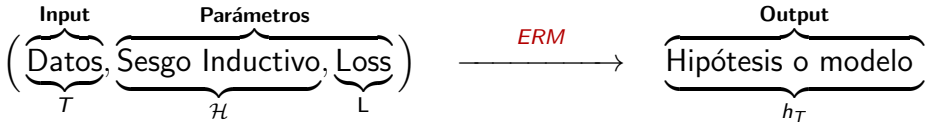
Machine Learning para Inteligencia Artificial

Selección y Validación

Universidad ORT Uruguay

9 de Abril, 2025

Recordar: algoritmo de ML



- Un algoritmo de ML busca la mejor hipótesis en el sesgo inductivo $\mathcal{H}$.
- $J_{\mathcal{D}}(h) = \boldsymbol{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\Big[\text{Loss}(h(\boldsymbol{x}), y)\Big]$; $\quad J_T(h) = \boldsymbol{E}_{(\boldsymbol{x},y)\in T}\Big[\text{Loss}(h(\boldsymbol{x}), y)\Big]$
- Idealmente buscamos h_{opt} que minimiza el Costo Verdadero $J_{\mathcal{D}}$.
- En la práctica buscamos h_T que minimiza el Costo Empírico (ERM) J_T.

Generalización y sobreajuste

- El margen de generalización de una hipótesis h es la diferencia

$$G(h) = J_{\mathcal{D}}(h) - J_T(h)$$

- Intuición: h_T no generaliza bien si el margen de generalización es "grande".
- Una posible causa de mala generalización es el **sobreajuste**:

  h sobreajusta los datos de entrenamiento T si existe h' tal que

$$J_T(h') > J_T(h) \quad \text{y} \quad J_{\mathcal{D}}(h') < J_{\mathcal{D}}(h)$$

  Notar que en dicho caso $G(h') < G(h)$.

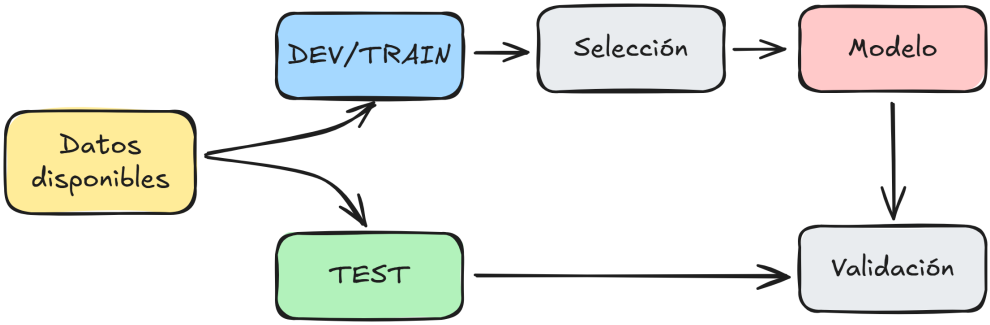
Selección y Validación

Selección

Proceso de elegir el sesgo inductivo con mejor rendimiento para una tarea específica, según criterios de rendimiento y otros factores como la complejidad.

Validación

Proceso de evaluar el rendimiento en datos no vistos de un modelo entrenado para garantizar que una generalización adecuada.


DEV/TRAIN
Selección
Modelo
Datos disponibles
TEST
Validación

Validación

- Problema: no sabemos calcular $J_{\mathcal{D}}(h_T)$.
- Solución: tomar una muestra V **independiente** de T, y calcular

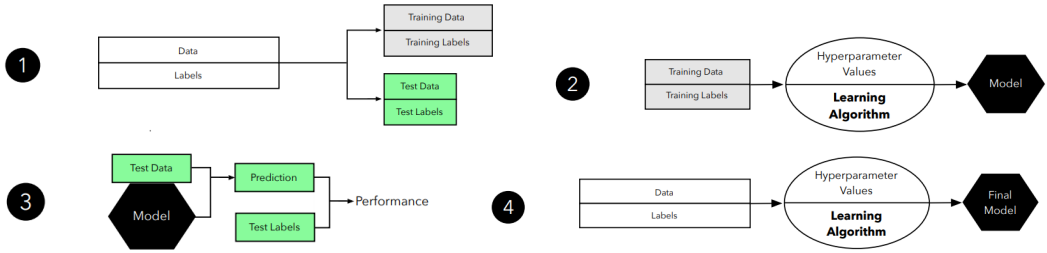
$$J_V(h_T) = \frac{1}{|V|} \sum_{(\boldsymbol{x},y)\in V} \text{Loss}\left(h_T(\boldsymbol{x}), y\right)$$

- Nuestra estimación será por tanto

$$\text{Estimación de } G(h_T) = J_V(h_T) - J_T(h_T)$$

- **Holdout** - En la práctica se particionan *al azar* los datos disponibles:
  - Train (70 %-80 %-90 %): entrada del algoritmo de aprendizaje.
  - Test (o Validation) (30 %-20 %-10 %): evaluar el margen de generalización.

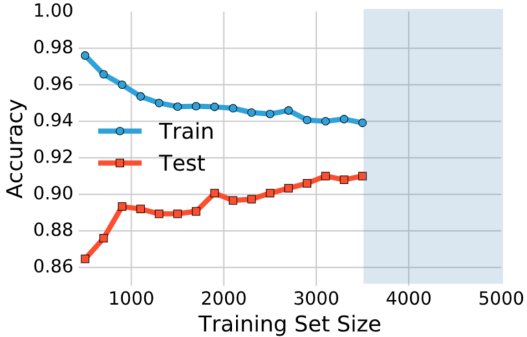
Diagrama de la técnica de Holdout



Fuente: Curso de ML de Sebastian Raschka

El algoritmo definitivo se entrena usando el dataset entero.

Pessimistic bias: Holdout y su dependencia con N



Fuente: Curso de ML de Sebastian Raschka

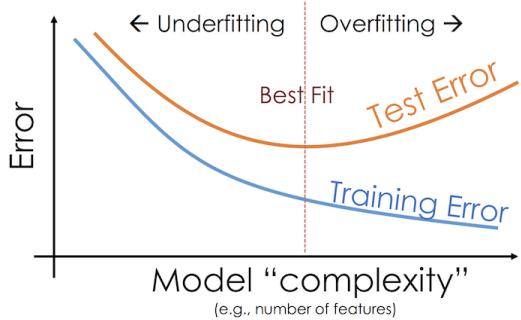
- Algoritmo de clasificación.
- 500 imágenes aleatorias de cada una de las diez clases del MNIST.
- Las 5000 imágenes se dividieron 70 % train y 30 % test con estratificación.
- Train se dividió en subconjuntos aún más pequeños, y estos subconjuntos se usaron para ajustar el clasificador.
- La curva se usa para evaluar el margen de generalización.

G tiende a disminuir con N, se produce un sesgo pesimista en la estimación.

Selección

- Es tentador usar Holdout para hacer model selection.
- Cuidado: optimistic bias.
- Para estimar correctamente el error de generalización del best fit particionar

$$\text{Datos} = \underbrace{\text{Train} + \text{Val}}_{\text{Dev}} + \text{Test}$$



Fuente: Principles and Techniques of Data Science

Costo verdadero **esperado** en $\mathcal{H}$

- El output h_T del algoritmo de ML depende del dataset $T \sim \mathcal{D}^N$.

- El costo verdadero **esperado** en $\mathcal{H}$ es

$$J_{\mathcal{D}}(\mathcal{H}) = \underset{T \sim \mathcal{D}^N}{\boldsymbol{E}} [J_{\mathcal{D}}(h_T)]$$

Selección con Cross-Validation

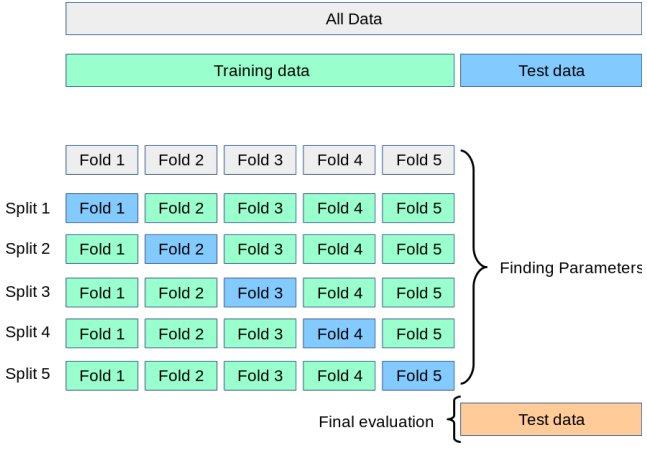
- Particionar *Train* en k (5-10) *folds* $F_1 \dots F_k$ de igual tamaño
- Para todo $i = 1 \dots k$
  - Usar $T_i = \cup_{j \neq i} F_j$ (todos salvo F_i) como entrada para el algoritmo

$$h_i = \arg\min_{h \in \mathcal{H}} \{J_{T_i}(h)\}$$

  - Evaluar el costo de la hipótesis obtenida h_i en F_i: $J_i = J_{F_i}(h_i)$
- Calcular el promedio de los costos

$$J_{CV}(\mathcal{H}) = \frac{1}{k} \sum_{i=1}^{k} J_i$$

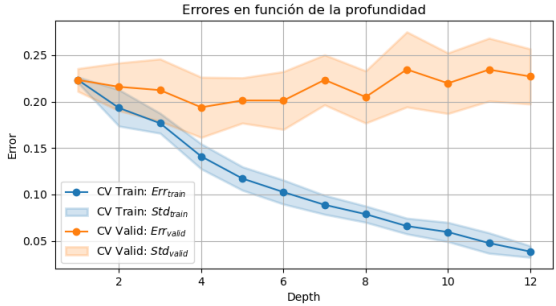
Selección con Cross-Validation



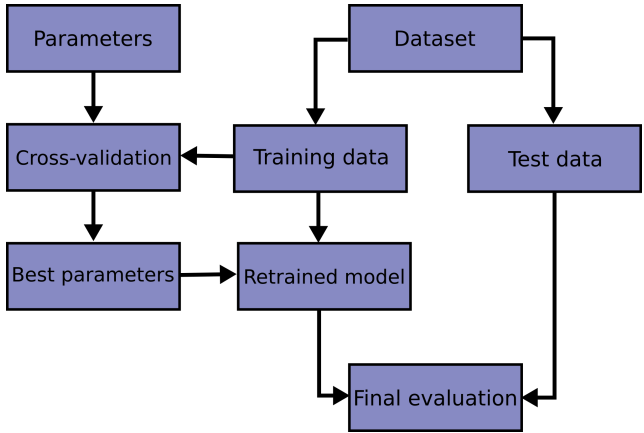
Fuente: sklearn User Guide Capítulo 3

Selección con Cross-Validation

- $J_{CV}(\mathcal{H})$ es una estimación de $J_{\mathcal{D}}(\mathcal{H})$
- Se usa para comparar sesgos inductivos: $\mathcal{H}$ y $\mathcal{H}'$
  - Aplicar validación cruzada con $\mathcal{H}$ y $\mathcal{H}'$
  - Comparar $J_{CV}(\mathcal{H})$ vs $J_{CV}(\mathcal{H}')$

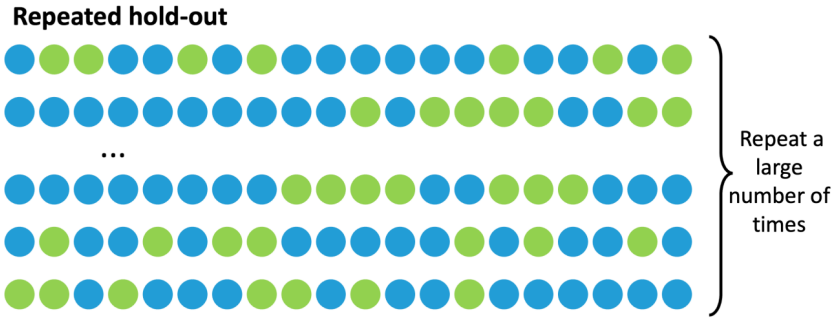


Selección con Cross-Validation



Fuente: sklearn User Guide Capítulo 3

Selección con Repeated Holdout



Fuente: Evaluating machine learning models and their diagnostic value

Bibliografía

- An introduction to statistical learning with applications in Python. Capítulo 5.1.

- Machine Learning - A First Course for Engineers and Scientists. Capítulo 4.