MÉTODO BAYESIANO EMPÍRICO PARA AJUSTE Y PREDICCIÓN DEL MODELO DE REGRESIÓN SEMIPARAMÉTRICO

Jaime Cuevas Domínguez[1], José Crossa Hiriart[2], Walter MagañaLandero[1]; Víctor Soberanis Cruz[1]

[1]Universidad de Quintana Roo, Unidad Chetumal, Departamento de Ciencias, Chetumal Quintana Roo México

[2]Centro Internacional de Mejoramiento de Maíz y Trigo. Departamento de Bioestadística, Texcoco Estado de México. México

Email: jaicueva@uqroo.edu.mx

RESUMEN

Se propone un método veloz y preciso para la predicción en un modelo de regresión kernel gaussiano. El método estima el ancho de banda del kernel gaussiano y el parámetro de encogimiento del modelo, maximizando la verosimilitud integrada previa integración de los parámetros de ruido. Se muestra su desempeño con una función de regresión no lineal simulada añadiéndole un nivel de ruido. Asimismo modo se anexan los seudocódigos.

Palabras clave: Regresión semiparamétrica, Bayesiano empírico, Kernel

INTRODUCCION

Para modelar una variable respuesta $(\boldsymbol{y})$, en función de una o más variables explicativas $(\boldsymbol{x})$ adicionando una variable aleatoria conocida como error $(\boldsymbol{e})$ que acumula lo que no puede explicar la función de las variables $\boldsymbol{x}$, se usa la siguiente función de regresión:

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{e} \qquad (1)$$

Donde la función $f(\boldsymbol{x})$, establece una relación entre las variables o explicativas $(\boldsymbol{x})$ y la variable respuesta $(\boldsymbol{y})$.

En la regresión paramétrica $f(\boldsymbol{x})$ es conocida y sigue una estructura rígida. En la regresión semiparamétrica, la idea básica es desarrollar un modelo con menos supuestos para predecir la respuesta en un rango de valores, captando las complejidades en el comportamiento de los datos con la idea de mejorar la predicción. Las técnicas semiparamétricas proporcionan una estimación suavizada de la relación para un conjunto de valores (ancho de banda) de la(s) variables explicativa(s). Estos valores son ponderados, de modo que los vecinos más cercanos tengan mayor peso que los más alejados en una ventana de datos. Diversas funciones (kernel) basados en una medida de distancia son utilizados. La función kernel y el ancho de banda (h) determinan

Manuscrito recibido el 05 de Septiembre de 2016

el ajuste del modelo con los datos, en este sentido varios autores muestran que la determinación del ancho de banda suele ser más importante que el kernel utilizado (Hardle 1990).

La decisión en torno al ancho de banda, consiste en seleccionar una tal que no sea pequeña o muy local y que sobreajuste el modelo, ni tampoco una ventana muy ancha o demasiado global, que impida captar las complejidades del modelo derivando en un ajuste bajo.

Una forma de entender la problemática de la regresión kernel es por medio de la predicción. Sea $\hat{y}_i$ el estimado para la $i-esima$ observación y_i $i = 1, \dots, n$. Entonces para una kernel determinado $f(\boldsymbol{x})$:

$$\hat{y}_i = \sum_{j=1}^{n} w_{ij} y_j$$

donde $\sum_{j=1}^{n} w_{ij} = 1$.O en términos matriciales

$$\hat{\boldsymbol{y}} = \boldsymbol{W}\boldsymbol{y} \quad (2)$$

Donde generalmente los pesos ponderados se escogen de tal modo que los w_{ij} sean cercanos a cero para todas las y_i, fuera de la proximidad del lugar de interés. Así mismo la función kernel es utilizada para definir los pesos o ponderaciones $\boldsymbol{W}$. Uno de los primeros y más comunes estimadores de los pesos fue propuesto por Nadaraya (1964) y Watson (1964).

$$w_{ij} = \frac{K(\frac{x_i - x_j}{h})}{\sum_{j=1}^{n} K(\frac{x_i - x_j}{h})} \quad (3)$$

Donde el parámetro h es el ancho de banda que define el ajuste del modelo y su selección es uno de los motivos de este artículo.

La primera idea en regresión para seleccionar h, es minimizar el cuadrado medio del error de predicción (PMSE), Allen (1974). Este método también conocido como validación cruzada es muy utilizado, sin embargo Gianola et al (2006) comenta que este método suele ser bastante demandante e incluso poco factible cuando se trata de una número muy grande de observaciones.

En este trabajo se usa un método basado en la verosimilitud integrada similar al propuesto por Pérez-Elizalde y Cols (2015) y se aplica a un ejemplo simulado para visualizar el desempeño del método.

EL MODELO DE REGRESIÓN KERNEL SEMIPARAMÉTRICO

El modelo (1), puede ser definido como:

$$\boldsymbol{y} = \mu\boldsymbol{1} + \boldsymbol{K}\boldsymbol{f} + \boldsymbol{e} \quad (4)$$

Donde se asume que el vector $\boldsymbol{y}$ sigue una distribución normal multivariada, μ es un intercepto o media común, $\boldsymbol{1}$ es un vector unos, $\boldsymbol{K}$ es una matriz semidefinida positiva construidas con las covariables $\boldsymbol{x}$, $\boldsymbol{f}$ y $\boldsymbol{e}$ son vectores aleatorios independientes con distribución normal con media cero y varianza $\sigma_f^2\boldsymbol{I}$,$\sigma^2\boldsymbol{I}$, respectivamente.

En el modelo semiparámetrico (4) la matriz $\boldsymbol{K}$ es construida con las covariable**s** $\boldsymbol{x}$ y el parámetro h . Una de las funciones que han mostrado mejor comportamiento

es el kernel gaussiano de la forma $\mathrm{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-h(\boldsymbol{x}_i'\boldsymbol{x}_j)/q_{0.5}\right)$, donde $\boldsymbol{x}_i, \boldsymbol{x}_j$ representan los valores de las covariables para el $i-esimo$ y $j-esimo$ observación respectivamente siendo $(\boldsymbol{x}_i'\boldsymbol{x}_j)$ las distancias al cuadrado de esos observaciones , y $q_{0.5}$ representa un factor de escala que se calcula con el cuantil 0.5 (mediana) de las distancias al cuadrado de todos las observaciones.

Se asume que los vectores $\boldsymbol{y}, \boldsymbol{f}$ son continuos, y la matriz $\boldsymbol{K}$ es un operador lineal, en espacios infinitos de Hilbert conocidos como Reproducing Kernel Hilbert Space (RKHS) [de los Campos et al, 2010, Gianola y Kaam (2008)]. Estos espacios tienen buenas propiedades, la principal es que la norma en el espacio de Hilbert es reproducible en espacios finitos por el álgebra lineal y se adapta bien a procesos gaussianos como el que supone el ruido $\boldsymbol{e}$ del modelo (4).

En este trabajo se usa un método basado en la verosimilitud integrada similar al propuesto por Pérez-Elizalde y Cols (2015), para estimar el ancho de banda *h*.

El MÉTODO DE ESTIMACIÓN CON LA MODA DE LA VEROSIMILITUD INEGRADA

La verosimilitud marginal, conocida también como verosimilitud marginal o distribución predictiva es la integral sobre todos los parámetros de la distribución a posteriori:

$$m_h(\boldsymbol{y}) = \int_{-\infty}^{\infty}\int_{R^n}\int_{0}^{\infty} p(\mu, \boldsymbol{f}, \sigma^2|\boldsymbol{y})\, d\mu\, df\, d\sigma^2$$

Es decir, de acuerdo con el teorema de Bayes

$$p(\mu, \boldsymbol{f}, \sigma^2|\boldsymbol{y}) = p(\boldsymbol{y}|\mu, \boldsymbol{f}, \sigma^2)\, p(\mu)p(\boldsymbol{f})p(\sigma^2)$$

$$m_h(\boldsymbol{y}) = \int_{-\infty}^{\infty}\int_{R^n}\int_{0}^{\infty} p(\boldsymbol{y}|\mu, \boldsymbol{f}, \sigma^2)p(\mu)p(\boldsymbol{f})p(\sigma^2)\, d\mu\, df\, d\sigma^2 \quad (5)$$

El primer paso es integrar sobre el intercepto, considerando una distribución a priori de μ, $p(\mu)$ no informativa o plana. En Berger et al (1999) se justifica plenamente el uso de las distribuciones no informativas e incluso impropias con fines de selección de modelos. Se define

$\boldsymbol{q} = \boldsymbol{y} - \bar{y}\boldsymbol{1}$ donde $\bar{y}$ es la la media de las observaciones $\boldsymbol{y}$ **,** de forma tal que :

$$(\boldsymbol{y} - \mu\boldsymbol{1} - \boldsymbol{Kf})^t(\boldsymbol{y} - \mu\boldsymbol{1} - \boldsymbol{Kf}) = n(\bar{\boldsymbol{y}} - \mu)^2 + (\boldsymbol{q} - \boldsymbol{Kf})^{\boldsymbol{t}}(\boldsymbol{q} - \boldsymbol{Kf})$$

Completando la distribución normal para μ se obtiene:

$$\int_{-\infty}^{\infty} p(\boldsymbol{y}|\mu, \boldsymbol{f}, \sigma^2)p(\mu)\, d\mu = \frac{n^{0.5}\exp\left(-\frac{(\boldsymbol{q}-\boldsymbol{Kf})^{\boldsymbol{t}}(\boldsymbol{q}-\boldsymbol{Kf})}{2\sigma^2}\right)}{(2\pi\sigma^2)^{(n-1)/2}} \quad (6)$$

En un segundo paso, se integra la expresión (6) con respecto de $\boldsymbol{f}$**,** previamente se descompone la matriz $\boldsymbol{K}$ en valores singulares, de tal forma que $\boldsymbol{K} = \boldsymbol{USV^t}$**,** donde las matrices $\boldsymbol{U}, \boldsymbol{V}$ son un matrices ortogonales de orden $\boldsymbol{nxn}$, y $\boldsymbol{S}$ es una matriz diagonal del mismo orden con los valores singulares en su diagonal

ordenados de mayor a menor. Las columnas de $\boldsymbol{U}$, son los componentes principales de $\boldsymbol{K}$ y forman una base para el espacio de los datos, $\boldsymbol{V}$ forma una base que rota los parámetros $\boldsymbol{f}$, de tal forma $\boldsymbol{b} = \boldsymbol{V^t f}$.

Por facilidad y dado que es igual, la integración se realiza a lo largo de $\boldsymbol{b}$ en lugar $\boldsymbol{f}$, la expresión de la integral con respecto a $\boldsymbol{f}$ al integrar la expresión (6), es decir:

$$\int_{-\infty}^{\infty}\int_{R^n} p(\boldsymbol{y}|\,\mu,\boldsymbol{f},\sigma^2)p(\boldsymbol{f}|\sigma^2)d\mu df$$

$$= n^{0.5}\int_{R^n} \frac{\exp\left(-\frac{(\boldsymbol{q}-\boldsymbol{K}\boldsymbol{f})^{\boldsymbol{t}}(\boldsymbol{q}-\boldsymbol{K}\boldsymbol{f})}{2\sigma^2}\right)}{(2\pi\sigma^2)^{(n-1)/2}} p(\boldsymbol{b})d\boldsymbol{b}$$

La distribución a priori $\boldsymbol{b}$ considera una distribución normal con media un vector de ceros y matriz de varianza $\varphi\sigma^2\boldsymbol{S}^{-1}$, donde φ es un parámetro de escala y $\boldsymbol{S}$ es la matriz diagonal con los valores singulares de la descomposición de $\boldsymbol{K}$

$$\int_{-\infty}^{\infty}\int_{R^n} p(\boldsymbol{y}|\,\mu,\boldsymbol{f},\sigma^2)p(\boldsymbol{f}|\sigma^2)d\mu df =$$

$$\frac{n^{0.5}|\varphi\boldsymbol{S}^{-1}|^{-0.5}}{(2\pi\sigma^2)^{(n-1)/2}(2\pi\sigma^2)^{(n)/2}}\int_{R^n}\exp\left(-\frac{(\boldsymbol{q}-\boldsymbol{USb})'(\boldsymbol{q}-\boldsymbol{USb})}{2\sigma^2}-\frac{\boldsymbol{b^t}\frac{1}{\varphi}\boldsymbol{Sb}}{2\sigma^2}\right)d\boldsymbol{b}$$

Completando la normal resulta que los términos restantes son:

$$\int_{-\infty}^{\infty}\int_{R^n} p(\boldsymbol{y}|\,\mu,\boldsymbol{f},\sigma^2)p(\boldsymbol{f}|\sigma^2)d\mu d\boldsymbol{f}$$

$$= \int_{-\infty}^{\infty}\int_{R^n} p(\boldsymbol{y}|\,\mu,\boldsymbol{f},\sigma^2)p(\boldsymbol{f}|\sigma_e^2)d\mu d\boldsymbol{f}$$

$$= \frac{A}{(\sigma^2)^{(n-1)/2}}\exp\left(-\frac{B}{2\sigma^2}\right) \qquad (7)$$

Donde: $A = \frac{n^{0.5}\prod_{i=1}^{n}(s_i^2\varphi+1)^{-0.5}}{(2\pi)^{(n-1)/2}}$ y

$$B = -\boldsymbol{q^t U}(1+\boldsymbol{S^{-2}}\tfrac{1}{\varphi})^{-1}\boldsymbol{U^t q} + \boldsymbol{q^t q}$$

Ahora, si se considera que la distribución a priori de σ^2 es $p(\sigma^2) = \frac{1}{\sigma^2}$ entonces:

$$\int_{-\infty}^{\infty}\int_{R^n}\int_{0}^{\infty} p(\boldsymbol{y}|\mu,\boldsymbol{f},\sigma^2)p(f|\sigma^2)\frac{1}{\sigma^2}d\mu d\boldsymbol{f}d\sigma^2$$

$$= \int_{0}^{\infty}\frac{A}{(\sigma^2)^{\frac{n-3}{2}}}\exp\left(-\frac{B}{2\sigma^2}\right)d\sigma^2$$

Completando la distribución gamma inversa, resulta que:

$$m(\boldsymbol{y}|\varphi,h) = \frac{A\Gamma(\frac{n-1}{2})B^{-(n-1)/2}}{2^{-(n-1)/2}} \qquad (8)$$

Ahora simplificamos la expresión

$$m(\boldsymbol{y}|\varphi,h) \propto \prod_{i=1}^{n}(1+\varphi s_i)^{-\frac{1}{2}}\left[\sum_{i=1}^{n}\frac{\tilde{d}_i^2}{(1+\varphi s_i)}\right]^{-\frac{n-1}{2}} \qquad (9)$$

donde s_i es el $i-esimo$ valor singular, $\tilde{d}_i^2$ es el $i-esimo$ valor de $\tilde{\boldsymbol{d}} = \boldsymbol{U^t q}$.

Puede notarse que h no está explícito, pero los valores singulares y la matriz ortogonal $\boldsymbol{U}$, dependen de h en la descomposición de la matriz $\boldsymbol{K}$ **.**

El método de estimación de los parámetros φ, h

Los parámetros φ, h pueden estimarse con la moda de la distribución a posteriori

$$p(\varphi, h|\boldsymbol{y}) \propto m(\boldsymbol{y}|\varphi, h)p(\varphi, h)$$

Si consideramos plana $p(\varphi, h)$, entonces $p(\varphi, h|\boldsymbol{y}) \propto m(\boldsymbol{y}|\varphi, h)$ y estimar los parámetros φ, h con su moda es equivalente a maximizar la expresión (9), a este método suele conocerse como empirical bayes.

En este trabajo, se estiman los parámetros φ, h que maximizan la expresión (9) con la función optim del paquete R (R Core Team. 2015), que ofrece varios métodos de optimización, basados principalmente en el gradiente descendiente.

Estimación de los parámetros $\boldsymbol{f}$

Por otra parte, la norma al cuadrado de $\boldsymbol{u} = \boldsymbol{Kf}$ en el espacio RKHS es:

$$||\boldsymbol{u}||^2_{\mathcal{H}} = \boldsymbol{f^t K f}$$

Con esta propiedad y usando la regresión Ridge (Hoerl y Kennard, 1970)

$$\hat{\boldsymbol{u}} = \boldsymbol{K}\hat{\boldsymbol{f}} = \boldsymbol{K}(\boldsymbol{K} + (\tfrac{\boldsymbol{1}}{\varphi})\boldsymbol{I})^{-1}(\boldsymbol{y} - \mu\boldsymbol{1}) \quad (11)$$

Predicción de las observaciones

Si estimamos $\hat{\mu} = \bar{y}$, entonces se puede estimar $\hat{\boldsymbol{f}}$, $\hat{\boldsymbol{u}}$ con las expresiones anteriores ya que no dependen de ningún otro parámetro. De tal forma que

$$(\hat{\boldsymbol{y}} - \bar{y}\boldsymbol{1}) = \hat{\boldsymbol{u}} \quad (12)$$

En las expresiones (10) y (11) estamos suponiendo que la matriz $\boldsymbol{K}$ es de entrenamiento de orden nxn, y el vector $\boldsymbol{y}$ de orden $nx1$ son observaciones conocidas (entrenamiento) y la estimación en (11) (12) puede servir para calcular los residuales del modelo contra los datos reales y realizar los análisis que convengan.

Pero si lo que se desea son predicciones de nuevos valores de salidas ante nuevas entradas, se puede demostrar que:

$$\hat{\boldsymbol{u}}_p = \boldsymbol{K}_p(\boldsymbol{K} + \left(\frac{\boldsymbol{1}}{\varphi}\right)\boldsymbol{I})^{-1}(\boldsymbol{y} - \mu\boldsymbol{1}) = \boldsymbol{K}_p\hat{\boldsymbol{f}} \quad (13)$$

$$\left(\hat{\boldsymbol{y}}_p - \bar{y}\boldsymbol{1}\right) = \hat{\boldsymbol{u}}_p \quad (14)$$

Donde $\hat{\boldsymbol{y}}_p$ es un vector de $mx1$ valores predichos de las salidas, $\hat{\boldsymbol{u}}_p$ es también un vector de $mx1$ valores predichos y $\boldsymbol{K}_p$ es una matriz de orden mxn .

EJEMPLO DEL USO DEL MÉTODO

En el apéndice se muestran los códigos en R (R Core Team. 2015), para el método. Los códigos funcionan para un vector respuesta y múltiples vectores de covariables. Para el ejemplo se usa una función no lineal de una sola covariable que tiene como soporte el intervalo de 0 a 10, con la finalidad de mostrar el uso del método, se propone el modelo:

$$\boldsymbol{y} = 2 * \sin(\boldsymbol{x}) + \boldsymbol{e}$$

En este ejemplo cada valor de entrada se añade un ruido que sigue una distribución normal con media cero y desviación estándar igual a 0.10. En total se generaron 200 observaciones. Se tomó una muestra aleatoria de tamaño 160 para entrenamiento y 60 para prueba del método.

La Figura 1, muestra que existe un buen ajuste del modelo y que las predicciones son muy precisas con un cuadrado medio del error de apenas 0.011.

Por supuesto el lector puede usar los códigos para variar la función y ajustar el modelo, o usarlos con datos reales incluso con muchas covariables definidas en una matriz diseño $\boldsymbol{X}$ y con situaciones de colinealidad o de sobreparametrización.

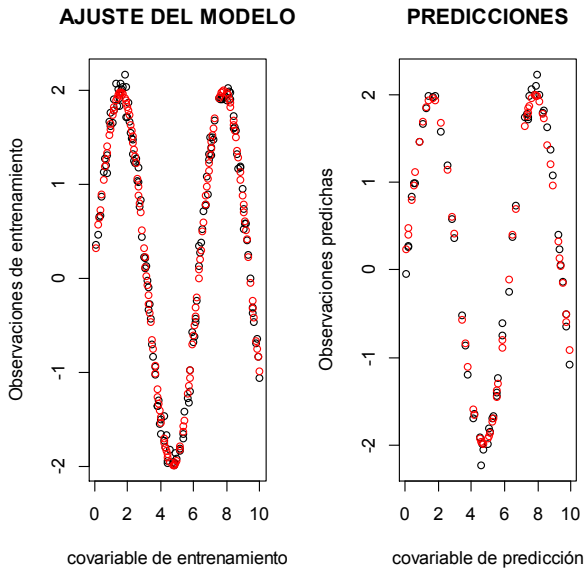

Figura 1.- Gráficas de ajuste de datos de entrenamiento y predicciones con datos de prueba, los puntos negros son los verdaderos y los rojos los estimados con el modelo.

CONCLUSIONES

El método está basado en la distribución predictiva, eliminando inicialmente mediante integración los parámetros de ruido, en particular μ, σ^2, $\boldsymbol{f}$, estableciendo la distribución predictiva de $\boldsymbol{y}$ en función del parámetro φ , que está explicita y del parámetro h que está implícito en los valores singulares y en la rotación de los valores conocidos de la variable respuesta. La predicción se realiza con la moda de la distribución predictiva, auxiliado con un método numérico para estimar en forma conjunta el ancho de banda del kernel (h) y el factor de encogimiento (φ) .

El método funciona en forma rápida y precisa cuando el objetivo es predecir nuevas observaciones. Su aplicación puede ser amplia en muchas disciplinas cuyos problemas puedan modelarse como procesos gaussianos y en problemas de regresión con matrices diseño de rango completo e incompleto.

Sin embargo si el objetivo es encontrar los coeficientes de las covariables, el método es de poca utilidad.

REFERENCIAS

Allen (1974).-Mean square error of prediction as a criterion for selecting variables. Technometrics, 13, 469-475

Berger, J.O., Liseo B., and Wolpert, R. L., (1999), Integrated Likehood Methods for eliminating Nuisance Parameters, Statistical Scienc. Vol 14, No 1, 1-28.

de los Campos, G., Gianola, G., Rosa, G.J.M., Weigel, K.A., Crossa, J. (2010). Semi parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert space methods. Genet. Res 92(4) 295-308

Gianola, D., Fernando, R., Stella, A. (2006). Genomic assisted prediction of genetic value with semiparametric procedure. Genetics 173 (3), 1761-1776

Gianola, D., van Kaam, J., (2008).Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantative Traits.Genetics 178 (4) ; 2289-2303

Hardle B. W. (1990). Applied Non parametric Regression, Cambridge. U.K. Cambridge University Press.

Hoerl, E. A., Kennard, W. R. (1970). Ridge regression: Biased estimation for non-orthogonal problems. Technometrics 12: 55-67.

Nadaraya E. A. (1964).- On estimating regression. Theory prob. Appl. , 9, 141-142

Pérez- Elizalde, S., J. Cuevas, P. Pérez-Rodríguez, and J. Crossa. 2015. Selection of the Bandwidth Parameter in a Bayesian Kernel Regression Model for Genomic-Enabled Prediction. Journal of Agricultural, Biological, and Environmental Statistics (JABES) 5(4):512-532.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Watson, G. S. (1964).-Smooth regression analysis , Shankhya. Serv. A , 26, 359-372