Sesión 7: Modelos GAM y Regularización

Modelos Aditivos Generalizados y Técnicas de Regularización

Cristian E García

cegarcia@uao.edu.co

Maestría en Inteligencia Artificial y Ciencias de Datos
Facultad de Ingeniería
2025

Universidad Autónoma de Occidente

Outline

1. Modelos GAM
   1.1. Introducción
   1.2. Definición
   1.3. Modelos

2. Regularización

3. Ejemplos

Introducción

En el estudio de las variables bidimensionales, y en general, de las multidimensionales puede resultar interesante investigar la posible existencia de una relación de dependencia entre las variables implicadas y la construcción de algún modelo matemático que permita describir dicha relación, en el supuesto de que ésta exista

Introducción

En el estudio de las variables bidimensionales, y en general, de las multidimensionales puede resultar interesante investigar la posible existencia de una relación de dependencia entre las variables implicadas y la construcción de algún modelo matemático que permita describir dicha relación, en el supuesto de que ésta exista

El propósito del estudio de los modelos de regresión es construir modelos matemáticos que permitan explicar la relación de dependencia existente entre una variable respuesta Y y una o más variables independientes. Podemos utilizar estos modelos como herramienta para predecir nuevos valores de la variable respuesta a partir de cierto valor particular que ha tomado la variable explicativa. Es imprescindible el empleo de técnicas de regresión no paramétrica cuando se pretende predecir una variable respuesta que es imposible o muy costosa de medir

uao

Introducción

Dichas relaciones entre una variable de respuesta y sus predcitores líneales se han abordado desde diferentes perpectivas y en el caso de que la relación no sea propiamente lineal (en el sentido de la recta), se han planteado diversas variantes del modelo convencional, en este trabajo se aborda los modelos aditivos generalizados (GAM) y sus variantes en cuanto a las suavizadores y las adaptaciones a diversas aplicaciones.

Modelo de Regresión no Paramétrico

Los objetivos del análisis de regresión no paramétrica son los mismos que su contraparte parámetrica, vale decir, estimar y probar las características de la función de regresión. Se dispone de ciertas variables explicativas x_i, de una variable respuesta y_i y un termino aleatorio asociado al error ε_i.

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1,...,n$$

Modelo de Regresión no Paramétrico

Los objetivos del análisis de regresión no paramétrica son los mismos que su contraparte parámetrica, vale decir, estimar y probar las características de la función de regresión. Se dispone de ciertas variables explicativas x_i, de una variable respuesta y_i y un termino aleatorio asociado al error ε_i.

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1,...,n$$

El procedimiento para estimar la función de regresión f en el marco del análisis de regresión no-paramétrica se llama *suavización*.

Teniendo en cuenta lo anterior el modelo aditivo generalizado se define como:

$$(y \mid x = (x_1, x_2, \cdots, x_p)) = \beta_0 + f_1(x_1) + \cdots + f_p(x_p) + \varepsilon$$

uao

Modelo de Regresión no Paramétrico

La función de regresión se presenta como:

$$\sum_{j=1}^{p} f_j(x_j)$$

En donde se define a $f_j(x_j)$ como las funciones no especificadas suaves (no paramétricas en el sentido de que se estimarán mediante procedimientos no paramétricos). Un modelo aditivo generalizado difiere de un modelo lineal generalizado en que un predictor aditivo sustituye al predictor lineal. Específicamente, se supone que la respuesta y tiene una distribución dada con la media de $\mu = E(y \mid \mathbf{x} = (x_1, x_2, \cdots, x_p)$vinculado a través de predictores.

uao

Algoritmo Backfitting

Partiendo del modelo

$$y = \beta_0 + \sum_{i=1}^{p} f_i(x_i) + \epsilon$$

Se tiene que si este es cierto, entonces una forma de obtener f_i es a través de

$$f_i(x_i) = E\left[y - \beta_0 - \sum_{k \neq i} f_k(x_k) \,|\, x_i \right]$$

En cuyo caso se debe plantear un proceso iterativo para obtener una estimación de las funciones f_i

uao

Algoritmo Backfitting

[P1:] Iniciar con $\hat{\beta}_0 = \bar{y}, \hat{f}_j \equiv 0 \forall i, j$

[P2:] Plantear ciclos $j = 1,2,\cdots,p,\cdots,1,2,\cdots,p,\cdots,$

$$\hat{f}_j \longleftarrow S_j \left[\left\{ y - \hat{\beta}_0 - \sum_{k \neq i} \hat{f}_k(x_{ik}) \right\}_1^n \right]$$

$$\hat{f}_j \longleftarrow \hat{f}_j - \frac{1}{n} \sum_{i=1}^{n} \hat{f}_j(x_{ij})$$

Repetir hasta que los $\hat{f}_j$ no cambién con la diferencia entre ellos sea menor a un umbral previamente establecido.

uao

Versión General Backfitting

Sin perdida de la generalidad plantearemos f de la sigueinte manera

$$\hat{f}_j = S_j(Y - \sum_{l \neq j} \hat{f}_l - X\hat{\beta})$$

De esta manera si asumimos que se cuenta con una estimación de β se puede ver

$$\hat{f}_j = S_j(Y - \sum\nolimits_{l \neq j} \hat{f}_l - X\hat{\beta})$$

como un vector residual.

De lo anterior podemos plantear que todos los estimadores que se basan en el enfoque de bases de funciones, es decir Smoothing tienen la forma,

$$\hat{f}_j = V_j(V_j^T V_j + \lambda_j K_j)^{-1} V_j^T (Y - \sum_{l \neq j} \hat{f}_l - X\hat{\beta})$$

Regularización

¿De que se trata?

Pensemos en algunas situaciones que se dan en la practica, ¿qué pasa cuando el número de variables es mayor que en número de observaciones ?

Por otro lado preguntemonos por un problema que se escucha comúnmente, la llamada “Multicolinialidad”

Regresión Lineal Regularizada

Regularización (un concepto muy importante)

- Añadir una penalización en valores grandes de $\beta_1, \ldots, \beta_p$
- Añadir penalización a la función objetivo
- Resolver para $\hat{\beta}$!

La nueva función objetivo:

$$\hat{\beta} = \text{argmin}_{\beta} \left[\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \text{penalty}(\beta_j) \right]$$

λ es el peso de la penalización: los valores bajos significan pocos coeficientes cerca de 0, los valores altos significan muchos coeficientes cerca de 0.

uao

Regresión Lineal Regularizada

La nueva función objetivo:

$$\hat{\beta} = \text{argmin}_{\beta} \left[\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \text{penalty}(\beta_j) \right]$$

¿Cuándo y por qué puede ser un mejor predictor?

- Si bien agrega sesgo (no estamos brindando lo mejor a los datos)
- reduce en gran medida la varianza

Nota 1: ¡necesitamos reescalar tanto X como Y! ¿Por qué?

Nota 2: ¿será este siempre un mejor predictor? ¿Por qué no?

uao

Regresión Rige

La regresión Ridge es una técnica de regularización estadística clásica que corresponde al enfoque de regularización de Tikhonov, y se logra añadiendo la norma-L_2 (al cuadrado) del vector de parámetros a un criterio de optimalidad. En el contexto de la regresión lineal, esto se traduce en agregar una penalización proporcional a la suma de los cuadrados de los coeficientes:

Regresión Rige

La regresión Ridge es una técnica de regularización estadística clásica que corresponde al enfoque de regularización de Tikhonov, y se logra añadiendo la norma-L_2 (al cuadrado) del vector de parámetros a un criterio de optimalidad. En el contexto de la regresión lineal, esto se traduce en agregar una penalización proporcional a la suma de los cuadrados de los coeficientes:

$$\text{pen}(\boldsymbol{\beta}) = \sum_{j=0}^{k} \beta_j^2 = \boldsymbol{\beta}'\boldsymbol{\beta}$$

El criterio penalizado de mínimos cuadrados se expresa como:

$$\text{PLS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}$$

Regresión Rige

Siguiendo una derivación análoga a la del estimador de mínimos cuadrados ordinarios, se toma la derivada con respecto a $\boldsymbol{\beta}$:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}}\text{PLS}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}}\left(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}\right) \\ &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta}\end{aligned}$$

Regresión Rige

El estimador ridge:

$$\hat{\boldsymbol{\beta}}_{\text{PLS}} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}$$

Este estimador se diferencia del estimador de mínimos cuadrados ordinarios por el término adicional $\lambda \mathbf{I}_p$, que actúa como una penalización. Si λ es cercano a cero, el estimador ridge se aproxima al estimador clásico. Sin embargo, si λ es grande, el término $\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p$ es siempre invertible (incluso si $\mathbf{X}'\mathbf{X}$ no lo es), y los coeficientes estimados tienden a reducirse hacia cero.

$$\text{pen}(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{\beta} \Rightarrow \boldsymbol{\beta} \to \mathbf{0}$$

Regresión LASSO (Least Absolute Shrinkage and Selection Operator)

Mientras que la regresión Ridge permite la estimación de coeficientes en contextos de alta dimensionalidad o con matrices de diseño cercanas a la colinealidad, esta no produce una solución dispersa: todos los coeficientes estimados serán distintos de cero (con probabilidad uno).

Desde el punto de vista interpretativo, sería deseable no solo reducir los coeficientes pequeños hacia cero, sino también tener la posibilidad de establecer algunos exactamente en cero. Esto permitiría combinar la estimación del modelo con la selección de variables en un solo paso.

uao

Regresión LASSO (Least Absolute Shrinkage and Selection Operator)

Este enfoque se puede lograr reemplazando la penalización de mínimos cuadrados (como en Ridge) por una penalización basada en los valores absolutos:

$$\text{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{k} |\beta_j|$$

Así, el estimador LASSO se define como:

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{k} |\beta_j| \right)$$

Regresión LASSO (Least Absolute Shrinkage and Selection Operator)

donde, nuevamente, el intercepto no se penaliza. Al igual que en Ridge, el criterio penalizado de mínimos cuadrados busca un balance entre el ajuste del modelo a los datos y la regularización inducida por la penalización. Este balance está gobernado por el parámetro de suavizado λ.

Dado que $\hat{\boldsymbol{\beta}}_{\text{LASSO}}$ se define en términos de una penalización L_1, permite seleccionar covariables de forma automática (tipo selección de variables), por lo que se le conoce como el **operador de contracción y selección absoluta mínima**, o **LASSO**

uao

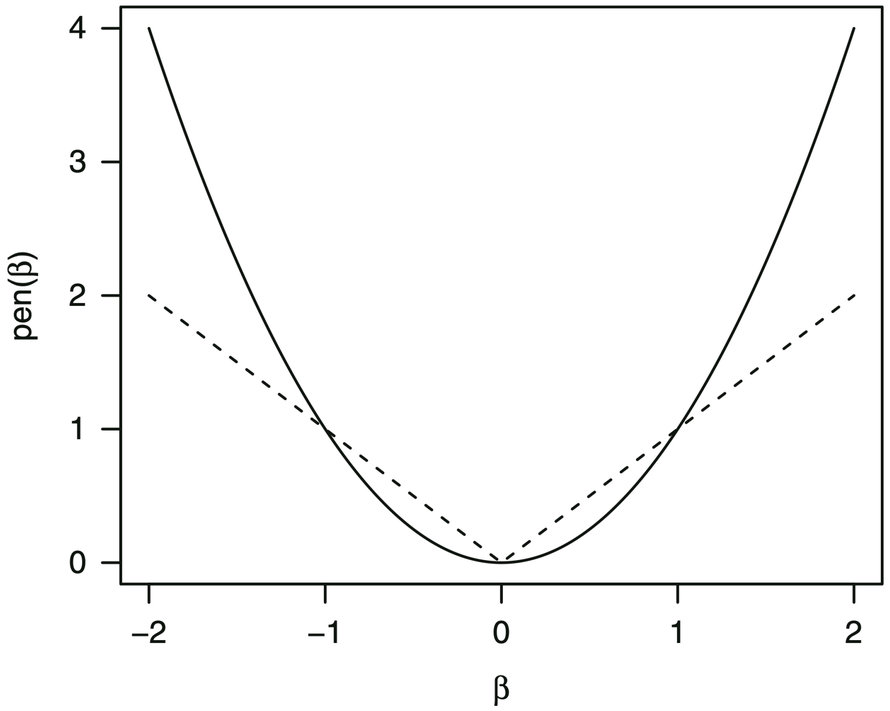
Regresión LASSO (Least Absolute Shrinkage and Selection Operator)

La diferencia clave entre Ridge y LASSO radica en la forma de la penalización. Ridge impone una penalización cuadrática, que tiene un fuerte impacto sobre coeficientes grandes, pero penaliza débilmente los coeficientes cercanos a cero.

Por el contrario, la penalización L_1 de LASSO crece más lentamente para coeficientes grandes, pero tiene un mayor efecto en aquellos cercanos a cero. Como resultado, se espera que LASSO reduzca a cero coeficientes pequeños (produciendo modelos dispersos), mientras que los coeficientes grandes se ven menos afectados.

uao

Regresión LASSO (Least Absolute Shrinkage and Selection Operator)



Penalizaciones para regresión Ridge (línea discontinua) y LASSO (línea continua)}. La penalización de Ridge corresponde a una función cuadrática ($\lambda\beta^2$), mientras que la penalización LASSO usa el valor absoluto $\lambda|\beta|$. Obsérvese cómo la penalización LASSO crece más lentamente para valores grandes de β, pero es más pronunciada cerca de cero, lo cual permite que algunos coeficientes se reduzcan exactamente a cero.

Estimación Regularizada

Criterio de Mínimos Cuadrados Penalizados

La estimación regularizada en modelos lineales se basa en el criterio de mínimos cuadrados penalizados:

$$\text{PLS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \cdot \text{pen}(\boldsymbol{\beta})$$

donde $\lambda \geq 0$ es un parámetro de suavizado, y $\text{pen}(\boldsymbol{\beta})$ penaliza la complejidad del modelo.

Estimación Regularizada

Regresión Ridge

En la regresión Ridge, la penalización se da como la suma de los cuadrados de los coeficientes:

$$\text{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{k} \beta_j^2 = \boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}$$

donde $\mathbf{K} = \text{diag}(0,1,\ldots,1)$ es la matriz de penalización que excluye el intercepto. El estimador de mínimos cuadrados penalizado es:

$$\hat{\boldsymbol{\beta}}_{\text{PLS}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{K})^{-1}\mathbf{X}'\mathbf{y}$$

Estimación Regularizada

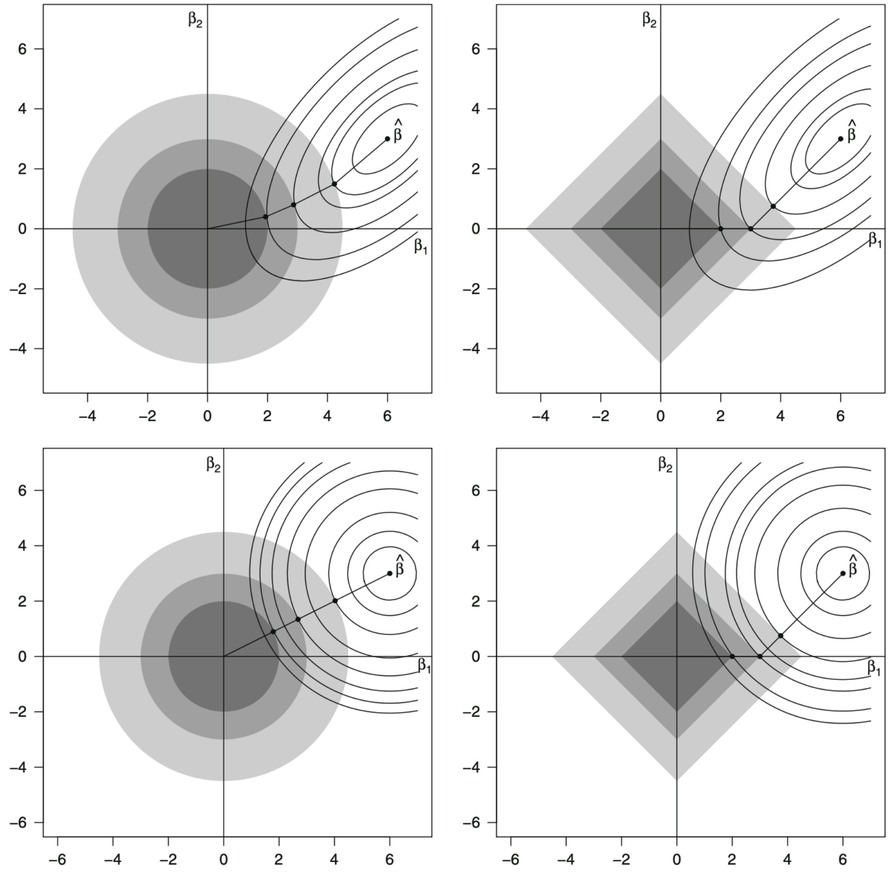
LASSO

Para el método LASSO, la penalización corresponde a la suma de los valores absolutos de los coeficientes:

$$\text{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{k} |\beta_j|$$

Este estimador no tiene una solución en forma cerrada y debe calcularse numéricamente, por ejemplo mediante programación cuadrática.

Estimación Regularizada



En Ridge (izquierda), la penalización genera regiones de contorno circulares(ℓ_2-norma). En LASSO (derecha), la penalización genera regiones con forma de rombo o diamante (ℓ_1-norma).

La solución penalizada $\hat{\boldsymbol{\beta}}$ ocurre donde la línea de nivel del error toca la región factible. En el caso de LASSO, esto suele ocurrir en los vértices, generando coeficientes exactamente cero y favoreciendo la **selección de variables**.

Aplicaciones

¡Veámoslo en R!

uao

¿Preguntas?