EST-25134: Aprendizaje Estadístico

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Remuestreo.
Objetivo: Estudiaremos estrategias para cuantificar el error de generalización utilizando la información disponible. Revisaremos conceptos separación de muestras, validación cruzada y Bootstrap.
Lectura recomendada: Capítulo 5 de [1].

1. INTRODUCCIÓN

Estudiaremos métodos de `remuestreo`: validación cruzada y *bootstrap* como mecanismos computacionales para obtener información adicional sobre el modelo ajustado.

Hasta ahora, hemos considerado $\mathcal{D}_n$ y $\mathcal{T}_m$ como la separación de datos en dos conjuntos *provenientes* de la misma población. Nos permiten cuantificar:

- El **error de entrenamiento** $\mathcal{L}(\mathcal{D}_n)$.
- El **error de prueba** o `generalización` $\mathcal{L}(\mathcal{T}_m|\mathcal{D}_n)$.

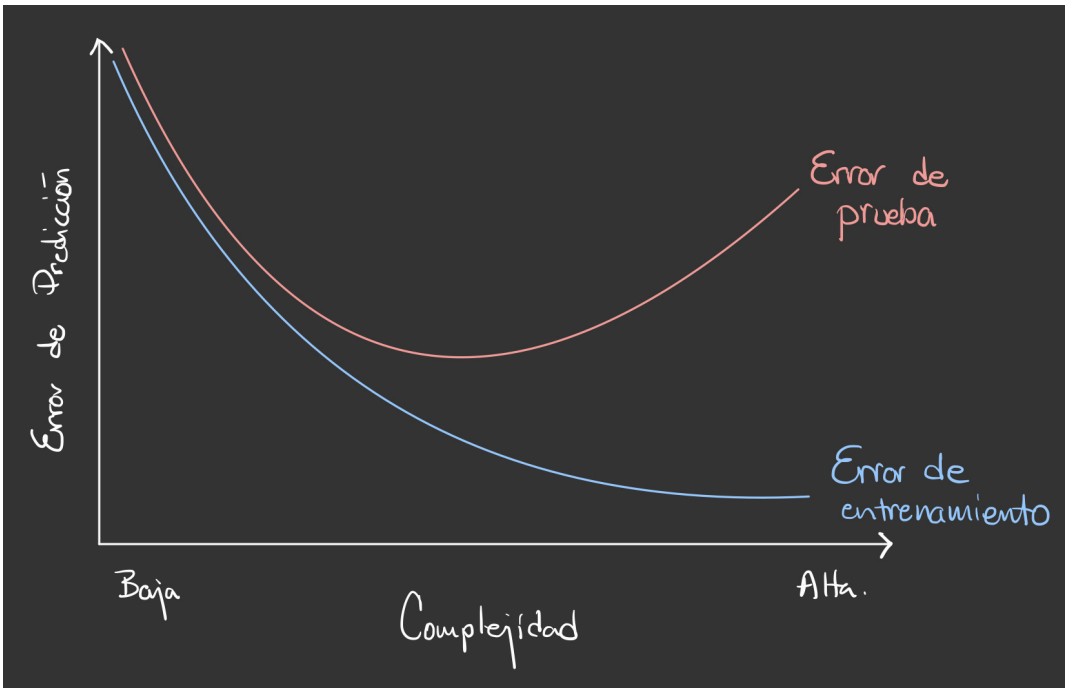

FIGURA 1. *Comportamiento típico del error de generalización y entrenamiento.*

1.1. ¿Cuál es la mejor manera para estimar de manera confiable el error de generalización?

(LGN) Un conjunto de datos grande que permita cuantificar *bien* el error.

1.2. ¿Qué alternativas tenemos?

1. Utilizar métodos que hacen correcciones al sesgo por utilizar el conjunto de datos para generalizar.
2. Utilizar un conjunto `fuera de entrenamiento`.

2. UTILIZANDO UN CONJUNTO DE VALIDACIÓN

Utilizamos una partición de nuestro conjunto de datos: `datos de entrenamiento` y `datos de validación`.

> El error cuantificado en el conjunto de validación nos da un estimador del error de prueba. Típicamente utilizamos el mismo criterio de error con el que entrenamos (RMSE, precisión, etc.).

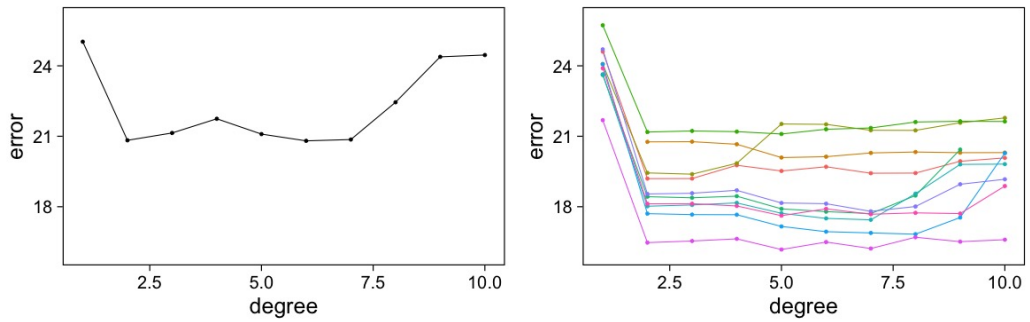

FIGURA 2. *Error de validación utilizando una partición del conjunto de datos (izquierda). Error de validación con distintas particiones de los datos (derecha).*

> Nota que las estimaciones son muy variables. Son altamente sensibles al conjunto de entrenamiento y validación que se utillizaron. Podríamos estar **sobre-estimando** el error de validación pues dejamos de lado un conjunto de datos para entrenar el modelo.

3. VALIDACIÓN CRUZADA

Es una técnica que nos permite elegir la *mejor* configuración de un modelo y nos da indicios del error de *generalización*. La idea es:

1. Dividir el conjunto de datos en K bloques.
2. Utilizar un método iterativo para ajustar modelos con $K-1$ bloques y registrar el error de ajuste con el bloque fuera del entrenamiento.

3.1. Detalles

Sean K bloques y utilicemos $C_1, C_2, \ldots C_K$ para denotar con C_k el conjunto de índices en el bloque k. En total tenemos n_k observaciones en cada bloque. El caso particular es $n_k = n/K$.

Calculamos

$$\mathsf{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \mathsf{MSE}_k \,. \tag{1}$$

3.2. Caso especial: LOO-CV

Si utilizamos $K = n$, entonces tenemos lo que se conoce como *leave-one out cross-validation* (LOO-CV).

En el caso de estimadores lineales por mínimos cuadrados (como regresión polinomial) tenemos un *atajo* para calcular con *un sólo ajuste*

$$\mathsf{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2 . \tag{2}$$

Donde h_i es el estadístico de `anclaje` de la observación i.

La definición de este estadístico lo puedes encontrar en el Capítulo 3 de [1] en la página 99,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x}_n)^2}{\sum_{j=1}^{n}(x_j - \bar{x}_n)^2} . \tag{3}$$

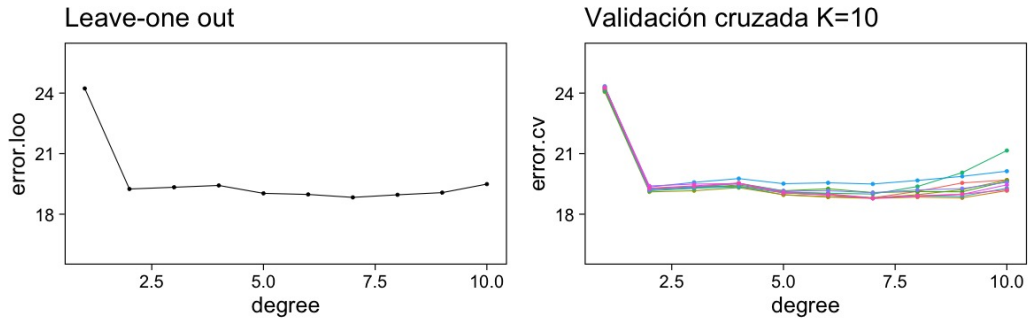

FIGURA 3. *Métricas de error bajo validación cruzada.*

3.3. Observaciones

Utilizamos conjuntos de datos mas pequeños para entrenar. Por lo tanto tenemos un sesgo en el error mas grande de los que hubiéramos querido.

El sesgo se *puede eliminar* al tomar $K = n$ pero tiene una *gran varianza*.

En la práctica un *buen compromiso* se puede establecer con $K = 5$ ó 10 (experimentación empírica).

3.4. ¿Y para clasificación?

3.5. Un caso para pensar

Consideremos que tenemos un conjunto de datos con pocas muestras y muchos atributos, $p \gg n$. Para ajustar un modelo lo que hacemos es:

1. Encontrar los $p = 20$ predictores con mayor correlación con la respuesta.
2. Utilizar validación cruzada para entrenar un modelo con esos $p = 20$ predictores y cuantificar su error de generalización.

¿Está bien esta estrategia?

4. BOOTSTRAP

Es una técnica de remuestreo que nos permite cuantificar incertidumbre sobre un *estimador* o un *procedimiento de estimación*.

Lo usamos muchas veces para estimar el *error estándar* de un estimador o poder reportar intervalos de confianza basados en percentiles. (No utilizamos supuestos asintóticos).

Si pudiéramos generar muestras de la población no tendríamos problemas. Pero en muchas ocasiones no tenemos acceso al generador de datos.

Resolvemos estos problemas tomando `re-muestras` de las observaciones que tenemos utilizando `muestreo aleatorio con reemplazo`.

De esta manera, creamos conjuntos de datos ficticios (a partir de los datos observados) que nos permiten estimar las cantidades de interés. Con un número suficiente de réplicas podemos obtener una distribución de estimadores de la cual podemos extraer percentiles para construir un intervalo de confianza.

4.1. Observaciones

Muchas veces hay que tener cuidado con la forma en que generamos las remuestras. Por ejemplo, en situaciones con datos temporales o geográficos.

4.2. Cuantificando el error de generalización

En validación cruzada los bloques no tienen traslape. Esto es ventajoso para cuantificar el error y su variación.

Si utilizáramos *bootstrap* entonces los bloques ocasionarían problemas con los estimadores.

REFERENCIAS

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, New York, NY, 2021. ISBN 978-1-07-161417-4 978-1-07-161418-1. . 1, 3