Índice general

Capítulo 1

Modelos aditivos generalizados

1.1. Introducción

Como bien podemos intuir por su nombre, los modelos aditivos generalizados no son más que la fusión entre los modelos lineales generalizados y los modelos aditivos, los cuales se introducen con una sección en este capítulo. Podemos ver estos dos tipos de modelos como extensiones del modelo lineal. Por un lado, como vimos en el capítulo anterior, el MLG hace uso de una función de enlace entre el predictor lineal y el valor esperado de la variable dependiente para poder expresar relaciones más complejas y relaja la hipótesis distribucional permitiendo que tal variable siga distribuciones de la familia exponencial. Por otro lado, los modelos aditivos, además de también relajar esta hipótesis de distribución, introducen las funciones de suavizado en el modelo, estas proporcionan más flexibilidad a la hora de relacionar las variables explicativas con la de respuesta.

Luego, como ya hemos mencionado, y como se plantea en Hastie and Tibshirani [1990], el MAG reune estas dos propuestas de modo que generaliza el modelo aditivo de la misma forma que el MLG generalizaba el modelo lineal. Sin embargo, la flexibilidad que proporciona este modelo da lugar a dos nuevos problemas teóricos: cómo estimar las funciones de suavizado y cómo de "suaves" deben ser.

En este capítulo nos adentramos en los modelos no paramétricos, es decir, en aquellos que en vez de expresar la relación del valor esperado de la variable de respuesta con las variables predictoras mediante un predictor lineal, lo hacen mediante funciones f, como se vió en **??**, pero ahora sin hacer ninguna suposición sobre ella. Esto conllevará en muchas ocasiones un mejor ajuste del modelo y traerá a la mesa una nueva cuestión conocida como sobreajuste que, aunque ya aparecía para los modelos paramétricos, ahora jugará un papel fundamental a la hora de querer predecir datos fuera de los observados. Este concepto refleja el hecho de que el modelo ajusta tan bien los datos proporcionados para la estimación de sus parámetros que es incapaz de mostrar la verdadera relación entre las variables que se estudian y, por tanto, da lugar a predicciones de nuevos datos que no serán las idóneas.

Tal y como se hace en Wood [2017], comenzaremos viendo cómo construir los modelos aditivos generalizados, es decir, qué bases de funciones podemos elegir para obtener las funciones de suavizado y qué parámetro de suavizado se debe seleccionar o cómo se puede estimar. Luego se introduce el modelo aditivo, en el que se utilizarán los resultados

vistos a lo largo del capítulo. Tras todo ello se propone la forma final del modelo aditivo generalizado.

Definición 1.1.1 (Estructura básica del modelo aditivo generalizado)**.**

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} E[Y_1] \\ \vdots \\ E[Y_n] \end{pmatrix} = E[Y]$$

$$g(\mu_i) = A_i\theta + f_1(x_{1i} + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots \quad , \forall i = 1, \dots, n \tag{1.1}$$

Donde:

- Y_i es la variable de respuesta y sigue una distribución de la familia exponencial de media μ_i y parámetro de escalado ϕ. A partir de ahora esto lo denotaremos por: $Y_i \backsim EF(\mu_i, \phi)$.
- A_i es la fila i-ésima de la matriz del modelo para aquellas componentes del modelo que son estrictamente paramétricas.
- θ es el correspondiente vector de parámetro, que antes denotábamos por β, para las variables predictoras mencionadas en el anterior punto.
- Las f_i son las funciones de suavizado para las covariables x_k.

1.2. Suavizado univariante

Dicho esto, partiremos considerando modelos que, aunque no sean adecuados para un uso práctico general, nos permitirán estudiar el marco teórico de una forma más sencilla. Es decir, en esta sección consideraremos un modelo con una sola función de suavizado, f, y una sola covariable, x, de la forma:

$$y_i = f(x_i) + \epsilon_i \tag{1.2}$$

Donde y_i es la variable de respuesta y los ϵ_i son variables aleatorias independientes e identicamente distribuidas como $N(0, \sigma^2)$ que representan el error.

1.2.1. Bases de funciones

Nos proponemos en esta sección obtener una estimación de la función de suavizado a partir de una base de un espacio de funciones, en el que también se encontrará f (o una aproximación suya). Elegir una base equivale a tomar un conjunto de funciones $\{b_j(x)\}_{j=1}^k$ y, por tanto, podemos representar la función de suavizado como:

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j \tag{1.3}$$

Para ciertos parámetros β_j a determinar.

Base polinómica

Si consideramos la base $\mathcal{B}$ del espacio de polinomios de grado k, es decir, $\mathcal{B} = \{1, x_i, x_i^2, \ldots, x_i^k\}$, la función de suavizado toma la forma:

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \cdots + \beta_{k+1} x^k$$

Y, por tanto, el modelo 1.2 queda:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \cdots + \beta_{k+1} x_i^k + \epsilon_i$$

Observación 1.2.1 (Problema de la base polinómica). Notemos que por el teorema de Taylor, la base polinomial nos será útil cuando nuestro interés sea el de estudiar las propiedades de la función de suavizado en el entorno de un punto concreto, pero nos encontramos con problemas cuando queremos hacerlo en todo el dominio de f.

El principal problema se debe a que la interpolación de los datos puede resultar en una función muy oscilante o que no ajuste bien la indormación, dependiendo del valor de k. Esto se puede solucionar de cierta manera con el siguiente tipo de base de funciones.

Base lineal por partes

Consideremos ahora una partición de nodos $\{x_j^* : j = 1, \ldots, k\}$ del rango de la variable predictora x tal que $x_j^* > x_{j+1}^*$ y la base de funciones $\mathcal{B} = \{b_j(x)\}_{j=1}^k$ donde:

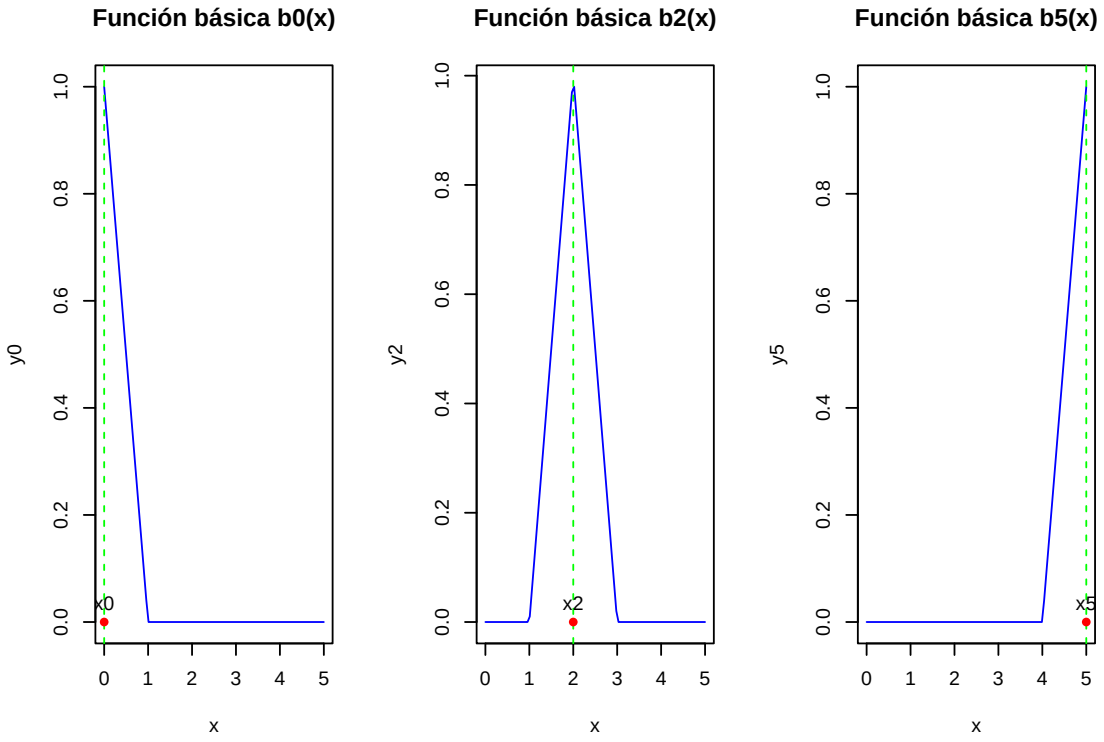
$$b_1(x) = \begin{cases} \frac{x_2^* - x}{x_2^* - x_1^*} & , si \;\; x < x_2^* \\ 0 \quad c.c. \end{cases}$$

$$b_j(x) = \begin{cases} \frac{x - x_{j-1}^*}{x_j^* - x_{j-1}^*} & , si \;\; x_{j-1}^* < x < x_j^* \\ \frac{x_{j+1}^* - x}{x_{j+1}^* - x_j^*} & , si \;\; x_j^* < x < x_{j+1}^* \\ 0 \quad c.c \end{cases}$$

$$b_k(x) = \begin{cases} \frac{x - x_{k-1}^*}{x_k^* - x_{k-1}^*} & , si \;\; x > x_{k-1}^* \\ 0 \quad c.c. \end{cases}$$

Es decir, la base de funciones $b_j(x)$ que son 0 en todo su dominio excepto entre los nodos a izquierda y derecha de x_j^*, donde crece y decrece de forma lineal hasta llegar a 1 en tal nodo. Este tipo de funciones se conocen como *tent functions*.

Ejemplo 1.2.1. Supongamos que el rango de x va de 0 a 5 y consideremos 6 nodos: $\{0, 1, 2, 3, 4, 5\}$, entonces podemos representar las funciones $b_0(x)$, $b_2(x)$ y $b_5(x)$ como:



De momento sólo planteamos estas formas de estimar las funciones de suavizado para tener una idea inicial y sencilla de cómo hacerlo pero más adelante dedicamos una sección a mejorar estas estimaciones mediante *splines*.

1.2.2. Control del suavizado

Nos interesará ahora controlar el grado de suavizado del GAM. Para ello tendremos en cuenta que el modelo aproxime de forma correcta los datos a la vez que la curvatura se mantiene controlada. Consideramos un nuevo parámetro λ, denominado parámetro de suavizado, el cuál tiene como principal función el compensar entre la fidelidad a los datos del modelos y el grado de suavizado del mismo.

Notemos primero que podemos representar la penalización a la curvatura de f como:

$$\int (f'')^2$$

Y en el caso de utilizar la base de funciones lineales por partes se puede aproximar[1] por:

$$\sum_{j=2}^{k-1} (f(x^*_{j-1}) - 2f(x^*_j) + f(x^*_{j+1}))^2$$

Es fácil observar que cuando f es una línea recta la penalización es 0 y cuando presenta muchas fluctuaciones en su curvatura este término es mayor.

[1]Se supone que se los nodos están espaciados de manera uniforme, pues en el caso de no que no lo estuvieran habría que añadir pesos a la suma.

Luego, en vez de ajustar el modelo por mínimos cuadrados, ahora se hará añadiendo la anterior penalización, es decir, minimizando:

$$||y - X\beta||^2 + \lambda \sum_{j=2}^{k-1} (f(x^*_{j-1}) - 2f(x^*_j) + f(x^*_{j+1}))^2 \tag{1.4}$$

Observación 1.2.2. Mientras mayor sea λ más importancia le estaremos dando a que la función f sea suave y menos a que a aproxime bien los datos. De hecho, cuando $\lambda \to \infty$ la función de suavizado f que minimiza la anterior expresión será una línea recta y cuando $\lambda = 0$ resultará en una estimación no penalizada.

1.2.3. Elección del parámetro de suavizado

Cómo hemos visto en la observación anterior: si el parámetro de suavizado es muy grande, el modelo será demasiado simple como para ajustarse bien a los datos y si es muy pequeño, la función de suavizado tendrá una curvatura muy alta. En cualquiera de los casos se tendrá que la estimación de f no se parecerá a la función real que ajusta los datos. Por ello, debemos dar un criterio para la elección de λ.

Un primer criterio planteado en Wood [2017] es el de elegir λ de forma que minimice la siguiente expresión para $x_1, \ldots, x_n$ unas observaciones dadas.

$$M = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_i - f_i)^2$$

Donde $\hat{f}_i = \hat{f}(x_i)$ es la evaluación de los puntos dados en la estimación de la función y $f_i = f(x_i)$ son sus evaluaciones en la función real.

Sin embargo, como la función f es desconocida, no es posible utilizar este criterio directamente. Daremos entonces una primera versión **método de validación cruzada**.

Definición 1.2.1 (Validación cruzada ordinaria)**.** Sea $\hat{f}_i^{[-i]}$ la estimación de la función de suavizado que ajustada por todos los datos $\{(x_j, y_j)\}_{j=1}^n$ menos el i-ésimo, se define la validación cruzada ordinaria como:

$$\nu_0 = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_i^{[-i]} - y_i)^2 \tag{1.5}$$

Se puede entender como que se ajusta el modelo sin utilizar la observación (x_i, y_i), se predice la variable de respuesta con este modelo en el punto x_i y luego se calcula la diferencia al cuadrado entre la estimación y el valor observado $\forall i = 1, \ldots, n$.

Observación 1.2.3. Podemos ver que tomar λ de modo que minimice ν_0 es una buena manera de abordar que minimice M. Para ello veamos que $E[\nu_0] \approx E[M] + \sigma^2$. Sustituyendo en 1.5 que $y_i = f_i + \epsilon_i$ nos queda que:

$$\nu_0 = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_i^{[-i]} - f_i + \epsilon_i)^2 = \frac{1}{n} \sum_{i=1}^{n} [(\hat{f}_i^{[-i]} - f_i)2 - 2(\hat{f}_i^{[-i]} - f_i)\epsilon_i + \epsilon_i^2]$$

Entonces, tomando valor esperado y teniendo en cuenta que $E[\epsilon_i] = 0$ y que ϵ_i y f_i son independientes:

$$E[\nu_0] = \frac{1}{n} E[\sum_{i=1}^{n} (\hat{f}_i^{[-i]} - f_i)^2] + \sigma^2 = E[M] + \sigma^2$$

Por lo tanto, cuando $n \to \infty$ se tienen las igualdades $E[\nu_0] = E[M] + \sigma^2$ y $\hat{f}^{[-i]} = \hat{f}$.

Si los modelos sólo fueran juzgados por su capacidad de ajustar los datos que les aportamos, entonces siempre se elegirían los modelos más complejos, pero el elegir el modelo que maximice la capacidad de predecir nuevos datos no tiene este problema.

Sin embargo, como se indica en Wood [2017], p.171, este método es costoso computacionalmente, ya que se deben realizar n ajustes de los datos, por ello se propone un nuevo método el cuál hace uso de la matriz de influencia A.

Definición 1.2.2 (Validación cruzada generalizada)**.** Dadas unas observaciones $\{(x_i, y_i)\}_{i=1}^n$ se elige λ tal que minimice:

$$\nu_g = n\frac{\sum_{i=1}^n (y_i - \hat{f}_i)^2}{(n - tr(A))^2}$$

1.3. Modelos aditivos

Como ya hemos mencionado previamente, el modelo aditivo es una extensión del modelo de regresión lineal. Su principal característica, la cual da lugar a su nombre, es que los efectos de las variables predictoras sobre la variable de respuesta son aditivos, es decir, una vez ajustado el modelo aditivo se pueden examinar tales efectors por separado. Veremos primero la forma general del modelo aditivo tal y como la introduce Hastie and Tibshirani [1990] y luego desarrollaremos la teoría alrededor de él para el caso de dos variables predictoras.

Definición 1.3.1 (Modelo aditivo)**.** Supongamos el contexto de las anteriores de las definiciones de modelos, el modelo aditivo se expresa como:

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon$$

Donde α es el término independiente, ϵ son errores aleatorios independientes de los X_j tales que $E[\epsilon] = 0$ y $Var(\epsilon) = \sigma^2$, y las f_j son funciones que conviene suponer univariadas y suaves pero no es necesario.

Observación 1.3.1. Además se debe tener que $E[f_j(X_j)] = 0 \quad \forall j = 1, \ldots, n$, pues de otro modo las funciones f_j añadirían términos independientes constantes adicionales.

Suele ser útil el pensar el modelo aditivo como un método que primero estima los parámetros adecuados en los que medir las variables y luego realiza el análisis lineal estándar sobre las variables transformadas. La principal motivación a priori tras este tipo de modelos es que, al representar por separado el efecto de cada variable predictora, mantienen la interpretabilidad del modelo lineal.

En lo que sigue, para poder ajustar más fácilmente el modelo como en Wood [2017], supondremos que se tienen sólo dos variables predictoras $X = (X_1, \ldots, X_p)$ y $V = (V_1, \ldots, V_p)$ y consideraremos el modelo aditivo:

$$y_i = \alpha + f_1(x_i) + f_2(v_i) + \epsilon_i \tag{1.6}$$

Observación 1.3.2. Los principales problemas del modelo aditivo son:

- La suposición de los efectos aditivos sobre **??** es bastante restrictiva.
- Existen problemas de identificabilidad pues las f_j son estimables con una precisión de una constante aditiva.

Sin embargo, si suponemos resueltos estos problemas, el modelo aditivo puede ser representado por splines de regresión penalizados, los cuales serán estimados mediante mínimos cuadrados penalizados, y el grado de suavizado, que se obtendrá por validación cruzada.

Modelo aditivo por regresión penalizada por partes

En lo que sigue, consideraremos la base del espacio de funciones lineales por partes vista en la sección anterior, es decir, expresamos las funciones f_1 y f_2 como:

$$f_1(x) = \sum_{j=1}^{k_1} b_j(x)\delta_j f_2(v) = \sum_{j=1}^{k_2} \beta_j(v)\gamma_j$$

Donde δ_j y γ_j son parámetros conocidos y las b_j y β_j son las funciones básicas de tipo carpa para los nodos x_j^* y v_j^* respectivamente, los cuales están espaciados uniformemente en el rango de x y v.

Definimos ahora los vectores n-dimesionales $\vec{f_1} = (f_1(x_1), \ldots, f_1(x_n))^T$ y $\vec{f_2} = (f_2(v_1), \ldots, f_2(v_n))^T$ como:

$$\vec{f_1} = X_1\delta = \begin{pmatrix} b_1(x_1) & \ldots & b_{k1}(x_1) \\ \vdots & \ldots & \vdots \\ b_1(x_n) & \ldots & b_{k1}(x_n) \end{pmatrix} \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_{k1} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{k1} b_j(x_1)\delta_j \\ \vdots \\ \sum_{j=1}^{k1} b_j(x_n)\delta_j \end{pmatrix} = \begin{pmatrix} f_1(x_1) \\ \vdots \\ f_1(x_n) \end{pmatrix}$$

$$\vec{f_2} = X_2\gamma = \begin{pmatrix} \beta_1(v_1) & \ldots & \beta_{k2}(v_1) \\ \vdots & \ldots & \vdots \\ \beta_1(v_n) & \ldots & \beta_{k2}(v_n) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_{k1} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{k2} \beta_j(v_1)\gamma_j \\ \vdots \\ \sum_{j=1}^{k2} \beta_j(v_n)\gamma_j \end{pmatrix} = \begin{pmatrix} f_2(v_1) \\ \vdots \\ f_2(v_n) \end{pmatrix}$$

Proposición 1.3.1. *En el caso de considerar la base lineal por partes, los coeficientes β_j que definen a una función f coinciden con los valores de la función en los nodos, es decir, $\beta_j = f(x_j^*)$.*

Gracias a esto, se tiene que el problema de ajuste de la regresión penalizada se reduce a minimizar la siguiente expresión respecto de *beta*:

$$||y - X\beta||^2 + \lambda\beta^T S\beta$$

Donde $S = D^T D$ con $\begin{pmatrix} 1 & -2 & 1 & 0 & \ldots & \ldots \\ 0 & 1 & -2 & 1 & 0 & \ldots \\ 0 & 0 & 1 & -2 & 1 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ddots \end{pmatrix}$

Bibliografía

T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1990. ISBN 9780412343902. URL https://books.google.it/books?id=qa29r1Ze1coC.

S.N. Wood. *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2017. ISBN 9781498728348. URL https://books.google.it/books?id=HL-PDwAAQBAJ.