Selección de Modelos y Regularización

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

Agenda

Regularización: Motivación

- Las técnicas econometricas estándar no están optimizadas para la predicción porque se enfocan en la insesgadez.
- OLS por ejemplo es el mejor estimador lineal *insesgado*
- OLS minimiza el error *"dentro de muestra"*, eligiendo β de forma tal que

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 \tag{1}$$

- pero para predicción, no estamos interesados en hacer un buen trabajo dentro de muestra
- Queremos hacer un buen trabajo, **fuera de muestra**

Agenda

Ridge

- Asegurar cero sesgo dentro de muestra crea problemas fuera de muestra: trade-off Sesgo-Varianza
- Las técnicas de machine learning fueron desarrolladas para hacer este trade-off de forma empírica.
- Vamos a proponer modelos del estilo

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} R(\beta_j) \tag{2}$$

- donde R es un regularizador que penaliza funciones que crean varianza
- Explícitamente en la minimización incluimos un termino de sesgo y un termino de varianza.

Ridge

- Para un $\lambda \geq 0$ dado, consideremos ahora el siguiente problema de optimización

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} (\beta_j)^2 \tag{3}$$

App

Formalmente

- Las Xs estan estandarizadas (x_i con media 0 ($\bar{x} = 0$) y varianza 1 ($\sum x_i^2 = 1$))
- Regresión: $y = \beta x + u$
- OLS

$$\hat{\beta}_{ols} = \sum x_i y_i$$

- Ridge

$$\hat{\beta}_{ridge} = \frac{\sum x_i y_i}{(1+\lambda)} = \frac{\hat{\beta}_{ols}}{(1+\lambda)}$$

Formalmente

- En regresión multiple (X es una matriz $n \times k$)
- Regresión: $y = X\beta + u$
- OLS

$$\hat{\beta}_{ols} = (X'X)^{-1}X'y$$

- Ridge

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1}X'y$$

Ridge vs OLS

- Ridge es sesgado $E(\hat{\beta}_{ridge}) \neq \beta$
- Pero la varianza es menor que la de OLS
- Para ciertos valores del parámetro λ $MSE_{OLS} > MSE_{ridge}$
- Mostremos esto para el caso de 1 variable

Ridge vs OLS

- OLS:
  - Sesgo $E(\hat{\beta}_{ols}) - \beta =$
  - Varianza $V(\hat{\beta}_{ols}) =$
  - $MSE(\hat{\beta}_{ols}) =$
- Ridge:
  - Sesgo $E(\hat{\beta}_{ridge}) - \beta =$
  - Varianza $V(\hat{\beta}_{ridge}) =$
  - $MSE(\hat{\beta}_{ridge}) =$

Ridge vs OLS

$$MSE(\hat{\beta}_{ols}) - MSE(\hat{\beta}_{ridge}) = \tag{4}$$

Agenda

Escala de las variables

- La escala de las variables importa en Ridge, mientras que en OLS no.
- Tiene consecuencias
  - En la solución ($\hat{\beta}$)
  - En la predicción ($\hat{y}$)

Escala de las variables

Ridge no es invariante a las escala

- Supongamos $z = c * x$
- Vamos a mostrar que $\hat{y}_i^z = \hat{y}_i^x$
- Partamos del modelo

$$y_i = \beta_0^z + \beta_1^z z_i + u \tag{5}$$

$$\hat{\beta}_1^z = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})^2} \tag{6}$$

Escala de las variables

Ridge no es invariante a las escala

- Continuando

$$\hat{\beta}_1^z = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})^2} \tag{7}$$

- Pero $z = c * x$

$$\hat{\beta}_1^z = \frac{\sum(cx_i - c\bar{x})(y_i - \bar{y})}{\sum(cx_i - c\bar{x})^2} \tag{8}$$

$$= \frac{1}{c}\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{9}$$

$$= \frac{1}{c}\hat{\beta}_1^x \tag{10}$$

- En Ridge?

Escala de las variables

Ridge no es invariante a las escala

- Para un $\lambda \geq 0$ dado, el problema de optimización

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0^z - \beta_1^z z_i)^2 + \lambda (\beta_1^z)^2 \tag{11}$$

- Demo: baticomputer, math: Homework

Escala de las variables

Ridge no es invariante a las escala

- En la predicción

$$\hat{\beta}_1^z z_i = \hat{\beta}_1^z c x_i \tag{12}$$

$$= \frac{1}{c}\hat{\beta}_1^x c x_i \tag{13}$$

$$= \hat{\beta}_1^x x_i \tag{14}$$

Escala de las variables

Ridge no es invariante a las escala

- En términos generales, si $Z = cX$

$$
\begin{aligned}
\hat{\beta}^{Z}_{OLS} &= (Z'Z)^{-1}Z'y \\
&= ((cX)'(cX))^{-1}(cX)'y \\
&= \frac{c}{c^2}(X'X)^{-1}X'y \\
&= \frac{1}{c}(X'X)^{-1}X'y \\
&= \frac{1}{c}\hat{\beta}^{X}_{OLS}
\end{aligned}
$$

Escala de las variables

Ridge no es invariante a las escala

- Entonces

$$
\begin{aligned}
\hat{\beta}^{Z}_{OLS} Z &= \frac{1}{c} \hat{\beta}^{X}_{OLS} c X \\
&= \hat{\beta}^{X}_{OLS} X
\end{aligned}
$$

- Con Ridge esto no funciona

$$
\hat{\beta}^{Z}_{Ridge} Z \neq \hat{\beta}^{X}_{Ridge} X
$$

- Es importante estandarizar las variables (la mayoría de los softwares lo hace automáticamente)

Agenda

1. Ridge
   - Trade-off Sesgo-Varianza
   - Escala de las variables
   - More predictors than observations
   - Selección de λ
2. Lasso
3. Recap

More predictors than observations ($k > n$)

- What happens when we have more predictors than observations ($k > n$)?
  - OLS fails
  - Ridge ?

OLS when $k > n$

- Rank? Max number of rows or columns that are linearly independent
  - Implies $rank(X_{k \times n}) \leq min(k, n)$
- MCO we need $rank(X_{k \times n}) = k \implies k \leq n$
- If $rank(X_{k \times n}) = k$ then $rank(X'X) = k$
- If $k > n$, then $rank(X'X) \leq n < k$ then $(X'X)$ cannot be inverted
- Ridge and Lasso work when $k \geq n$

Ridge when $k > n$

$$min_{\beta}R(\beta) = \sum_{i=1}^{n}(y - x\beta)^2 + \lambda(\beta)^2 \tag{15}$$

- Solution $\rightarrow$ data augmentation
- Intuition: Ridge "adds" k additional points.
- Allows us to "deal" with $k \geq n$

Ridge when $k > n$

$$min_\beta R(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda(\beta_s)^2 \tag{16}$$

Ridge when $k > n$

$$min_{\beta}R(\beta) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{k} x'_{ij}\beta_j)^2 + \lambda(\sum_{j=1}^{k}\beta_j)^2 \tag{17}$$

Agenda

1. Ridge
   - Trade-off Sesgo-Varianza
   - Escala de las variables
   - More predictors than observations
   - Selección de λ
2. Lasso
3. Recap

Selección de λ

- Asegurar cero sesgo dentro de muestra crea problemas fuera de muestra: trade-off Sesgo-Varianza
- Ridge hace este trade-off de forma empírica.

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} R(\beta_j) \tag{18}$$

- λ es el precio al que hacemos este trade off
- Como elegimos λ?

Selección de λ

- λ es un hiper-parámetro y lo elegimos usando validación cruzada
  - Partimos la muestra de entrenamiento en K Partes:
    $MUESTRA = M_{fold\,1} \cup M_{fold\,2} \cdots \cup M_{fold\,K}$
  - Cada conjunto $M_{fold\,K}$ va a jugar el rol de una muestra de evaluación $M_{eval\,k}$.
  - Entonces para cada muestra
    - $M_{train-1} = M_{train} - M_{fold\,1}$
    - $\vdots$
    - $M_{train-k} = M_{train} - M_{fold\,k}$

Selección de λ

- Luego hacemos el siguiente loop
  - Para $i = 0, 0.001, 0.002, \dots, \lambda_{max}$ {

    - Para $k = 1, \dots, K$ {

      - Ajustar el modelo $m_{i,k}$ con λ_i en $M_{train-k}$

      - Calcular y guardar el $MSE(m_{i,k})$ usando M_{eval-k}

    } # *fin para k*

    - Calcular y guardar $MSE_i = \frac{1}{K} MSE(m_{i,k})$

    } # *fin para* λ

- Encontramos el menor MSE_i y usar ese $\lambda_i = \lambda^*$

Agenda

1. Ridge
   - Trade-off Sesgo-Varianza
   - Escala de las variables
   - More predictors than observations
   - Selección de λ

2. Lasso

3. Recap

Lasso

Agenda

1. Ridge
   - Trade-off Sesgo-Varianza
   - Escala de las variables
   - More predictors than observations
   - Selección de λ
2. Lasso
3. Recap

Recap

- Próxima clase: detalles de lasso