Overfit, cross validation y bootstrap

Walter Sosa-Escudero

Universisad de San Andrés y CONICET

Cuestiones preliminares

Ley de grandes numeros

Sea z_n una sucesion de variables aleatorias escalares. Consideremos la siguiente sucesion

$$\bar{z}_n = \frac{\sum_{i=1}^{n} z_i}{n}$$

Ley de grandes numeros (Kolmogorov): $\{z_n\}$ iid con $E(z_i) = \mu$ finita. Entonces $\bar{z}_n \xrightarrow{p} \mu$.

- Funcion de distribucion acumulada (FDA): Y una variable aleatoria.

$$F(y) = Pr(Y \leq y)$$

- Funcion de distribucion acumulada empirica: $Y_1, Y_2, \ldots, Y_n$ una muestra aleatoria iid de la poblacion Y, con fda $F(y)$.

$$F_n(y) \equiv \frac{1}{n} \sum_{i=1}^{n} 1 [Y_i \leq y]$$

Teorema Fundamental de la Estadistica

Para cualquier y dado ($y = y_0$)

$$F_n(y_0) \equiv \frac{1}{n}\sum_{i=1}^{n} 1\left[Y_i \leq y_0\right]$$

es un promedio. Entonces por la LGN:

$$F_n(y_0) \xrightarrow{p} E\left(1\left[Y \leq y_0\right]\right) = P(Y \leq y_0) = F(y_0)$$

Teorema Fundamental de la Estadistica (Glivenko-Cantelli): version fuerte y uniforme de $F_n(y) \to F(y)$.

Complejidad y overfit

El trade-off sesgo/varianza

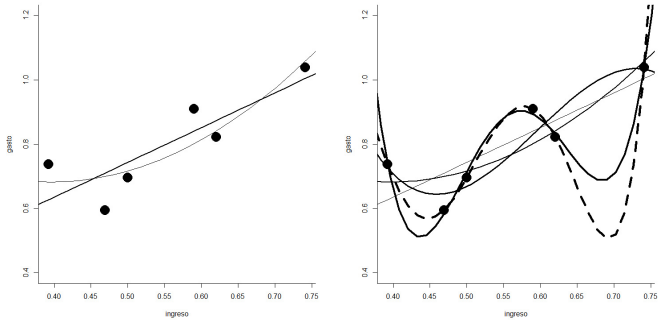
Trade off-sesgo varianza en econometria:

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

- Omision de variables relevantes, inclusion de variables irrelevantes.
- *Trade off:* modelos mas 'complejos' tienden a ser menos sesgados pero con mayor varianza.
- Complejidad: numero de variables?.
- Preferencia 'lexicografica' por la insesgadez: minimizar ECM es minimizar varianza.
- Desafio: tolerar sesgos para bajar considerablemente la varianza.

Overfit

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



Fuente: WSE, 2019, *Big data*, Siglo XXI Editores, capitulo 5

Caso: polinomio

- Modelo verdadero: $Y = f(x_0) + u$, f es un polinomio de grado finito p^*, pero desconocido.
- Estimamos polinomios con grado creciente $p = 1, 2, \ldots$.
- Recordar: $Err(Y - \hat{f}) = \sigma^2 + Sesgo^2\left(f, \hat{f}\right) + V\left(\hat{f}\right)$

Que pasa cuando aumenta el grado del polinomio estimado?

- Sesgo?
- Varianza?: $\hat{f}(x_0) = \sum_{s=0}^{p} x_0^s \, \hat{\beta}_s \equiv x_0^{'}\hat{\beta}, \qquad x_0' \equiv (1, x_0, x_0^2, \ldots, x_0^p)$

$$V\left(\hat{f}(x_o)\right) = V(x_0^{'}\hat{\beta}) = x_0^{'}V(\hat{\beta})x_0 = \sigma^2 x_0^{'}(X'X)^{-1}x_0$$

Promedio de las varianzas para todos los x_i :

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

Resultado (overfit): a partir de p^*, aumentar la complejidad no reduce el sesgo, mientras que la varianza aumenta monotonicamente, para σ^2 y n dados.

Proof: $A_{m\times m}$ with typical A_{ij}. The trace of A ($tr(A)$) is the sum of the elements of its diagonal: $tr(A) \equiv \sum_{s=1}^{m} A_{ii}$.

Properties:

1. For any square matrices A, B and C: $tr(A+B) = tr(A) + (B)$
2. The cyclic property: $tr(ABC) = tr(BCA) = tr(CAB)$
3. A rather trivial property is that if $m = 1$, $tr(A) = A$.

The fitted model for the case of the p degree polynomial is

$$\hat{Y}_i = \sum_{s=o}^{p} \hat{\beta}_s X_{ip}^p = x_i \hat{\beta}$$

with $x_i \equiv (1, X_i, X_i^2, \ldots X_i^p)$. Then $V\left(\hat{f}(X_i)\right) = V\left(x_i' \hat{\beta}\right) = \sigma^2 x_i'(X'X)^{-1} x_i$. Now:

$$\text{Average } V\left(x_i' \hat{\beta}\right) = \frac{1}{n} \sum_{i=1}^{n} \sigma^2 x_i'(X'X)^{-1} x_i$$

Now we invoke traces. Note that $x_i'(X'X)^{-1}x_i$ is a scalar and hence equal to its trace, so

$$\text{Average } V\left(x_i'\hat{\beta}\right) = \frac{\sigma^2}{n}\sum_{i=1}^{n} tr\left(x_i'(X'X)^{-1}x_i\right)$$

Now, using the 'cyclic' property, $tr\left(x_i'(X'X)^{-1}x_i\right) = tr\left((X'X)^{-1}x_ix_i'\right)$, and using the first property of traces, we get

$$\sum_{i=1}^{n} tr\left((X'X)^{-1}x_ix_i'\right) = tr\left(\sum_{i=1}^{n}(X'X)^{-1}x_ix_i'\right) = tr\left((X'X)^{-1}(X'X)\right) = p.$$

Replacing above we get the final result.

Cross validation

Overfit y prediccion fuera de la muestra

- 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
- Como medir el error de pronostico *fuera* de la muestra?
- R^2 no funciona: mide prediccion dentro de la muestra, no decreciente en complejidad.

Conceptos

- Perdida: $L(Y, \hat{Y})$
  - Regresion: $L(Y, \hat{Y}) = (Y - \hat{Y})^2$
  - Clasificacion: $L(Y, \hat{Y}) = 1(Y \neq \hat{Y})$
- Error de prediccion esperado: $\text{Err} = E[L(Y, \hat{Y})]$.
- Error de prediccion esperado en la muestra de test:

$$\text{Err}_{\mathcal{T}} = E[L(Y, \hat{Y})|\mathcal{T}]$$

- Error promedio en la muestra de entrenamiento:

$$\hat{\text{err}} = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \hat{Y}_i)$$

Problema: como elegir $\mathcal{T}$?

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 $k = 1, \ldots, K$.

La estimacion por cross-validation del error de prediccion es

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

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

- 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.

Por qué funciona? Ley de los grandes numeros.

$$
\begin{aligned}
CV(\hat{f}) &= \frac{1}{N} \sum_{i=1}^{n} L\left(Y_i - \hat{Y}_{-k}(x_i)\right) \\
&= \frac{1}{K} \sum_{j=1}^{K} \hat{\text{err}}_j
\end{aligned}
$$

- Computamos $\hat{\text{err}}_j$ para cada particion y luego promediamos.
- Para cualquier K finito, $n \to \infty$ implica que dentro de cada sumando hay infinita informacion.
- $K = n$ la suma es infinita.

Eleccion de K

- K chico: maximiza datos para estimar, sensible a los valore particulares.
- K grande: maximiza datos para evaluar, modelo estimado menos precisamente.
- Regla: 5 o 10 (ver Kohavi (1995)

Cross validation para eleccion de modelos

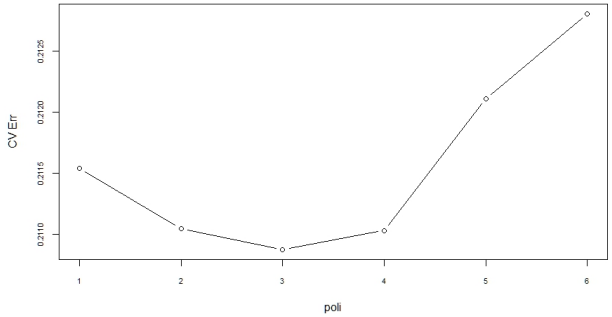
- Supongamos que α parametriza la complejidad de un modelo.
- Ejemplo: $\alpha =$ grado de polinomio en nuestro caso inicial
- Cross validation para un modelo indizado por α:

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

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

Ejemplo: error de CV para distintos grados de polinomio, elegir el que minimiza $CV(\hat{f}, \alpha)$

Ejemplo: eleccion de grado de un polinomio



Vinos: Vinos malos. Potencias de volatile.acidity. Metodo de estimacion: logit. K=10-fold Cross validation.

Bootstrap

Bootstrap

- $Y_1, Y_2, \ldots, Y_n$ iid $Y \sim (\mu, \sigma^2)$, ambas finitas.
- Queremos estimar $V(\bar{Y}) = \sigma^2/n$ (varianza de la media muestral)
- Formula: $\hat{\sigma}^2/n$ con

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}(Y_i - \bar{Y})^2$$

Metodo alternativo 'sin formula':

1. De los n datos originales $y_1, y_2, \ldots, y_n$, tomar una muestra *con reemplazo*, de tamaño n.
2. Computar la media muestral con esta 'pseudomuestra'
3. Repetir B veces. Al finalizar tendremos B estimaciones de la media.
4. Computar la varianza de las B medias.

En terminos generales

$Y_i, i = 1, \ldots, n$ y θ es una magnitud de interes.

1. Muestra de tamaño n con reemplazo de la mueestra original (muestra *bootstrap*).
2. Computar $\hat{\theta}_j$, $j = 1, \ldots, B$.
3. Repetir B veces.
4. $$\hat{V}(\hat{\theta})_B = \frac{1}{B} \sum_{j=1}^{B} (\hat{\theta}_j - \bar{\hat{\theta}})^2$$

Ideas

- Por que?: en la mayoria de los casos NO hay una formula para la varianza.
- Mucho mas que la varianza: desvio estandar, mediana, coeficiente de Gini,....
- Intervalo de confianza: percentiles empiricos de las B estimaciones bootstrap.
- Regresion?

$$\ln Y = \beta X + u$$

X es una variable binaria. Recordar que $\beta \sim e^{\beta} - 1$. Supongamos que realmente nos interesa $e^{\beta} - 1$.

Podemos estimar β por MCO de $\ln Y$ en X y su varianza por boostrap:

1. Tomar una muestra de $(X_i, Y_i) i = 1, \ldots, n$ de tamaño n con reemplazo.
2. Estimar $\hat{\beta}_j$
3. Computar $\hat{\theta}_j = e^{\hat{\beta}_j)} - 1, j = 1, \ldots, B$
4. Computar la varianza muestral de los $\hat{\theta}_j$

Por que funciona?

- 'Muestra de la muestra'
- n es grande: es como si estuviesemos tomando una muestra de la poblacion.
- Teorema fundamental de la estadistica (estamos reemplazando $F(y)$ por $F_n(y)$.

Ejemplo: desigualdad

Table 1

Gini coefficients and bootstrapped confidence intervals and standard errors

Greater Buenos Aires and Neuquén, 1992-1997

Region	Year	Gini	Confidence Interval				Standard	Coef. Of
			0.025	0.05	0.95	0.975	Error	Variation
GBA	92	0.4415	0.4264	0.4285	0.4529	0.4548	0.0072	1.6%
	93	0.4430	0.4298	0.4308	0.4520	0.4529	0.0062	1.4%
	94	0.4570	0.4418	0.4440	0.4691	0.4708	0.0077	1.7%
	95	0.4843	0.4687	0.4704	0.4955	0.4971	0.0074	1.5%
	96	0.4840	0.4705	0.4721	0.4956	0.4982	0.0073	1.5%
	97	0.4797	0.4670	0.4686	0.4938	0.4958	0.0074	1.6%
Neuquén	92	0.4632	0.4393	0.4423	0.4798	0.4830	0.0113	2.4%
	93	0.4422	0.4251	0.4266	0.4608	0.4619	0.0109	2.5%
	94	0.4574	0.4374	0.4406	0.4759	0.4801	0.0110	2.4%
	95	0.4813	0.4534	0.4582	0.5047	0.5087	0.0148	3.1%
	96	0.4973	0.4608	0.4651	0.5297	0.5357	0.0195	3.9%
	97	0.4718	0.4458	0.4482	0.4910	0.4929	0.0137	2.9%

Source: Authors' calculations based on the EPH, October 1992-1997.

Fuente: Sosa Escudero y Gasparini (2001). Cada año calculamos Gini, intervalo de confianza, error estandar y coeficiente de variacion usando bootstrap.