatives, we have

$$J_{22} = \int \int \left[\frac{\partial^2 f}{\partial u^2} \right]^2 + \left[\frac{\partial^2 f}{\partial v^2} \right]^2 + 2 \left[\frac{\partial^2 f}{\partial v \partial u} \right]^2 dudv$$

Multidimensional Splines(Thin plate splines)

- With J_{22} penalty, optimizing leads to a smooth 2-dimensional surface, thin-plate spline.
  - $\lambda \to 0$: the solution approaches an interpolating function.
  - $\lambda \to \infty$: the solution approaches the least squares plane.

- The solution has the form

$$\hat{f}(\mathbf{x}) = \beta_0 + \beta^t \mathbf{x} + \sum_{j=1}^{N} \alpha_j h_j(\mathbf{x})$$

  - where $h_j(\mathbf{x}) = ||\mathbf{x} - \mathbf{x_j}||^2 \log ||\mathbf{x} - \mathbf{x_j}||$ and h_j are examples of *radial basis function*.

Multidimensional Splines(Tensor product)

- Suppose $X \in \mathbb{R}^2$, and we have a basis of functions $h_{1k}(X_1), k = 1, 2, \cdots, M_1$ for representing functions of coordinate X_1, and likewise a set of M_2 functions $h_{2k}(X_2)$ for coordinate X_2. Then the $M_1 \times M_2$ dimensional **tensor product basis** defined by

$$g_{jk}(X) = h_{1j}(X_1)h_{2k}(X_2), \quad j = 1, \cdots, M_1, k = 1, \cdots, M_2$$

  can be used for representing a two-dimensional function:

$$g(X) = \sum_{j=1}^{M_2} \sum_{k=1}^{M_2} \theta_{jk} g_{jk}(X)$$

- The coefficients can be fit by least squares, as before.

Algorithms for estimating GAM

Backfitting

- The basic idea behind backfitting is to estimate each smooth component of an additive model by iteratively smoothing partial residuals from the AM, with respect to the covariates that the smooth relates to.
- The partial residuals relating to the j_{th} smooth term are the residuals resulting from subtracting all the current model term estimates from the response variable, except for the estimate of j_{th} smooth.
- Almost any smoothing method (and mixtures of methods) can be employed to estimate the smooths.

Backfitting

Algorithm

1. Initialize : $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{f}_j = 0$ for all j
2. Cycle over j until the functions $\hat{f}_j$ changes less than a pre-specified threshold.
   1. Compute partial residuals $\tilde{y}_i = y_i - \hat{\alpha} - \sum_{k \neq j} f_k(x_{ik})$ for all i
   2. Apply the 1-dimensional smoother to $\{x_{ij}, \tilde{y}_i\}_{i=1}^{n}$ to obtain $\hat{f}_j$ i.e. $\hat{f}_j \leftarrow S_j(\{\tilde{y}_i\}_{i=1}^{n})$
   3. Set $\hat{f}_j$ equal to $\hat{f}_j - \frac{1}{n}\sum_{i=1}^{n} \hat{f}_j(x_{ij})$ i.e. $\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{n}\sum_{i=1}^{n} \hat{f}_j(x_{ij})$

Backfitting(Remark)

- This same algorithm can accommodate other fitting methods in exactly the same way, by specifying appropriate smoothing operators S_j.
  - Models in which some terms are fit via local polynomials and others fit via splines
  - Models that mix parametric and nonparametric terms.
  - Models that include 2D smooth functions to model nonparametric interactions of terms.

- For identifiaility, the standard assumption is $\sum_{i=1}^{n} f_j(x_{ij}) = 0 \quad \forall j$

Backfitting(Remark)

- Computing degrees of freedom is also a simple extension of earlier results. the df for the j_{th} term are computed as $df_j = tr(S_j) - 1$.
- Backfitting is equivalent to a Gauss–Seidel algorithm for solving a certain linear system of equations.

Local scoring

- We extend additive models to generalized additive models in a similar way to the extension of linear models to generalized linear models.

- Y has conditional distribution from an exponential family and the conditional mean of the response $\mu_i = E(Y_i|X_{i1}, \cdots X_{ip}) = \mu(X_{i1}, \cdots, X_{ip})$ is related to an additive model through some link functions.

$$g(\mu_i) = \eta_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij})$$

Local scoring

- This motivates the use of the IRLS procedure used for GLMs but incorporating the backfitting algorithms used for estimation in AM.
- As seen for GLM the estimation technique is again motivated by the approximation

$$g(y_i) \approx g(\mu_i) + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$$

Local scoring

- This motivates a weighted regression setting of the form z_i, where with the ϵs, the working residuals, independent with $E(\epsilon_i) = 0$ and V_i is the variance of Y_i.

$$z_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i, \quad i = 1, \cdots, n$$

$$Var(\epsilon_i) = \omega_i^{-1} = \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2 V_i$$

Local scoring

Algorithm

1. Initialize : Find initial values for our estimate

$$\alpha^{(0)} = g\left(\sum_{i=1}^{n} \frac{y_i}{n}\right)$$

$$f_1^{(0)} = \cdots = f_p^{(0)} = 0$$

2. Update
   1. Construct an adjusted dependent variable

$$z_i = \eta_i^{(0)} + (y - \mu_i^{(0)}) \left(\frac{\partial \eta_i}{\partial \mu_i}\right)_0$$

with $\eta_i^{(0)} = \alpha^{(0)} + \sum_{j=1}^{p} f_j^{(0)}(x_{ij})$ and $\mu_i^{(0)} = g^{-1}(\eta_i^{(0)})$

Local scoring

Algorithm

1. Update(continued)
   1. Construct weights
   $$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)_0^2 (V_i^{(0)})^{-1}$$
   2. Fit an additive model to the tagets z_i with weights w_i, using a weighted backfitting algorithm. This gives new estimated functions $f_j^{(1)}$, additive predictor $\eta^{(1)}$ and fitted values $\mu_i^{(1)}$.
   3. Compute the convergence criteria.
2. Repeat previous step replacing $\eta^{(0)}$ by $\eta^{(1)}$ until the difference of $\eta^{(k)}$ and $\eta^{(k+1)}$ is less than a pre-specified threshold.

Fitting GAMs in R

- The two main packages in R that can be used to fit generalized additive models are gam(written by Trevor Hastie) and mgcv(written by Simon Wood).
- mgcv package is much more general because it considers GAM to be any penalized GLM.

Differences between gam and mgcv.

Component	gam	mgcv
Implementation	Backfitting	Lanczos algorithm
Confidence intervals	Frequentist	Bayesian
Splines	Smoothing splines, loess	Not support loess or smoothing splines, but supports a wide array of regression splines
Parametric terms	Supported	Supported, and you can penalize or treat as random effects.
Selecting smoothing parameters	No default approach	Find smoothing parameters by default. Supports both REML and GCV.
Missing values	Clever approach to dea -ling with missing value sthrough na.action=gam. -replace	Omits observations with missing values.(No special treatment)
Multi dimensional smoothers	Supported with loess	Supported with tensors and thin plate splines.

Reference

1. Simon N. Wood, Generalized Additive Models : an introduction with R, 581–683 (2006).
2. Arthur Charpentier, Computational Actuarial Science with R. Sociometry 40(1), 35–41 (2014).
3. Kim kion, The comparison of IRLS with spline basis and local algorithm in the estimation of generalized additive model, M.S. thesis at Yonsei university (2004).
4. Simon N. Wood and Giampiero Marra, Coverage properties of confidence intervals for generalized additive model components (2011).
5. Kim Larsen, GAM:The predictive modeling silver bullet, `http://http://multithreaded.stitchfix.com/assets/files/gam.pdf`
6. Hastie, Friedmann and Tibshirani, The Elements of Statistical Learning, Springer Series in Statistics, 2nd edition 295-304, (2008).