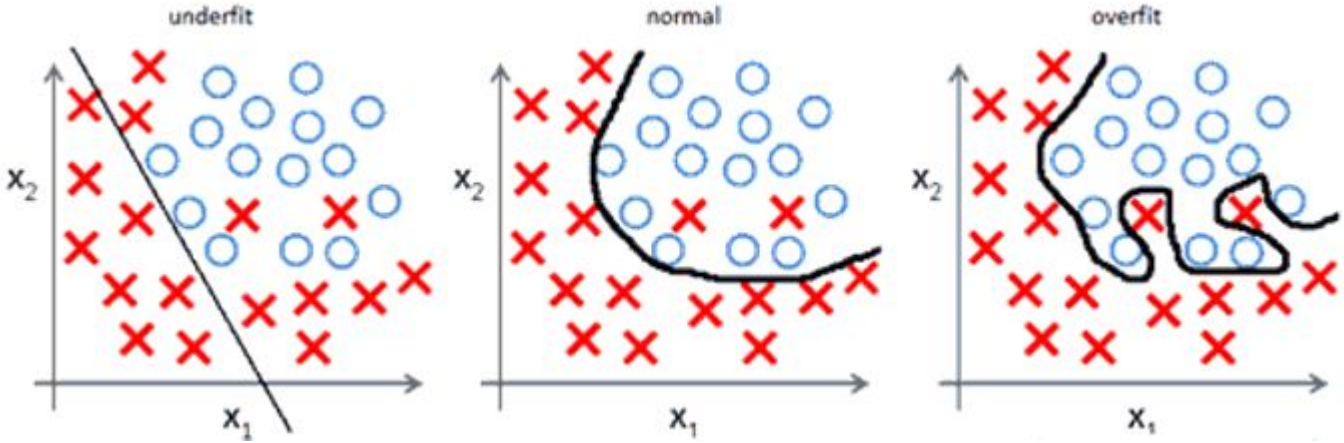
Selección del modelo (Model selection)

William Esteban Gómez
Estudiante de doctorado

¿Qué es la selección del modelo?

Los modelos muy complejos (muchos parámetros) tienden a memorizar la estructura de los datos, lo cual es llamado sobreajuste del modelo (en inglés, overfitting). Mientras que los modelos muy sencillos, tienden a realizar un ajuste pobre de la estructura de los datos (sub ajuste, ó en inglés, underfitting).

Entonces cómo escoger el mejor modelo para nuestros datos?



¿Cómo obtenemos un modelo general?

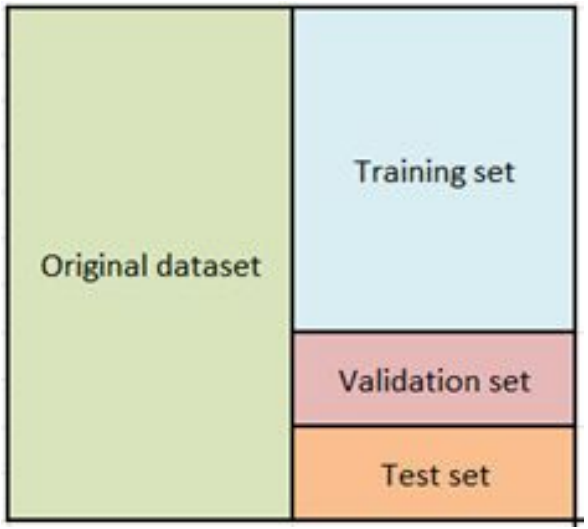
Aquí es donde la palabra **Validación** toma importancia. El procedimiento de validación intenta estimar cómo funcionará nuestro modelo con nuevos datos.

Los datos deben separarse en tres conjuntos:

Entrenamiento: Datos que permiten entrenar el modelo.

Validación: Datos que permiten evaluar el rendimiento del modelo.

Prueba: Datos que nos permiten estimar la eficiencia del modelo con datos futuros.



¿Cómo hacer la división de los datos?

Las dos metodologías más usadas para hacer la validación son:

Boostrapping:

En esta metodología sólo se dividen los datos en un conjunto de entrenamiento aleatorio que tiene el 70% (80%) de los datos y se prueba con el 30% (20%) restante.

Entrenamiento	Prueba

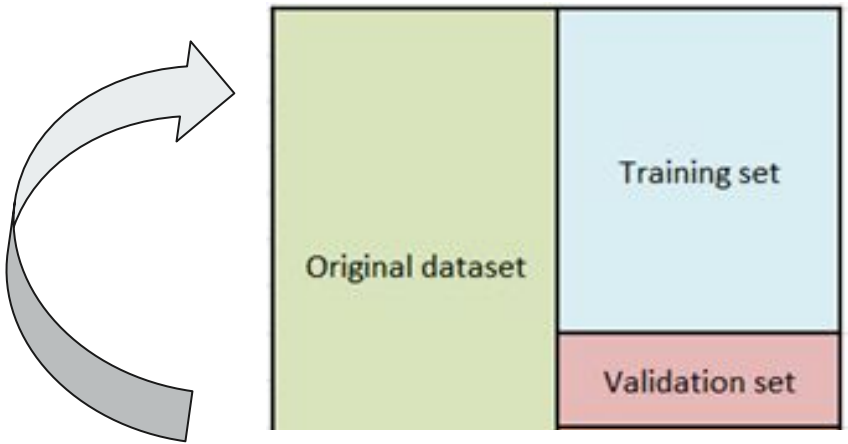
Validación cruzada (Cross-validation):

También llamada n fold cross validation. Y qué es un fold? Es sólo una división aleatoria de los datos.

Conjunto de datos dividido en 6 folds

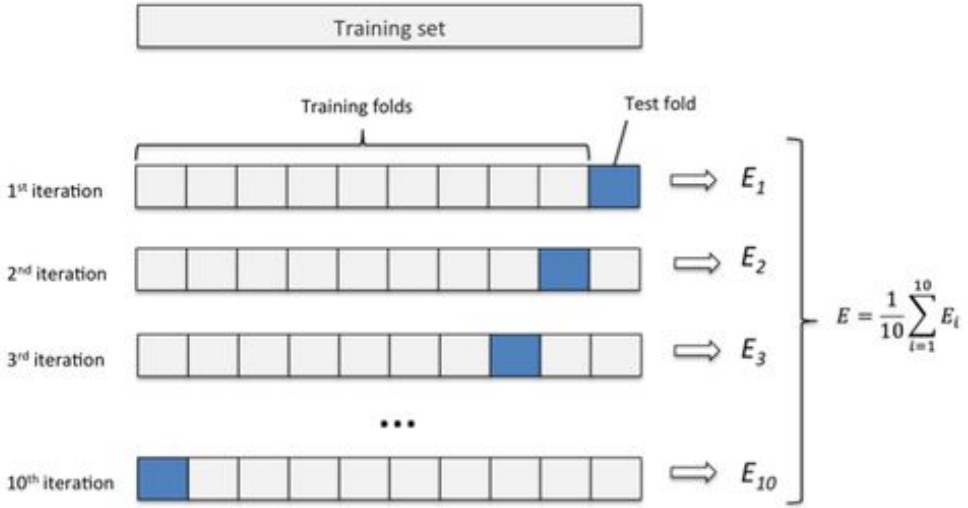
Bootstrap:

También llamada **Remuestreo** debido a que su enfoque es muestrear del conjunto de datos un conjunto para entrenamiento y uno para prueba. Este muestreo se hace de manera aleatoria y se repite muchas veces. Al final se hace un promedio de los errores de prueba para obtener el estimado del error de prueba.



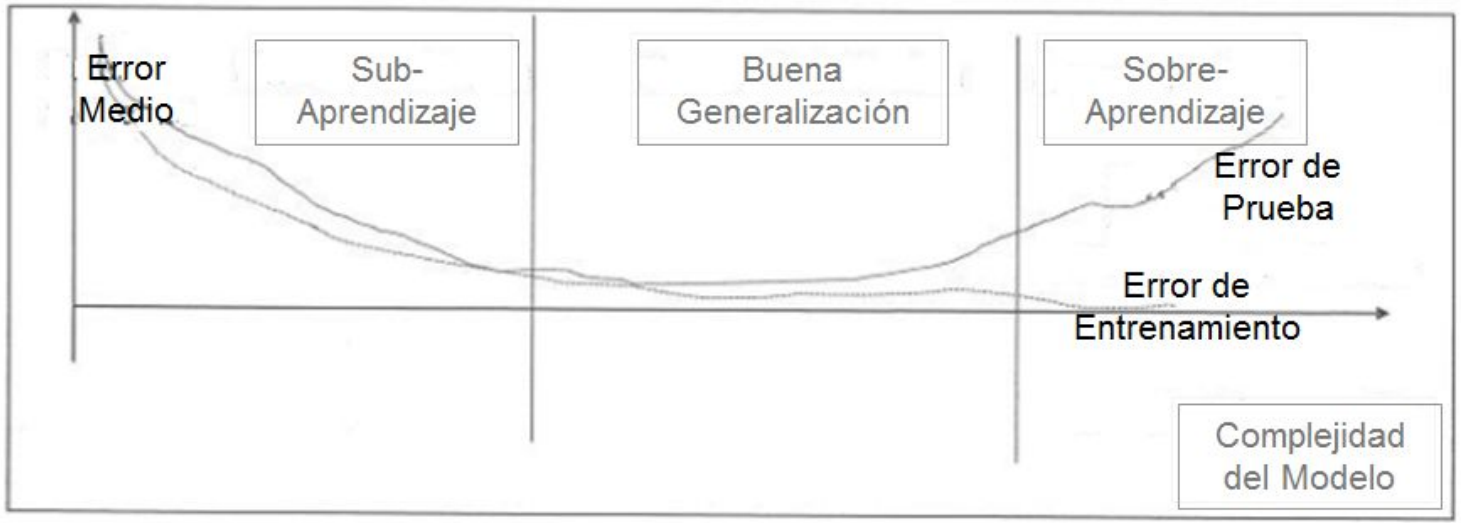
Validación cruzada:

Por ejemplo, si es una validación cruzada 10 folds, el procedimiento es realizar 10 iteraciones donde en cada iteración 9 folds son utilizados para entrenar el modelo y 1 para la prueba. Al final el promedio de las iteraciones es el estimado del error de prueba.



Un modelo potente siempre podrá memorizar...

Un modelo lo suficientemente complejo puede ajustarse a cualquier tipo de datos, lo que se debe hacer es saber cuando parar de entrenar.



Procedimiento general para abordar un problema de machine learning

–Identificación y extracción de descriptores.

–División de los datos.

–Selección de caracteristicas

–Método de aprendizaje

–Validación