Articulo Original Páginas: 11-15 ISSN
Paradigma Volumen XXII Abril 2021

Métodos De Machine Learning En Regresión Lineal (Métodos De Regularización)

Lic. Noe Panozo Jimenez
SENIOR Carrera de Ingeniería de Sistemas, Escuela Militar de Ingeniería
La Paz, Bolivia
npanozoj@doc.emi.edu.bo

Machine Learning Methods In Linear Regression (Regularization Methods)

Resumen: En este artículo, se pretende motivar a los investigadores sobre el uso de nuevas técnicas de ayuden a mejorar las predicciones de un modelo multivariante. Uno de los aspectos principales del entrenamiento de su modelo de aprendizaje automático es evitar el sobreajuste. El modelo tendrá una precisión baja si esta sobreajustado. Esto sucede porque un modelo se esfuerza demasiado por capturar a los puntos de datos que realmente no representan las verdaderas propiedades de sus datos, sino que son tomados al azar. El aprendizaje de estos puntos de datos hace que su modelo sea más flexible, con el riesgo de sobreajuste, ahí entra el concepto de equilibrio entre sesgo y varianza que es útil para comprender el fenómeno del sobreajuste.

Palabras Clave- *Regresión Lineal, Machine Learning, Regularizacion.*

Abstract: In this article, it is intended to motivate researchers on the use of new techniques to help improve the predictions of a multivariate model. One of the main aspects of training your machine learning model is to avoid overfitting. The model will have a low accuracy if it is overfitted. This happens because a model tries too hard to capture data points that don't really represent the true properties of its data, but are taken at random. Learning these data points makes your model more flexible, with the risk of overfitting, that's where the concept of balance between bias and variance comes in that is useful for understanding the phenomenon of overfitting.

Keywords- *Linear Regression, Machine Learning, Regularization.*

I. INTRODUCCIÓN

En la actualidad, existe una tendencia creciente de almacenar ingentes cantidades de datos con el fin de analizar y extraer algún tipo de información útil de ellos. Sin embargo, el tratamiento de los mismos no resulta trivial y la aplicación de métodos de análisis de datos puede sufrir multitud de problemas tales como sobreajuste o problemas de multicolinealidades causados por la existencia de variables altamente correlacionadas. Por ello, una etapa previa de extracción de características que permita reducir la dimensionalidad de los datos y eliminar dichas multicolinealidades perjudiciales entre variables es crucial para poder aplicar de manera adecuada y eficiente dichas técnicas de análisis de datos. En particular, los métodos de análisis multivariante (MVA) – que permiten extraer un nuevo conjunto de características representativas del problema– gozan de amplia popularidad y han sido aplicados con éxito en una gran cantidad de aplicaciones del mundo real. No obstante, cuando el objetivo consiste en obtener conocimiento de los datos capturados, no solo se requieren buenas prestaciones del sistema diseñado, sino también la capacidad de producir soluciones interpretables que permitan una mejor comprensión del problema. Este artículo permite una extensión de dicho marco general que facilita la inclusión de restricciones adicionales con el fin de proporcionarles habilidades adicionales, donde se reduzca el margen de error de proyección del modelo.

II. OBJETIVO

i) ¿Cómo funcionan los métodos de análisis de datos en el ámbito de los modelos lineales, con relación a la predicción?
ii) ¿Cuán “mejores” son estos nuevos métodos en relación con los tradicionales?
iii) ¿Se podrá hallar un equilibrio del sesgo y la varianza para controlar los errores en el aprendizaje automático?

Método clásico de regresión lineal para predicción

- ✓ Regresión lineal clásica

Método de Machine Learning en Regresión Lineal

- ✓ Regresión Ridge

- ✓ Regresión Lasso
- ✓ Regresión Red Elástica

El objetivo fundamental del Método de Regularización es mejorar el error de predicción del modelo reduciendo la variabilidad de los estimadores (coeficientes) y reduciendo las estimaciones hacia 0, es decir, esta técnica desalienta el aprendizaje de un modelo más complejo o flexible, para evitar el riesgo de sobreajuste.

Una de las formas de evitar el sobreajuste es mediante la validación cruzada, que ayuda a estimar el error sobre el conjunto de pruebas y a decidir que parámetros funcionan mejor para el modelo.

Ridge: Reduce los coeficientes introduciendo un término de penalización igual a la suma de sus cuadrados a través de los coeficientes de penalización, este coeficiente varia de 0 a 1 (sin penalización).

Lasso: el termino de penalización es la suma de coeficientes en valor absoluto, es más extenso que Ridge.

Red elástica: Combina las penalizaciones de Lasso y Ridge. Busca en la cuadricula de valores especificada para encontrar los "mejores" coeficientes de penalización de Lasso y Ridge.

III. MODELO REGRESIÓN ESTÁNDAR (CLÁSICA)

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i \qquad i = 1, \dots, n$$

- A medida que crece la correlación entre las variables, los estimadores de mínimos cuadrados ordinarios (MCO) son más inestables. Por lo tanto, eso hace que los errores de predicción sean mayores. Por esta situación nace la regresión stepwise, para reducir estos errores, que en la actualidad lo siguen usando.

Función objetivo a ser minimizada (no-penalizada)

Suma de cuadrados de los errores o Error cuadrático medio

$$\varphi = \sum_{i=1}^{n} \varepsilon_i^2 \longrightarrow \varphi = \varepsilon\grave{\varepsilon}$$

$$\varphi = \sum_{i=1}^{n} (Y_i - \beta_0 - X_i'\beta)^2 \longrightarrow \varphi = (Y - X\beta)'(Y - X\beta)$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$rango(X) = p + 1$$

$$n > p + 1$$

IV. MODELO DE REGRESIÓN RIDGE

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i \qquad i = 1, \dots, n$$

- Un método para mantener variables altamente correlacionadas en un modelo de regresión con fines de predicción.
- No realiza selección de variables, por tanto, no produce un modelo más interpretable.

Funcion objetivo a ser minimizada (penalizada)

$$\varphi = \sum_{i=1}^{n} (Y_i - \beta_0 - X_i'\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Error de Predicción — Coeficiente de penalización — Función de penalización

$$\boldsymbol{\lambda > 0}$$

función de penalización

$$\text{norma}L2 = \|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2} \text{ norma euclidiana}$$

- Mientras más grande son los $\beta's$, mas grande es la función φ.
- Para minimizar φ se requiere que los $\beta's$ sean pequeños.
- A mayor valor de λ, mayor penalización, mayor contracción de los $\beta's$ ($\beta's$ más pequeños). Parametro de ajuste que decide cuanto queremos penalizar la flexibilidad del modelo.
- El algoritmo de estimación determina el valor de λ que hace φ sea mínimo.
- El algoritmo de estimación de λ se denomina **validación cruzada**.

- El estimador de β es aquél vector que minimiza la función φ para algún valor de λ.
- Si $\lambda \to \infty$, el impacto de la penalización por contracción aumentan y las estimaciones del coeficiente de regresión se acercan a cero, es fundamental la selección de este.

Modelo en términos matriciales

$$\varphi = \sum_{i=1}^{n}(Y_i - \beta_0 - X_i'\beta)^2 + \lambda\sum_{j=1}^{P}|\beta_j|$$

Error de Predicción — Coeficiente de penalización — Función de penalización

Función objetivo a ser minimizada en términos matriciales

$$\varphi = (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta$$

Estimador Ridge de β

$$\hat{\beta}_r = (X'X + \lambda I_{p*p})^{-1}X'Y$$

- Tiene una función objetivo φ diferenciable.
- Tiene una solución de forma cerrada.
- Ninguno de los estimadores Ridge es exactamente cero.

V. MODELO DE REGRESIÓN LASSO

Lasso, (least absolute shrinkage and selection operator = Operador de contracción y selección mínimo absoluto) es un método de análisis de regresión usado para mejorar la precisión de las predicciones y la interpretación del modelo estimado.

Permite:

- Predecir la respuesta dentro y fuera de la muestra.
- Solo se penaliza los altos coeficientes.
- Seleccionar el modelo, que ayuda a una mejor inferencia.
- Selecciona las variables que se ajustan a los datos y testea si esas mismas variables predicen bien la respuesta en otros conjuntos (menor error cuadrático medio).

- Está diseñada para usar cienes o miles, de covariables.
- Se puede incluir más covariables que observaciones en los datos.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$
$$i = 1, \dots, n$$

Función objetivo a ser minimizada (penalizada)

$$Y = X\beta + \varepsilon$$

Función de penalización

$$\text{norma}L1 = \|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$$

$\lambda > 0$ es el parámetro de penalidad Lasso.

Problema

- La función φ no es derivable
- No se tiene una solución explicita de β

Solución

- Lasso elige el valor de λ que minimiza el error cuadrático medio fuera de la muestra con el método de **validación cruzada** (cross - validation), CV selecciona el λ que minimiza el ECM fuera de la muestra
- El β estimado corresponde al λ que minimiza el ECM
- β se estima numéricamente. Método de **coordenada descendente**.

VI. MODELO DE REGRESIÓN ELASTIC-NET (REDES ELÁSTICAS)

Redes Elásticas

- Método de machine learning usado para seleccionar covariables y ajustar modelos lineales.
- Combina las penalidades de Ridge y Lasso.
- Los coeficientes estimados son más robustos ante la presencia de covariables altamente correlacionadas que los coeficientes Lasso.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$
$$i = 1, \dots, n$$

Funcion objetivo a ser minimizada (penalizada)

$$\varphi = \sum_{i=1}^{n}(Y_i - \beta_0 - X_i'\beta)^2 + \lambda \sum_{j=1}^{P}((1-\alpha)\beta_j^2 + \alpha|\beta_j|)$$

Error de Predicción — Coeficiente de penalización — Función de penalidad Ridge — Función de penalidad Lasso

$$\alpha = \begin{cases} 0\ regresión\ Ridge\ (\lambda > 0) \\ 1\ regresión\ Lasso\ (\lambda > 0) \end{cases}$$

$\lambda = 0$ no hay termino de penalidad (estimador MCO)

Redes Elasticas busca un consenso entre Lasso y Ridge a través de α.

La penalidad de Redes Elásticas es un promedio ponderado entre Lasso y Ridge.

- Fue propuesto para variables altamente correlacionadas, especialmente para grupos de variables altamente correlacionadas.
- Al combinar las dos penalidades (de Lasso y Ridge), retiene la propiedad de Lasso de que muchos coeficientes son cero.
- El estimador de β es aquél vector que minimiza la función φ para algún valor de λ y α.

VII. VALIDACIÓN CRUZADA

- Permite encontrar el modelo que minimiza el cuadrado medio del error (error de predicción) fuera de la muestra.
- Es un algoritmo para encontrar el λ^* de modo que los coeficientes correspondientes a λ^* predicen mejor fuera de la muestra.

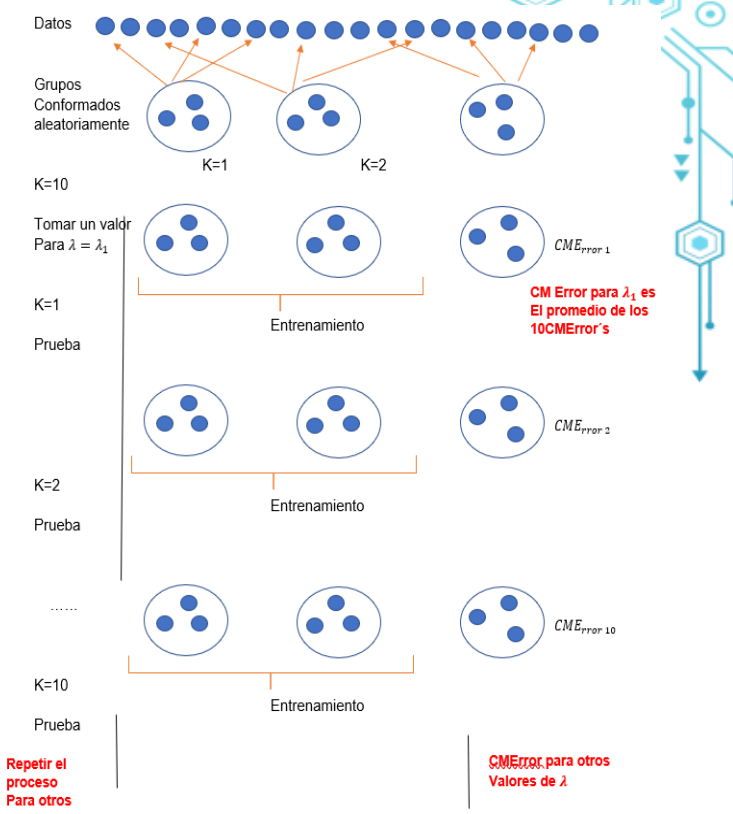
Algoritmo de Validación Cruzada (CV)

i. Dividir aleatoriamente los datos en K grupos

ii. Tomar un λ especifico. Luego, para cada grupo k $\epsilon \{1,2,\dots,K\}$
   a. Estimar los parámetros del modelo (coordenada descendente) con los datos que no están en el grupo K
   b. Usar los estimadores del anterior paso obtener las predicciones con los datos del grupo k (fuera de la muestra)

iii. Para cada λ, calcular el error cuadrático medio fuera de la muestra (en el grupo k)

iv. El valor de λ que general el error cuadrático medio más pequeño es el que minimiza la función CV (función φ)



VIII. CONCLUSIONES

Este Articulo, se propone un marco general MVA que engloba algunos de los métodos de análisis multivariante que ayudan a reducir el error de predicción, por lo tanto, tener una predicción con mayor aproximación a la realidad. Este tipo de soluciones resultan de suma importancia en problemas donde se ha capturado de manera indiscriminada una gran cantidad de datos y se quiere saber cuáles de ellos son relevantes para una determinada tarea. Este tipo de problemas se suelen encontrar en escenarios de "Big Data".

Las ventajas que presenta este marco general MVA son esencialmente las siguientes: **Eficiencia**: Permite obtener soluciones eficientes en función del tamaño de los conjuntos de entrada y salida, reduciendo considerablemente el costo computacional cuando la diferencia entre sus dimensiones es alta.
Flexibilidad o versatilidad: Permite incluir restricciones adicionales en función de las necesidades del problema, de modo que aporta soluciones especializadas para una tarea concreta.

Un modelo estándar de mínimos cuadrados tiende a tener alguna variación, es decir, este modelo no se generalizará bien para un conjunto de datos diferente a sus datos de entrenamiento. La regularización reduce

significativamente la varianza del modelo, sin un aumento sustancial de su sesgo. Por lo tanto, el parámetro de ajuste **λ**, utilizado en las técnicas de regularización descritas, controla el impacto sobre el sesgo y la varianza. A medida que aumenta el valor de **λ**, reduce el valor de los coeficientes y, por lo tanto, reduce la varianza. Hasta cierto punto, este aumento en **λ** es beneficioso ya que solo reduce la varianza (por lo tanto, evita el sobreajuste), sin perder propiedades importantes en los datos. Pero después de cierto valor, el modelo comienza a perder propiedades importantes, lo que genera sesgos en el modelo y, por lo tanto, se desajusta, el valor de **λ se debe seleccionar cuidadosamente**.

Referencias

[1] Allen, G. I., Peterson, C., Vannucci, M. y Maletic-Savatic, M. Regularized partial least squares with an application to NMR spectroscopy. Statistical Analysis and Data Mining, vol. 6(4), páginas 302–314, 2013

[2] Arenas-García, J., Petersen, K., Camps-Valls, G. y Hansen, L. K. Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods. IEEE Signal Process. Mag., vol. 30(4), páginas 16–29, 2013

[3] Bach, F., Jenatton, R., Mairal, J. y Obozinski, G. Convex optimization with sparsity-inducing norms. Optimization for Machine Learning, páginas 19–53, 2011.

[4] Boutsidis, C. y Gallopoulos, E. SVD based initialization: A head start for nonnegative matrix factorization. Journal of Pattern Recognition, vol. 41(4), páginas 1350–1362, 2008.

[5] Dhanjal, C., Gunn, S. R. y Shawe-Taylor, J. Efficient sparse kernel feature extraction based on partial least squares. IEEE Trans. Pattern Anal. and Mach. Intell., vol. 31(8), páginas 1347–1361, 2009.

[6] Guyon, I. y Elisseeff, A. An introduction to variable and feature selection. Journal of Machine Learning Research 3 , páginas 1157–1182, 2003.

[7] Pauca, V. P., Piper, J. y Plemmons, R. J. Nonnegative matrix factorization for spectral data analysis. Linear algebra and its applications, vol. 416(1), páginas 29–47, 2006.

[8] De la Torre, F. A least-squares framework for component analysis. IEEE Trans. Pattern Anal. and Mach. Intell., vol. 34(6), páginas 1041–1055, 2012.

[9] Yamanishi, Y., Vert, J., Nakaya, A. y Kanehisa, M. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics, vol. 19(suppl 1), páginas i323–i330, 2003.

Fecha de Envió del Articulo: 8/10/2020
Fecha de Aceptación de artículo: 20/10/2020