Regularizacion y eleccion de modelos

Walter Sosa-Escudero

Universisad de San Andrés y CONICET

Eleccion de modelos

Trade off-sesgo varianza

$$Y = X_1\beta_1 + X_2\beta_2 + u$$

- Trade off: modelos mas 'complejos' tienden a ser menos sesgados pero con mayor varianza.
- Preferencia 'lexicografica' por la insesgadez: minimizar ECM es minimizar varianza.
- Desafio: tolerar sesgos para bajar considerablemente la varianza.

Eleccion de modelos

- $M_k,\ k = 1, \ldots, K$ serie modelos.
- Ordenables en complejidad? Trade off-sesgo varianza. Overfit
- Anidados?
- Existe un 'supermodelo' que contenga al verdadero?

Eleccion de modelos: busqueda en M_k del mejor modelo para predecir fuera de la muestra.

Ejemplo: modelo lineal estimado por MCO, k es la cantidad de regresores.

Busqueda exhaustiva

$$Y = X\beta + u, \quad X_{n \times p}$$

1. Estimar *todos* los posibles modelos con $k = 1, 2, \ldots, p$ predictores. 2^p modelos.
2. Para cada modelo computar el error de prediccion por cross validation.
3. Elegir el que minimiza CV.

Problema: computacionalmente inviable para p moderado ($p = 20$, 1.048.576 modelos)

Stepwise selection

Forward selection:

- Empezar sin ningun predictor
- Probar todos los modelos con 1 predictor. Elegir el que minimiza CV.
- Agregar de a 1, sin quitar los ya incorporados. $p(p+1)/2$ modelos.
- De los p modelos elegidos, elegir el que minimiza CV.

Backward selection: empieza con el modelo completo. Busqueda no exhaustiva. Los incoporados no 'salen'. Forward tiene una ventaja en modelos de alta dimension (mas adelante).

Regularizacion

¿No vale empezar con el modelo general y tachar los coeficientes no significativos?

- Backward selection aproxima esa idea (no exactamente ya que el criterio es de ajuste/prediccion.
- 'Tachar' variables/coeficientes es una forma extrema de 'achicarlos'.
- Lasso: una manera formal y algoritmica de realizar esa tarea.
- Regularizar: utilizar informacion de afuera del modelo para simplificarlo.
- Recuerden que el objetivo es minimizar ECM fuera de la muestra de entrenamiento.

LASSO

Para $\lambda \geq 0$ dado, consideremos la siguiente funcion objetivo (a minimizar):

$$R_l(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} |\beta_s|$$

(el primer coeficiente corresponde al intercepto).

- ¿Si $\lambda = 0$?
- ¿Si $\lambda = \infty$?

- $\sum_{i=1}^{n}(y_i - x_i'\beta)^2$ penaliza falta de ajuste.
- ¿ $\sum_{s=2}^{p} |\beta_s|$?

$$R_l(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} |\beta_s|$$

- LASSO magic: automaticamente elige que variables entran $(\beta_s \neq 0)$ y cuales no $(\beta_s = 0)$
- Por que? Coeficientes anulados como 'soluciones de esquina'
- $R_l(\beta)$ es una funcion no diferenciable.

Soluciones de esquina

Caso extremo (facil de generalizar)

$$R_l(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta|$$

- Un solo predictor, un solo coeficiente.
- Predictor estandarizado. De modo que $\sum x^2 = 1$.
- En este caso, el estimador MCO es

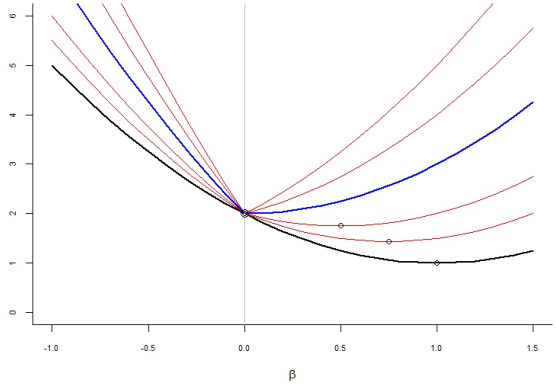
$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \sum x_i y_i$$

$$R_l(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda|\beta|$$

$$R_l(\beta) = \begin{cases} \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta, \text{ si } \beta \geq 0 \\ \sum_{i=1}^{n}(y_i - x_i\beta)^2 - \lambda\beta, \text{ si } \beta \leq 0 \end{cases}$$

- No diferenciable en $\beta = 0$.
- Cuadratica y diferenciable para $\beta \neq 0$.

Graficamente ($\hat{\beta} > 0$):



$\frac{dR_l(0)^+}{d\beta} > 0$, solucion de *esquina* $\hat{\beta}_l = 0$, caso contrario, solucion *interior*.

$$\begin{aligned}\frac{dR_l(\beta)^+}{d\beta} &= -2\sum y_i x_i + 2\beta \sum x_i^2 + \lambda \\ &= -2\sum y_i x_i + 2\beta + \lambda\end{aligned}$$

$$\frac{dR_l(0)^+}{d\beta} = -2\sum y_i x_i + \lambda$$

Entonces, si

$$\lambda \geq 2\sum y_i x_i, \quad \hat{\beta}_l = 0$$

Si $\lambda < 2\sum y_i x_i$, la solucion es *interior*. FOC:

$$-2\sum y_i x_i + 2\hat{\beta}_l + \lambda = 0$$

$$\begin{aligned} \hat{\beta}_l &= \sum x_i y_i - \lambda/2 \\ &= \hat{\beta} - \lambda/2 \end{aligned}$$

- Shrinkage: la solucion esta corrida hacia cero con respecto a $\hat{\beta}$ (MCO).
- El caso $\hat{\beta} < 0$ es completamente simetrico.

Intuiciones para economistas

$$\hat{\beta}_l = \begin{cases} 0 & \text{si } \lambda \geq 2\sum y_i x_i \\ \hat{\beta} - \lambda/2 & \text{si } \lambda \leq 2\sum y_i x_i \end{cases}$$

- Intuicion? En 0 el costo por λ sube y el de ajuste baja.
- Si uno sube mas rapido que lo que el otro baja: conviene quedarse en cero.
- En caso contrario conviene moverse afuera de cero.
- Moverse de cero si la relacion es lo suficientemente fuerte, sino evitar la penalizacion.

Ridge

Ridge

Para $\lambda \geq 0$ dado, consideremos la siguiente funcion objetivo (a minimizar):

$$R_r(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p}(\beta_s)^2$$

(el primer coeficiente corresponde al intercepto).

Las intuiciones coinciden con LASSO, pero el problema es completamente diferente.

Consideremos el caso simple (un predictor, normalizado)

$$R_r(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda\beta^2$$

FOC:

$$-\sum y_i x_i + 2\beta + 2\lambda\beta = 0$$

Despejando

$$\hat{\beta}_r = \frac{\sum y_i x_i}{1+\lambda} = \frac{\hat{\beta}}{1+\lambda}$$

- La solucion es siempre *interior* (comparar intuicion con LASSO)
- Nuevamente, la solucion esta 'corrida hacia cero' con respecto a $\hat{\beta}$ (shrinkage).

Ridge vs. MCO

Bajo los supuestos clasicos y en el caso simple:

- $E(\hat{\beta}) = \beta$ (insesgado)
- $V(\hat{\beta}) = \sigma^2 / \sum x_i^2 = \sigma^2$
- $ECM(\hat{\beta}) = \sigma^2$.

- $E(\hat{\beta}_r) = \beta/(1+\lambda)$ (sesgado)
- $V(\hat{\beta}_r) = \sigma^2/(1+\lambda)^2$ (menor varianza)
-

$$ECM(\hat{\beta}) = \left[\beta - \frac{\beta}{1+\lambda}\right]^2 + \frac{\sigma^2}{(1+\lambda)^2}$$

$$
\begin{aligned}
ECM(\hat{\beta}) - ECM(\hat{\beta}_r) &= \sigma^2 - \frac{\beta^2\lambda^2 + \sigma^2}{(1+\lambda)^2} \\
&= \frac{\lambda(2\sigma^2 - \beta^2\lambda + \lambda\sigma^2)}{(1+\lambda^2)}
\end{aligned}
$$

- Es suficiente que $\lambda < 2\sigma^2/\beta^2$ para que $ECM(\hat{\beta}) - ECM(\hat{\beta}_r) > 0$
- Para todo β y σ^2 existe λ de modo que ridge le 'gana' a MCO.
- En este caso es posible derivar una condicion necesaria y suficiente. Pero no es generalizable al caso de p variables (Theobald, 1974).

Comentarios tecnicos

1. No es posible derivar un resultado exacto similar para LASSO.
2. Ridge muchas veces aproxima a LASSO.
3. Muchas variables y no normalizadas? Hiper simple (usando el teorema de FWL: maestria).
4. El argumento de solucion de esquina en LASSO es un tanto informal. ¿Formal? Subgradientes, analisis convexo (Rockafellar, 1996, p. 264, muy intuitivo).
5. Usar con datos previamente estandarizados.

Alternativas y extensiones

- Elastic Net: $R_e(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda_1 \sum_{s=2}^{p} |\beta_s| + \lambda_2 \sum_{s=2}^{p} \beta_s^2$
- LASSO logit: $R_l(\beta) = L(\beta) + \lambda \sum_{s=2}^{p} |\beta_s|$

Ejemplo: bateadores

Ejemplo: Hitters

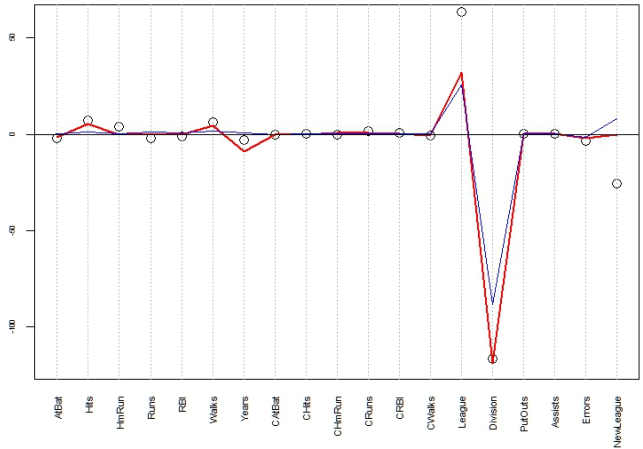
`Hitters`

Format

A data frame with 322 observations of major league players on the following 20 variables.

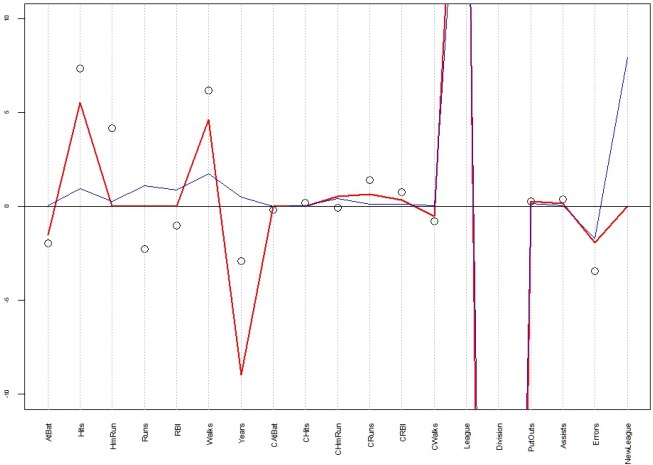
`AtBat` Number of times at bat in 1986
`Hits` Number of hits in 1986
`HmRun` Number of home runs in 1986
`Runs` Number of runs in 1986
`RBI` Number of runs batted in in 1986
`Walks` Number of walks in 1986
`Years` Number of years in the major leagues
`CAtBat` Number of times at bat during his career
`CHits` Number of hits during his career
`CHmRun` Number of home runs during his career
`CRuns` Number of runs during his career
`CRBI` Number of runs batted in during his career
`CWalks` Number of walks during his career
`League` A factor with levels `A` and `N` indicating player's league at the end of 1986
`Division` A factor with levels `E` and `W` indicating player's division at the end of 1986
`PutOuts` Number of put outs in 1986
`Assists` Number of assists in 1986
`Errors` Number of errors in 1986
`Salary` 1987 annual salary on opening day in thousands of dollars
`NewLeague` A factor with levels `A` and `N` indicating player's league at the beginning of 1987

Ejemplo: Hitters



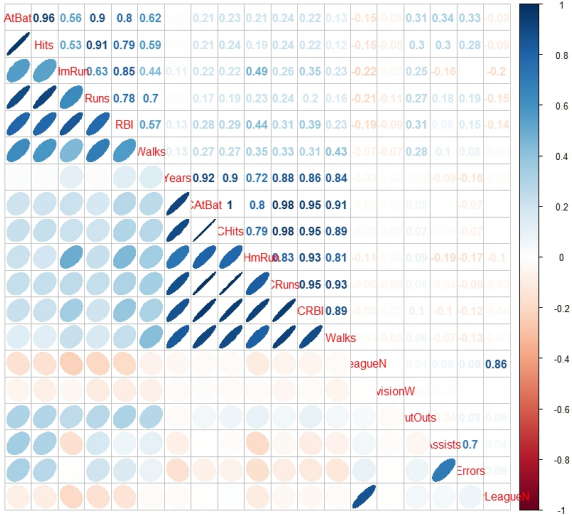
Puntos: OLS, rojo: LASSO, azul: ridge

Ejemplo: Hitters



Puntos: OLS, rojo: LASSO, azul: ridge

Ejemplo: Hitters



Apendice: LASSO y Ridge como optimizacion restringida

LASSO y Ridge como optimizacion restringida

Consideremos el problema de Ridge. Minimizar

$$R_r(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\ \beta^2$$

Llamemos $\hat{\beta}(\lambda)$ a la solucion, para un λ dado.

Resultado: existe $C \geq 0$ tal que $\hat{\beta}(\lambda)$ es la solucion a

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i\beta)^2 \text{ s.a. } \beta^2 \leq C$$

Recordar que las FOC para

$$R_r(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda\beta^2$$

es

$$-2\left(\sum(y_i - \beta x_i)x_i\right) + 2\lambda\beta = 0$$

y la solucion es

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

La funcion de Lagrange para el problema de optimizacion restringida

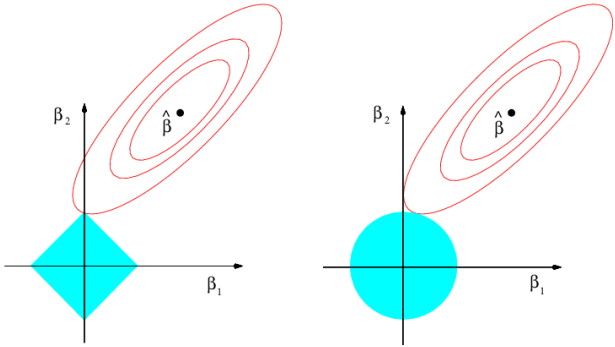
$$L(\beta, \alpha) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \alpha\ (\beta^2 - C)$$

CPO (Karush-Kuhn-Tucker)

$$\begin{aligned} -2\left(\sum_{i=1}^{n}(y_i - x_i\beta)x_i\right) + 2\alpha\beta &= 0 \\ \alpha(\beta^2 - C) &= 0 \end{aligned}$$

Notar que si $C = \hat{\beta}(\lambda)^2$, $\alpha = \lambda$ y $\hat{\beta}(\lambda)$ satisfacen las condiciones de KKT.

- Ridge puede ser visto como un problema de optimizacion restringida.
- Problema original: λ dado, restriccion 'determinada' por λ.
- Problema restringido: restriccion dada, multiplicador determinado por la restriccion.
- Dualidad.
- La formalizacion de LASSO es virtualmente identica (usando subgradientes)



Fuente: James, Witten, Hastie y Tibshirani (2013, pp. 222)