Selección de variables mediante regresiones penalizadas en grandes volúmenes de datos.

Leandro Kovalevski

Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística, Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario, Rosario, Argentina

lkovalevski@fcecon.unr.edu.ar

Abstract. Uno de los retos más importantes del análisis estadístico en esos grandes volúmenes de datos es detectar cuál es la información valiosa.
Las regresiones regularizadas, tales como Lasso o SCAD son alternativas usuales cuando los métodos usuales de selección de variables en los modelos lineales resultan no apropiados o no aplicables.
El objetivo de este trabajo es comparar el desempeño de las regresiones Lasso y SCAD en escenarios donde el número de variables importantes esté próximo al número de observaciones.
Se simularon distintos conjuntos de datos en los cuales las variables respuestas son funciones lineales de varios predictores independientes y errores que se simulan con distribución normal.
Las regresiones LASSO y SCAD se ofrecen como una alternativa válida para selección de variables con gran número de predictores, aunque pueden no ser las más adecuadas cuando la cantidad de efectos significativos se aproxima al número de observaciones.

Keywords: Regresiones regularizadas, LASSO, SCAD, selección de predictores.

1 Introducción

El manejo y análisis de grandes volúmenes de información no es un tema nuevo sino no que durante años ha acaparado mucha atención aunque hay algunas características de nuestros días que hacen que el tema cobre mucha notoriedad. Entre esas particularidades podemos destacar la casi instantánea digitalización de diversas fuentes, en su mayoría de datos no estructurados y las herramientas informáticas actuales que facilitan la aplicación de modelos estadísticos para analizar esos grandes volúmenes de información compleja.

Uno de los retos más importantes del análisis estadístico en esos grandes volúmenes de datos es detectar cuál es la información valiosa. Cuando las variables intervienen en un modelo predictivo, una de las técnicas preferidas son los modelos de regresión. En los casos en que el número de predictores de los modelos de regresión sea muy alto, los métodos usuales de selección de variables resultan no apropiados y a

adfa, p. 1, 2011.

veces no aplicables. Para solucionar esas situaciones surgen las regresiones regularizadas, tales como Lasso (Least Absolute Shrinkage and Selection Operator) o SCAD (Smoothly Clipped Absolute Deviation) que en el proceso de estimación de los efectos buscan penalizar por la magnitud de los coeficientes de regresión con el objetivo de reducir el número de los mismos.

Estas regresiones logran buenos resultados para detectar un pequeño conjunto de predictores significativos entre muchos predictores candidatos (Xie & Huang, 2009) (Castro, 2013) aunque no ha sido definida una regla para establecer la relación que deben guardar el número de predictores significativos y el número total de observaciones.

2 Objetivos

El objetivo de este trabajo es comparar el desempeño de las regresiones Lasso y SCAD en escenarios donde el número de variables importantes esté próximo al número de observaciones.

3 Materiales y métodos

3.1 Simulación de los datos

Se simularon distintos conjuntos de datos en los cuales las variables respuestas son funciones lineales de varios predictores independientes que siguen una distribución normal con media igual a cero y desviación estándar igual a 2. Los errores en el modelo lineal se simulan normales (0,1) y en todos se agregó un término independiente. La cantidad de predictores varía entre 10 y 490, de 10 en 10 y el número de observaciones es siempre igual a 500. Los coeficientes de los predictores fueron elegidos aleatoriamente de una distribución uniforme (-0,1; 3). Cada uno delos 49 escenarios se simularon una única vez.

3.2 Métodos

Las regresiones LASSO (Tibshirani, 1996) y SCAD (Fan, 2001) son modelos de regresión penalizados que tienen como objetivo controlar el sobreajuste en los modelos con gran cantidad de predictores. La selección de variables se realiza en el momento de la estimación de los parámetros, los cuales se estiman a través de:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

$$\hat{\beta}_{SCAD} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \sum_{j=1}^{p} \phi_\lambda(\beta_j) \right\},$$

$$con\ \phi_\lambda = \begin{cases} \lambda|\beta_j|, & |\beta_j| \le \lambda \\ -(\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2)/(2(a-1), & \lambda < |\beta_j| \le a\lambda \\ \dfrac{(a+1)\lambda^2}{2}, & |\beta_j| > a\lambda \end{cases}$$

Siendo $a > 2\ y\ \lambda > 0$ donde λ es elegido por validación cruzada y $a \approx 3{,}7$ es un valor sugerido por los autores.

El criterio elegido para comparar los modelos fue la raíz del error cuadrático medio (RECM)= $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$.

4 Resultados

Para comparar las distintas situaciones simuladas las RECM fueron evaluadas en función del cociente entre el número de observaciones y el número de predictores.

La menor RECM en los distintos escenarios se observó para el modelo de regresión lineal completo. En aquellos escenarios que la cantidad de observaciones por predictor significativos fueron mayores o iguales a 10, la RECM es muy similar para los tres modelos.

Cuando la cantidad de observaciones por predictor significativo disminuyen (menos de 2 por predictor), las estimaciones de la regresión LASSO son las primeras afectadas. La regresión SCAD tuvo resultados similares al modelo de regresión completo hasta que se tienen 1,13 observaciones por predictor, escenario a partir del cual su RECM comienza a incrementarse y volverse inestable.

Conclusiones

Las regresiones LASSO y SCAD se ofrecen como una alternativa válida para selección de variables cuando hay un gran número de predictores. El modelo de regresión lineal tiene sus limitaciones para encontrar los efectos significativos si bien conserva sus propiedades de mínima RECM aun cuando el número de observaciones por predictor es bajo.

Cuando la cantidad de efectos significativos se aproxima al número de observaciones, las estimaciones de los efectos y las predicciones de los modelos de regresiones regularizadas pueden no ser las más adecuadas.

Trabajos citados

Castro, S. (2013). *Análisis De Datos En Grandes Dimensiones. Estimación Y Selección De Variables En Regresión.* . Montevideo, Uruguay.: Universidad de La República.

Fan, J. a. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96 1348–1360.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. . Soc B. Vol. 58, No. 1, pages 267-288.

Xie, H., & Huang, J. (2009). Scad-Penalized Regression In High-Dimensional Partially Linear Models. *The Annals of Statistics*, Vol. 37, No. 2, 673–696.