Complejidad, overfit y regularizacion

Walter Sosa Escudero
wsosa@udesa.edu.ar
Banco Central del Uruguay, 2020

Recordar

- $Y, \hat{Y}$
- $Y = f(X) + u$
- $\hat{Y} = \hat{f}(X)$
- $Err\left(Y - \hat{f}\right) = Sesgo^2(\hat{f}) + V(\hat{f}) + \sigma^2$

Complejidad

$$Y = f(X) + u$$

- X: regresores, rezagos, potencias, interacciones, etc.
- Complejidad: dimension de X, cantidad de predictores.
- Machine learning: bastante mas que una intuicion (generalizable)

Trade-off sesgo/varianza

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

- Omision de variables relevantes, inclusion de variables irrelevantes.
- *Trade off:* modelos mas 'complejos' tienden a ser menos sesgados pero con mayor varianza.
- Econometria: Preferencia 'lexicografica' por la insesgadez: minimizar ECM es minimizar varianza.
- Machine learning: tolerar predictores sesgados puede producir una caida sustancial en el la varianza, que produzcan una caida en ECM.

Complejidad y overfit

- Modelo verdadero: $Y = f(x^*) + u$, f en donde x^* incluye p^* variables.
- Estimamos modelos aumentando la cantidad de variables $p = 1, 2, \ldots$.
- Recordar: $Err(Y - \hat{f}) = Sesgo^2\left(f, \hat{f}\right) + V\left(\hat{f}\right) + \sigma^2$

Que pasa cuando aumentamos la complejidad del modelo?

- Sesgo?

- Varianza?: $\hat{f}(x) = x'\hat{\beta}$

$$V\left(\hat{f}(x)\right) = V(x'\hat{\beta}) = x'V(\hat{\beta})x = \sigma^2 x'(X'X)^{-1}x$$

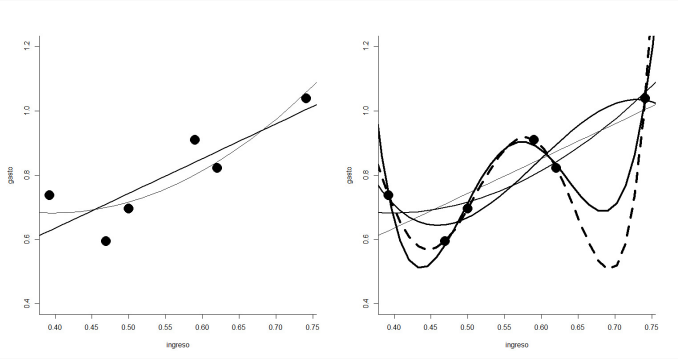
Promedio de las varianzas para todos los x_i :

$$\frac{1}{n}\sum_{i=1}^{n}\sigma^2 x_i'(X'X)^{-1}x_i = \sigma^2\,\frac{p}{n}$$

Overfit: mas alla de p^*, aumentar la complejidad no reduce el sesgo, mientras que la varianza aumenta monotonicamente, para σ^2 y n dados.

Overfit y prediccion fuera de la muestra

Overfit: 'sobreeliminar' el sesgo (problema 'menor' en econometria, central en ML).



- ML: prediccion *fuera* de la muestra (futura, condicional, contrafactica, etc.)
- Overfit: modelos extremadamente complejos predicen muy bien dentro de la muestra y muy mal fuera de ella.
- Elegir el nivel de complejidad optimo
- R^2 no funciona: mide prediccion dentro de la muestra, no decreciente en complejidad.

Como medir el error de pronostico fuera de la muestra?

Cross validation

- Prediccion fuera de la muestra
- Separar muestra de *entrenamiento* y de *test.*
- Estimar con la de entrenamiento y computar $Err(\hat{Y})$ en la muestra de test.
- Problema 1: perdida de eficiencia
- Problema 2: como partir la muestra?

K-fold cross validation

1. Partir los datos al azar en K partes.
2. Ajustar el modelo dejando afuera una de las particiones.
3. Computar el error de prediccion para los datos no utilizados.
4. Repetir para $k = 1, \ldots, K$.

Eleccion de K: 5 o 10 (insensible a la eleccion)

Error de prediccion por CV:

$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \hat{Y}_{-k}(x_i) \right)$$

$\hat{Y}_{-k}(x_i)$ prediccion cuando la observacion i no fue usada.

- Cada observacion es usada en dos roles: entrenamiento y test.
- $K = 1$: no test data
- $K = N$: 'leave one out'. Ir dejando de lado una obseracion por vez. Estima el modelo n veces con $n - 1$ datos.
- K: estima el modelo K veces con $n - K$ datos.

CV para eleccion de modelos

- α parametriza la complejidad de un modelo.
- Ejemplo: $\alpha =$ grado de polinomio.
- Cross validation para un modelo indizado por α:

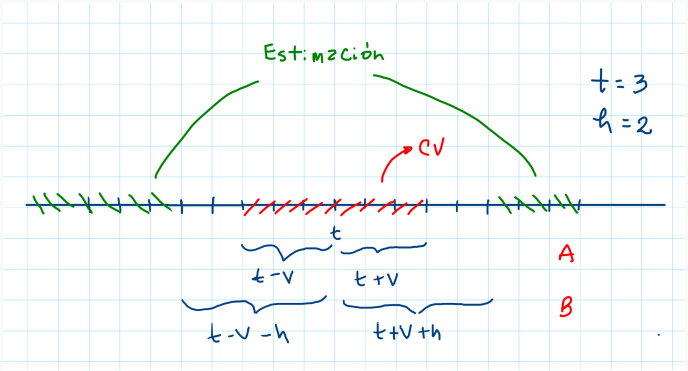
$$CV(\hat{f}, \alpha) = \frac{1}{N} \; L\left(Y_i - \hat{Y}_{-k}(x_i, \alpha)\right)$$

Idea: $CV(\hat{f}, \alpha)$ para una grilla de de α y minimizar.

Cross validation en series temporales

Problema: procesos dependientes, no se puede partir al azar. Racine (2000): 'hv cross validation'

- Datos: $t = 1, \ldots, T$
- Fijar dos numeros enteros, h y v
- Para un t en particular, fijar dos 'ventanas' alrededor de t: A) $(t - v, t + v)$ y B) $(t - v - h, t + v + h)$.
- Estimar el modelo usando los datos *fuera* de la ventana B.
- Computar el error de pronostico para los datos *dentro* de la ventana A.
- Repetir para todos los t



- h rompe la dependencia
- Racine: $h = 0{,}25T$, $v \sim \sqrt{T}$
- Ver Racine (2000) y Elliott y Timmerman (2016)

Seleccion de modelos

- Overfit: elegir el modelo de complejidad optima, el que minimiza ECM fuera de la muestra (cross validation).
- Estrategia obvia: busqueda exhaustiva. Estimar todos los modelos. Elegir el de menor ECM de acuerdo a CV.
- Problema: p predictores, 2^p modelos Ejemplo: $p = 20$ habria que estimar 1.048.576 modelos (comun en series temporales). Y...agregar CV!

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

- Como elegir la complejidad (tamaño) de un modelo sin tener que estimarlos todos?
- Reduccion de dimensionalidad (minimo modelo).
- Mejora en la capacidad predictiva.

Regularizar: introducir informacion ajena al modelo para 'disciplinar' sus complejidad.

LASSO

Para $\lambda \geq 0$ dado:

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

(el primer coeficiente corresponde al intercepto).

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

- $\sum_{i=1}^{n}(y_i - x_i'\beta)^2$ 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: en un paso:

- *Reduce dimensionalidad:* elige que variables entran ($\beta_s \neq 0$) y cuales no ($\beta_s = 0$). Solucion de esquina.
- *Mejora error de prediccion:* En general predice mejor que MCO.

$$R_l(b) = SRC(b) + \lambda|b|$$

- LASSO: poner variables solo si son suficientemente relevantes.
- En general, $\hat{b}_l = 0$ para variables irrelevantes, y $\hat{b}_l$ esta 'corrido hacia cero' para las relevantes.
- Shrinkage: estimacion *sesgada* por la regularizacion.
- Regularizar: 'disciplinar' el modelo hacia la hipotesis nula de no significatividad.
- Por que? Eliminar variables induce sesgo pero puede bajar dramaticamente la varianza, mejora ECM.

Ridge

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

Idea: intuicion similar pero ridge NO elimina variables. Por que?

Caso simple, una variable estandarizada:

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

FOC:

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

Despejando

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

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

Ridge vs. MCO

MCO:

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

Ridge:

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

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

- Si $\lambda < 2\sigma^2/\beta^2$, $ECM(\hat{\beta}) - ECM(\hat{\beta}_r) > 0$
- Para todo β y σ^2 existe λ de modo que ridge le 'gana' a MCO.
- Idea importante: sesgar la estimacion para reducir varianza.

En este caso es posible derivar una condicion necesaria y suficiente. Pero no es generalizable al caso de p variables (Theobald, 1974).

Regularizar

- Trade off sesgo-varianza: modelos mas complejos son menos sesgados pero mas volatiles.
- Shrinkage: LASSO 'sesga' los coeficientes a cero.
- Regularizar: controlar el comportamiento de parametros/variables con elementos de afuera del modelo (λ). λ es un hiperparametro.
- Aprendizaje: eleccion de un modelo optimo a traves de la eleccion de un hiperparametro.
- Eleccion de λ optimo? CV. Sin CV validation. Belloni y Chernozhukov (2011):

$$\lambda_T = 2c\sigma \sqrt{T}\, \Phi^{-1}(1 - \alpha/2p)$$

$\Phi(.)$ es la fda normal, σ^2 es la varianza de los residuos. Sugieren usar $\alpha = 0{,}1$ y $c = 1{,}1$.

Elastic net

$p > n$

- MCO: no funciona!
- LASSO y ridge... si!

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

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

Notar que X^r es una matriz $(n + p, p)$. Como $p \leq n + p$, entonces $(X^{r\prime} X^r)^{-1}$ 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 $p > n$ 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.

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.