Elastic Net

Walter Sosa-Escudero

Universisad de San Andrés y CONICET

Prediccion y eleccion de modelo

- Accuracy: minimizar error de pronostico (ridge, LASSO)
- Dimensionalidad: reducir el espacio de los predictores (LASSO)

Problema adidicional: $p \geq n$.

Problema: $p \geq n$

- Rango? Maximo numero de filas o columnas linealmente independientes.
- Entonces: $\rho(X_{p,n}) \leq \text{mín}(p, n)$
- MCO: $\rho(X)$ implica $n \geq p$
- Si $\rho(X) = p$, entonces $\rho(X'X) = p$
- $\hat{\beta} = (X'X)^{-1}X'Y$. $(X'X)$ tiene que tener rango p. Si $p > n$, $\rho(X'X) \leq n < p$, $(X'X)$ no es invertible.

Practica: gene data set

Punto: Ridge y LASSO funcionan cuando $p \geq n$

Intuicion para ridge ($p > n$)

Intucion para ridge. Supongamos que los datos fueron previamente estandarizados (no intercepto):

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

Definamos p datos artificiales $(x_s', y_s), s = 1, \ldots, p$ de la siguiente forma:

- $x_s' = (0, \ldots, \sqrt{\lambda}, \ldots, 0)$, donde $\sqrt{\lambda}$ esta en la $s-$esima posicion.
- $y_s = 0$.

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

Si 'apilamos' los datos originales y los p nuevos en (x_i^r, y_i^r), matricialmente

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

Notar que X^r es una matriz $(n+p, p)$. Como $p \leq n+p$, entonces $(X^{r'} X^r)^{-1}$ es invertible, aun cuando $(X'X)$ no.

Intuicion: ridge es como que 'agrega' p puntos adicionales. Esto permite lidiar con el problema de $p \geq n$. Magia.

Problemas con LASSO

1. Cuando $p > n$ elige como maximo n variables.
2. Cuando un grupo de variables esta muy correlacionada, tiende a elegir una sola, arbitrariamente. Lo hace muy inestable para la prediccion. Ridge no tiene esta problema. Tecnicamente: no unicidad por convexidad no estricta de la penalidad LASSO.
3. Cuando $n > p$ y hay alta correlacion en los predictores, ridge tiende a funcionar mejor que LASSO en terminos de ECM.

Elastic net: predice bien, reduce dimensionalidad, elige bien grupos de variables.

Naive elastic net

$$R_r(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda_2 \sum_{s=1}^{p}(\beta_s)^2 + \lambda_1 \sum_{s=1}^{p}|\beta_s|$$

- Mezcla ridge y LASSO
- La parte LASSO elige predictores.
- La convexidad estricta de la penalidad (ridge) resuelve el problema de inestabilidad por agrupamiento.

Ejercicio: usando el argumento de 'data augmentation' de ridge, mostrar que naive elastic net es un LASSO 'con mas datos'.

Caso particular: $\hat{\beta}_{MCO>0}$, un solo predictor estandarizado:

$$\hat{\beta}_{nen} = \frac{\left(\hat{\beta}_{mco} - \lambda_1/2\right)_+}{1+\lambda_2},$$

en donde $(z)_+$ es la parte positiva de z, decir, $(z)_+$ es z si $z > 0$ y 0 en caso contrario.

Chequear que LASSO y Ridge aparecen como casos particulares.

Elastic net

$$\hat{\beta}_{en} = \frac{1}{\sqrt{1+\lambda_2}}\,\hat{\beta}_{nen}$$

- Version reescalada
- Intuicion: elimina el 'double shrinkage' de ridge (demasiado sesgo)
- Funciona mejor en la practica.

- Eleccion de hiperparametros (λ_1, λ_2): cross validation bidimensional
- Fuente: Zou, H. y Hastie, T., 2005, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society,* 67, 2, 301-320.