Clase 3: Actividad - Módulo 2

Mauricio Alejandro Mazo Lopera

Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín

Métodos de remuestreo

DATOS DE ENTRENAMIENTO

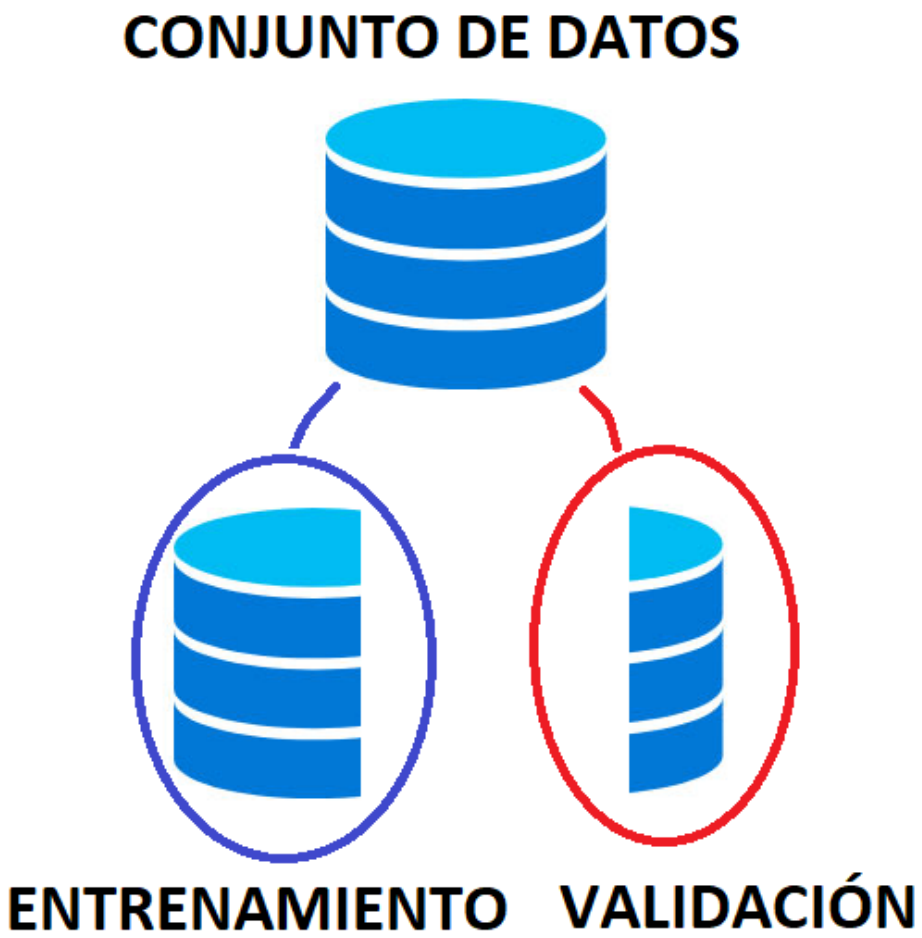
¿ES MEJOR AJUSTAR UN ÚNICO MODELO?

Ó

SUBMUESTRAS

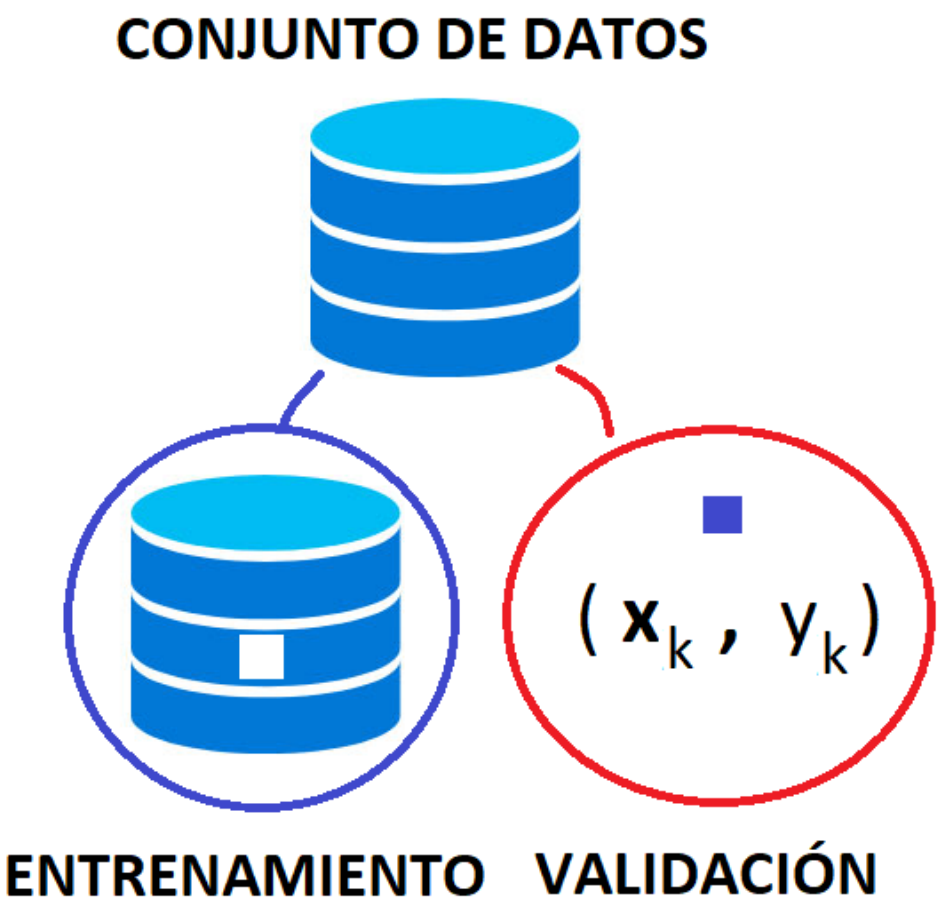
¿AJUSTAR VARIOS MODELOS CON SUBMUESTRAS?

1. Método del conjunto de validación:



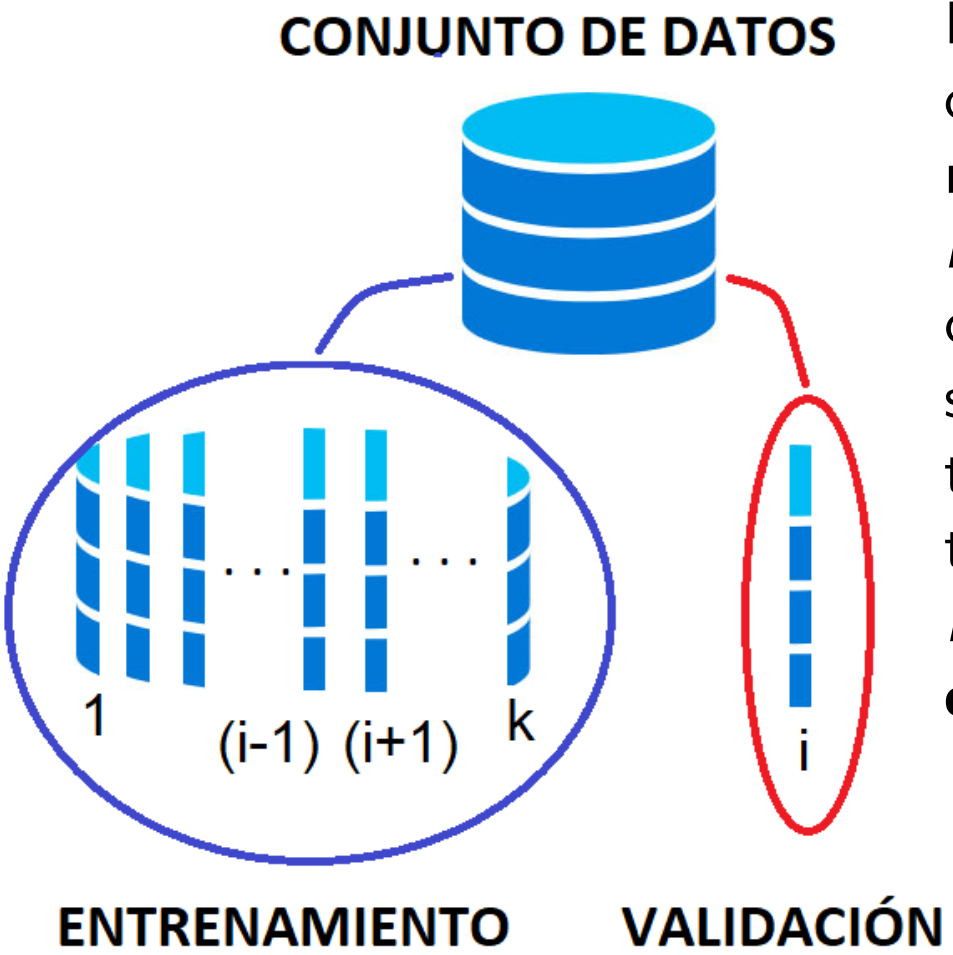
Como no siempre se cuenta con datos de prueba (*set test*) se parte la base de datos en dos subbases. Una se usa para ajustar el modelo (**datos de entrenamiento**) y la otra para validar el modelo (**datos de validación**). La tasa de error obtenida con el conjunto de validación (usualmente se usa el MSE cuando la variable respuesta es cuantitativa) es un estimador de la tasa de error de prueba, en inglés *test error rate*.

2. Validación cruzada dejando uno afuera:



Este tipo de validación que en inglés se escribe 'Leave-One-Out Cross-Validation' (LOOCV) consiste en dividir el conjunto de datos en dos subconjuntos: uno con un único dato (conjunto de prueba) y otro con todos los datos restantes (conjunto de entrenamiento). Este proceso se repite n veces, quitando en cada caso el k-ésimo dato y encontrando el $MSE_k = (y_k - \widehat{y}_k)^2$, para $k = 1, 2, \ldots, n$.

3. Validación cruzada de k pliegues (k-fold)



Los datos se dividen en k grupos o pliegues aproximadamente del mismo tamaño. Se selecciona el i-ésimo grupo ($i = 1, 2, \ldots, k$) como conjunto de validación y se entrena el modelo con los datos de los $(k-1)$ grupos restantes. Luego obtenemos el i-ésimo $MSE_{(k)i}$ y estimamos la **tasa de error de prueba** con:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_{(k)i}$$

En la práctica se suele usar $k = 5$ o $k = 10$.

Compensación Sesgo-Varianza:

CJTO. DE VALIDACIÓN ⇐ + SESGO, + VARIANZA, + RÁPIDO

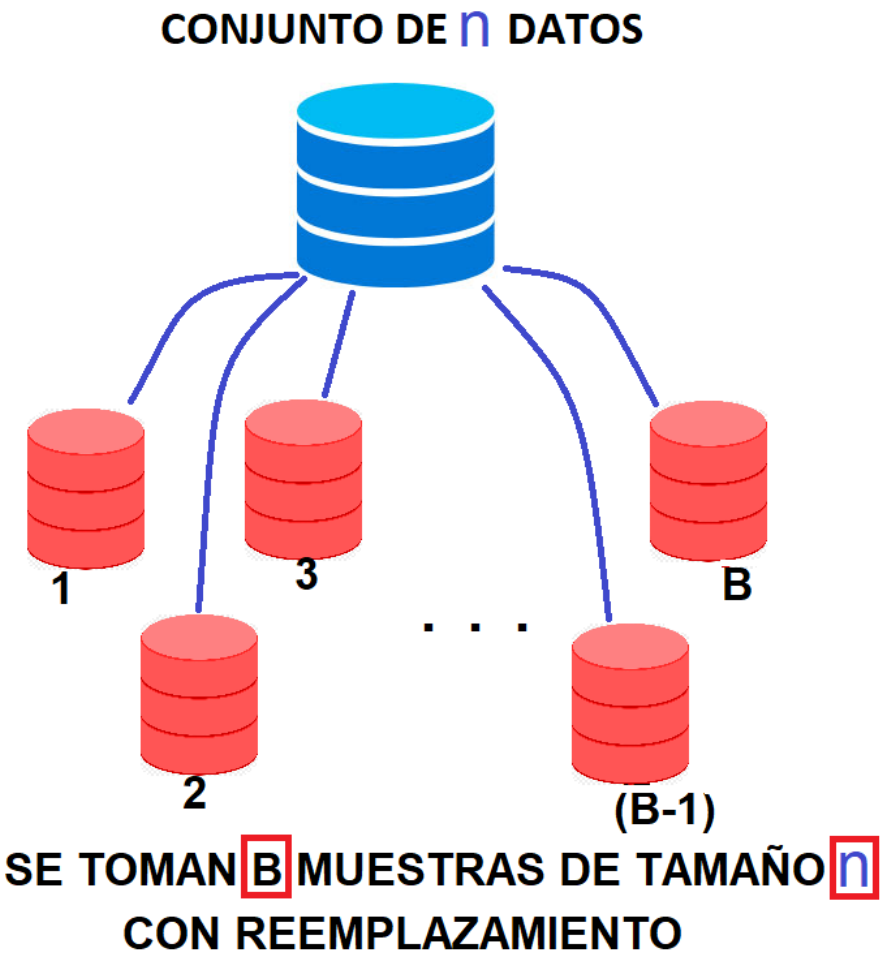
VERSUS

LOOCV ⇐ - SESGO, + VARIANZA, - RÁPIDO

VERSUS

***k*-fold** ⇐ + SESGO, - VARIANZA, ± RÁPIDO

El Bootstrap:



Para una base de datos de n individuos, se hace un submuestreo con reemplazamiento para obtener B muestras cada una de tamaño n. En este caso, la misma observación puede estar dos o más veces en cada submuestra.

Actividad:

1. Simule una base de datos considerando el modelo lineal $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$, teniendo en cuenta las siguientes condiciones:

   - Y: variable numérica dependiente.
   - X_1, X_2, X_3: variables numéricas independientes.
   - X_4: variable categórica independiente con tres categorías.
   - $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^\top = (1, 0.3, 0.6, -1, 1.5, -2)^\top$
   - $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, donde $\sigma = 2$.
   - $n = 500$.

Actividad:

a. Compare el modelo con todas las variables versus los modelos individuales (con una sola variable) utilizando los métodos de validación cruzada:

   - Conjunto de validación.
   - LOOCV.
   - k- fold.

b. Utilizando Bootstrap, obtenga intervalos de confianza para σ, para R^2 y para R^2-Ajustado.

Actividad:

c. Repita el item (b) pero simulando errores con distribución t-student con 9 grados de libertad. En este caso, ¿cuál es la varianza teórica esperada para los errores?

d. Simule de nuevo el modelo del item (a) pero ahora incluyendo un término cúbico para X_1 con coeficiente $\beta_{13} = 2$. Utilizando validación cruzada varíe, desde 1 hasta 10, el grado del polinomio que depende de X_1 y realice un gráfico de los MSE en función de dichos grados. ¿Coinciden los resultados con lo que se esperaba?