MDSF Predicción

REGULARIZACIÓN

Prof. Dr. Ricardo A. Queralt

Índice

Bibliografía

- An Introduction to Statistical Learning with Applications in R. Springer
  Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
- "Linear Model Selection & Regularization" `https://rpubs.com/ryankelly/reg`
- "Selección de predictores y mejor modelo lineal múltiple: subset selection, ridge regression, lasso regression y dimension reduction" `https://rpubs.com/Joaquin_AR/242707`
- "Regression with splines: Should we care about non-significant conponents?" `https://freakonometrics.hypotheses.org/47681`
- R for Statistical Learning. David Dalpiaz. Chapter 24 `https://daviddalpiaz.github.io/r4sl/regularization.html`

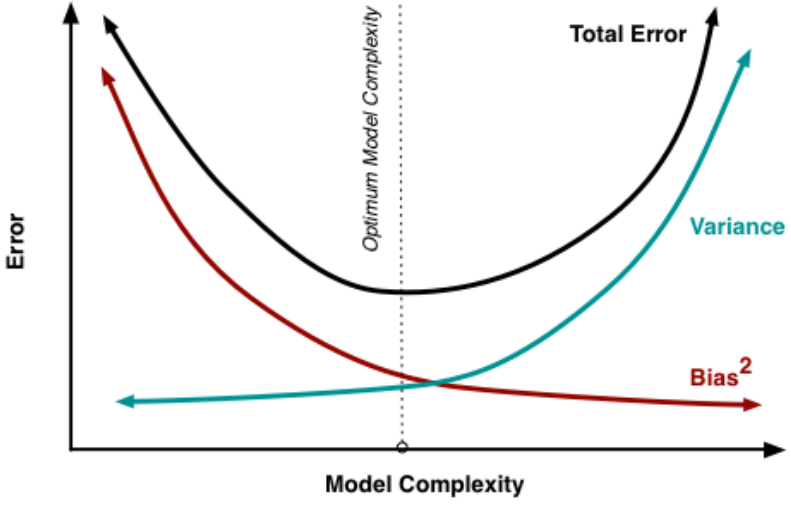
★ Glmnet Vignette. Trevor Hastie and Junyang Qian `https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html`

1. Modelos de Regularización

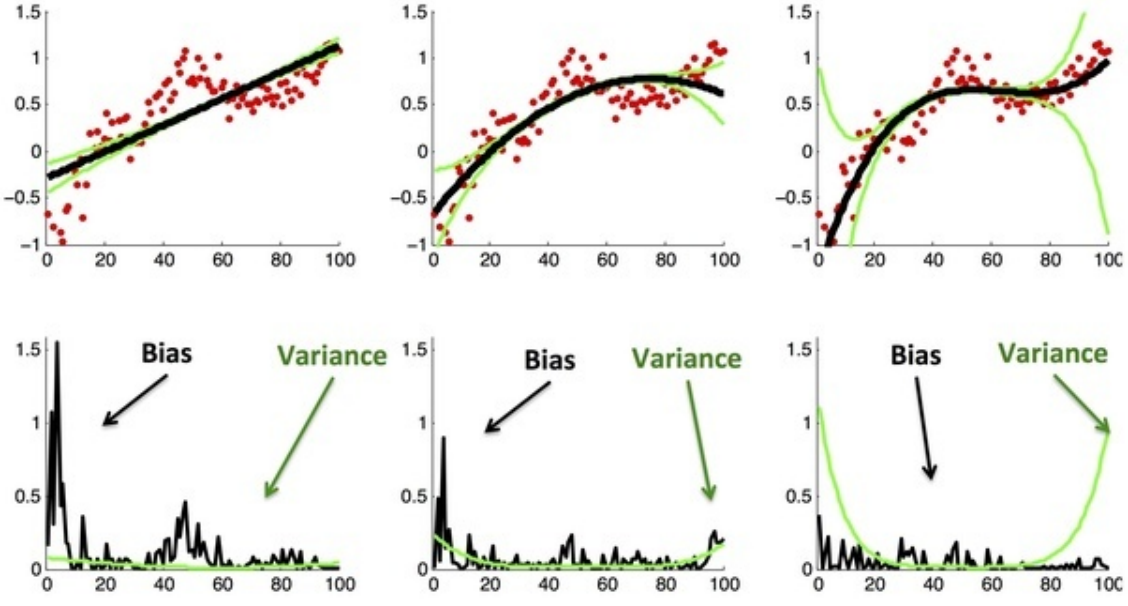
- ★ La relación entre la capacidad de ajuste y generalización de un modelo.
- ★ Cuando se logra un gran ajuste, la diferencia entre los datos reales y la estimación del modelo es pequeña, en este caso el sesgo también es pequeño, pero estos buenos resultados de ajuste, van de la mano con el aumento en la complejidad del modelo, cuando se aumenta la complejidad del modelo este se vuelve sensible a pequeñas variaciones en los datos de entrada, fluctuando en función de estos, es así cuando la varianza aumenta.
- ★ Esta claro que en Machine Learning se busca crear modelos que ofrezcan dos características esenciales: ajuste a los datos y generalización.

- ★ Esto acentúa la necesidad de encontrar un balance entre sesgo y varianza, o visto de otra forma, entre error y complejidad.
- ★ El MSE(Error Cuadrático Medio):

$$\text{MSE}(\hat{\theta}) = \text{E}_{\hat{\theta}}\left[(\hat{\theta} - \theta)^2\right] = \text{Var}_{\hat{\theta}}(\hat{\theta}) + \text{Bias}(\hat{\theta}, \theta)^2.$$



El mejor



2. Shrinkage Methods (Métodos de contracción)

- ★ Los métodos de selección de subconjuntos utilizan ajustes por mínimos cuadrados que contenían un subconjunto de los predictores para elegir el mejor modelo y estimar el error.
- ★ Aquí, discutimos una alternativa donde ajustamos un modelo que contiene todos los p predictores usando una técnica que restringe o regulariza las estimaciones de coeficientes, o equivalentemente, que reduce las estimaciones de coeficientes hacia cero.
- ★ La reducción de las estimaciones de los coeficientes tiene el efecto de reducir significativamente su varianza.

- ★ Ajustar el modelo incluyendo todos los predictores pero empleando un método que fuerce a que las estimaciones de los coeficientes de regresión tiendan a cero, es decir, que tienda a minimizar la influencia de los predictores menos importantes.
- ★ Dos de los métodos más empleados son:
  - ✓ **Ridge regression**: aproxima a cero los coeficientes de los predictores pero sin llegar a excluir ninguno.
  - ✓ **Lasso**: aproxima a cero los coeficientes, llegando a excluir predictores.

3. Ridge

- La regresión Ridge es similar a los mínimos cuadrados, excepto que los coeficientes se estiman minimizando una cantidad ligeramente diferente.
- La regresión de Ridge, como OLS, busca estimaciones de coeficientes que reducen el RSS, sin embargo, también tienen una penalización por contracción cuando los coeficientes se acercan a cero.
- Esta penalización tiene el efecto de reducir las estimaciones del coeficiente hacia cero.
- Un parámetro, λ, controla el impacto de la contracción. $\lambda = 0$ se comportará exactamente como la regresión OLS. Por supuesto, la selección de un buen valor para λ es crítica, y debe elegirse utilizando técnicas

de validación cruzada.

$$J(\beta) = \lambda\|\beta\|^2 + \sum_i (\beta^T x_i - y_i)^2. \quad (1)$$

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2 \quad (2)$$

$$\frac{\partial J}{\partial \beta} = \lambda\beta + \sum_i (\beta^T x_i - y_i)x = 0 \quad (3)$$

$$(XX^T + \lambda I)\beta = Xy \quad (4)$$

$$\beta = (XX^T + \lambda I)^{-1}Xy \quad (5)$$

$$= X\alpha. \quad (6)$$

¿Por qué la regresión cresta es mejor que los mínimos cuadrados?

- ★ La ventaja es evidente dentro de la disyuntiva de sesgo-varianza.
- ★ A medida que aumenta λ, la flexibilidad del ajuste de regresión de cresta disminuye. Esto conduce a disminuir la varianza, con un aumento menor en el sesgo. La regresión regular de MCO se fija con alta varianza, pero sin sesgo. Sin embargo, la prueba MSE más baja tiende a ocurrir en la intersección entre la varianza y el sesgo. Por lo tanto, ajustando adecuadamente λ y adquiriendo menos varianza a costa de una pequeña cantidad de sesgo, podemos encontrar un MSE potencial más bajo.
- ★ La regresión de Ridge funciona mejor en situaciones donde las estimaciones de mínimos cuadrados tienen alta varianza.
- ★ La regresión de Ridge es mucho más eficiente computacionalmente que

cualquier método de subconjunto, ya que es posible resolver simultáneamente todos los valores de λ.

4. Lasso

- La regresión Ridge tenía al menos una desventaja; incluye todos los predictores p en el modelo final. El término de penalización establecerá muchos de ellos cerca de cero, pero nunca exactamente a cero.
- En general, esto no es un problema para la precisión de la predicción, pero puede hacer que el modelo sea más difícil de interpretar los resultados.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| = RSS + \lambda \sum_{j=1}^{p}|\beta_j|$$

$$min_{\beta_0,\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2\right\} \text{ sujeto a } \sum_{j=1}^{p}|\beta_j|\leq t.$$

$$\min_{\beta_0,\beta}\left\{\frac{1}{N}\|y-\beta_0-X\beta\|_2^2\right\} \text{ subject to } \|\beta\|_1\leq t.$$

$$\|Z\|_p=\left(\sum_{i=1}^{N}|Z_i|^p\right)^{1/p}$$

es la p-norma ℓ^p

- ★ Lasso supera esta desventaja y es capaz de forzar algunos de los coeficientes a cero, dado que s es lo suficientemente pequeño. Como s = 1 da como resultado una regresión OLS regular, cuando s se acerca a 0,

los coeficientes se reducen a cero. Por lo tanto, la regresión de Lasso también realiza la selección de variables.

★ No hay un algoritmo dominante presente aquí, en general, es mejor probar las tres técnicas introducidas hasta ahora y elegir la que mejor se adapte a los datos usando estimaciones de error de prueba con validación cruzada.

5. Elastic Net (Red elástica)

★ La red elástica es otra penalización que incorpora la selección variable del lazo y la contracción de predictores correlacionados como la regresión de ridge.

$$\min_{\beta \in \mathbb{R}^p} \left\{ \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right\}$$

★ Es equivalente a:

$$\min_{\beta_0, \beta} \left\{ \|y - \beta_0 - X\beta\|_2^2 \right\} \text{ sujeto a } (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \leq t, \text{ donde } \alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$$

★ Si $\alpha = 1$ estamos en el caso Lasso y si $\alpha = 0$ estamos en el caso Ridge.

Colegio Universitario de Estudios Financieros (CUNEF)

MUCHAS GRACIAS

ricardo.queralt@cunef.edu

@raqueralt

Prof. Dr. Ricardo A. Queralt