Machine Learning para Inteligencia Artificial

Descomposición del Error

Universidad ORT Uruguay

30 de Abril, 2025

La predicción óptima

- Loss cuadrática: $\text{Loss}(\widehat{y}, y) = (\widehat{y} - y)^2$
- La relación $\boldsymbol{x} \rightsquigarrow y$ estocástica

$$y = f(\boldsymbol{x}) + \epsilon, \quad \boldsymbol{E}\left[\epsilon | \boldsymbol{x}\right] = 0$$

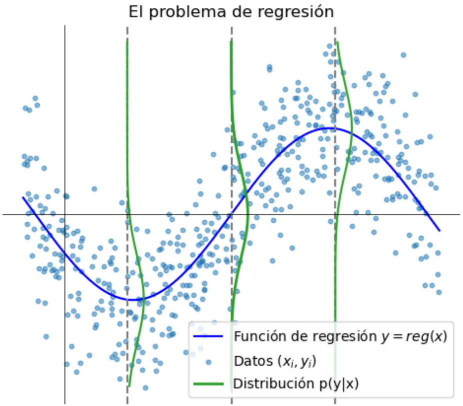
- $f : \mathcal{X} \to \mathcal{Y}$ es la función a aprender

$$f(\boldsymbol{x}) = \text{reg}(\boldsymbol{x}) = \boldsymbol{E}\left[y \mid \boldsymbol{x}\right]$$

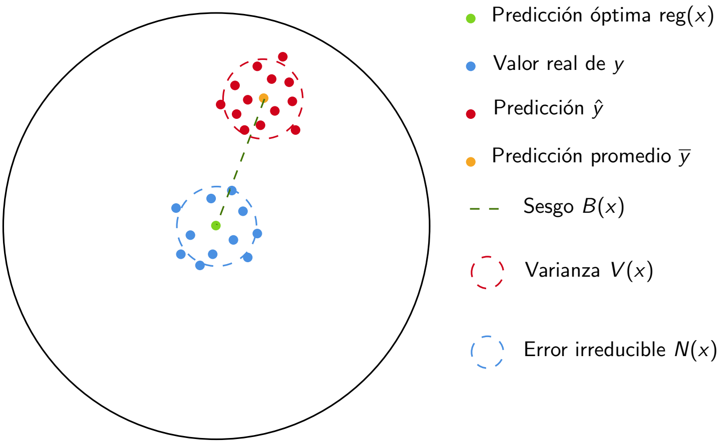
- Los datos son una muestra iid de la forma:

$$y_i = f\left(\boldsymbol{x_i}\right) + \epsilon_i$$

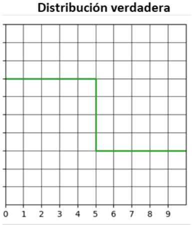
para $i = 1, \ldots, N$

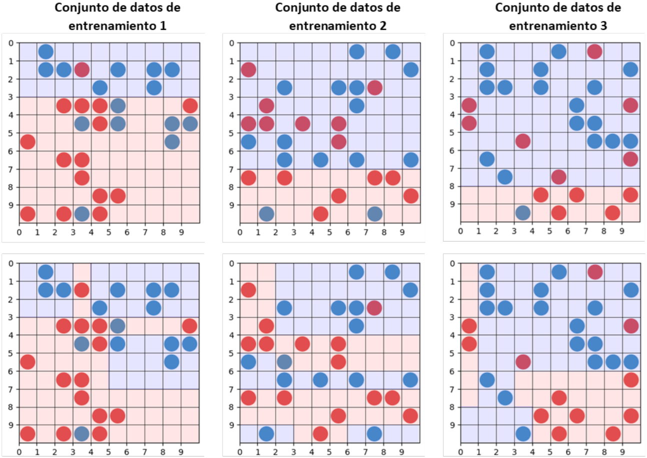


Idea intuitiva de sesgo y varianza



Sesgo y varianza: dependencia con $\mathcal{H}$





Sesgo Inductivo:
Árbol de profundidad 1

Baja varianza: la hipótesis varía poco con los datos

Alto sesgo: el modelo se ajusta peor a los datos

Sesgo Inductivo:
Árbol de profundidad 4

Alta varianza: la hipótesis varía mucho con los datos

Bajo sesgo: el modelo se ajusta bien a los datos

Error cuadrático medio (MSE)

- Un algoritmo A busca la mejor hipótesis en el sesgo inductivo $\mathcal{H}$.
- Su output h_T depende del dataset $T \sim \mathcal{D}^N$.
- Fijado un $\boldsymbol{x} \in \mathcal{X}$ el riesgo esperado (MSE) en $\boldsymbol{x}$ es

$$\boxed{\mathrm{MSE}(\boldsymbol{x}) = \underset{T \sim D^N}{\boldsymbol{E}} \left\{ \underset{y}{\boldsymbol{E}} \left[\left(h_T(\boldsymbol{x}) - y\right)^2 \,\middle|\, \boldsymbol{x} \right] \right\}}$$

Sesgo

Fijado un $\boldsymbol{x} \in \mathcal{X}$:

- $\bar{h}(\boldsymbol{x}) = \boldsymbol{E}_{T \sim D^N}\left[h_T(\boldsymbol{x})\right]$ la predicción promedio del algoritmo en $\boldsymbol{x}$
- El sesgo en $\boldsymbol{x}$ se define como

$$\boxed{\mathrm{B}(\boldsymbol{x}) = \left(\bar{h}(\boldsymbol{x}) - \mathrm{reg}(\boldsymbol{x})\right)^2}$$

Varianza

Fijado un $\boldsymbol{x} \in \mathcal{X}$, la varianza en $\boldsymbol{x}$ es:

$$\boxed{\mathrm{V}(\boldsymbol{x}) = \underset{T \sim D^N}{\boldsymbol{E}} \left[\left(h_T(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}) \right)^2 \right]}$$

Error Irreducible

Fijado un $\boldsymbol{x} \in \mathcal{X}$, el error irreducible (o ruido) en $\boldsymbol{x}$ es:

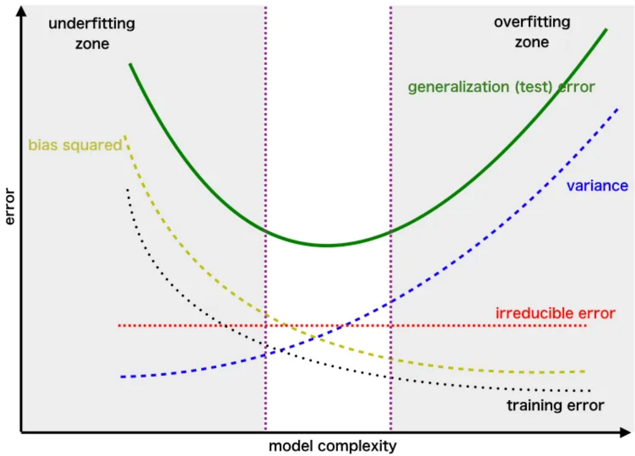
$$\boxed{\mathrm{N}(\boldsymbol{x}) = \underset{y}{\boldsymbol{E}}\left[(y - \mathrm{reg}(\boldsymbol{x}))^2 \;\middle|\; \boldsymbol{x}\right] = \mathrm{var}(y \mid \boldsymbol{x})}$$

Descomposición del error

$$\underbrace{\text{Error esperado}}_{\text{MSE}(\boldsymbol{x})} = \underbrace{\text{Sesgo}}_{\text{B}(\boldsymbol{x})} + \underbrace{\text{Varianza}}_{\text{V}(\boldsymbol{x})} + \underbrace{\text{Error Irreducible}}_{\text{N}(\boldsymbol{x})}$$

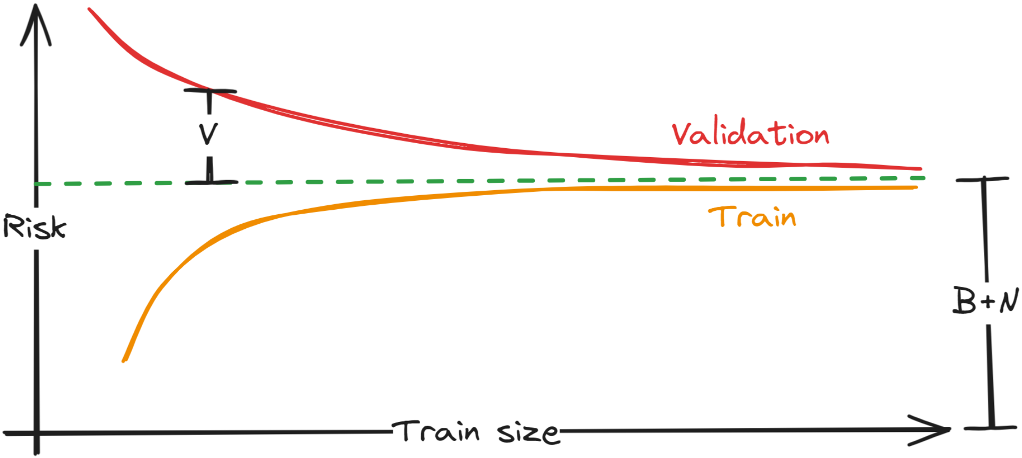
- El sesgo tiende a bajar cuando aumenta la *complejidad* (o *flexibilidad*)
- La varianza tiende a aumentar cuando aumenta la *complejidad*
- El error irreducible es la varianza de $y \mid \boldsymbol{x}$

Descomposición del error



Fuente: Blog TDS de Giorgos Papachristoudis

Descomposición del error - Learning curve



Opcional: Demostración de la descomposición del error

- Fijamos un $\boldsymbol{x} \in \mathcal{X}$. La MSE en ese $\boldsymbol{x}$ es

$$\mathrm{MSE}(\boldsymbol{x}) = \underset{T \sim D^N}{\boldsymbol{E}} \left\{ \underset{y}{\boldsymbol{E}} \left[\left(h_T(\boldsymbol{x}) - y\right)^2 \,\middle|\, \boldsymbol{x} \right] \right\}$$

- Sumamos y restamos la función de regresión $\mathrm{reg}(\boldsymbol{x})$:

$$\left(h_T(\boldsymbol{x}) - y\right)^2 = \left(h_T(\boldsymbol{x}) - \mathrm{reg}(\boldsymbol{x}) + \mathrm{reg}(\boldsymbol{x}) - y\right)^2$$

$$= \begin{matrix} \left(h_T(\boldsymbol{x}) - \mathrm{reg}(\boldsymbol{x})\right)^2 & (B + V) \\ + & \\ \left(\mathrm{reg}(\boldsymbol{x}) - y\right)^2 & (N) \\ + & \\ 2\left(h_T(\boldsymbol{x}) - \mathrm{reg}(\boldsymbol{x})\right)\left(\mathrm{reg}(\boldsymbol{x}) - y\right) & (0) \end{matrix}$$

Opcional: Demostración de la descomposición del error

- El término (N) al promediar en y da lugar a:

$$\underset{y}{\boldsymbol{E}}\left[(\mathrm{reg}(\boldsymbol{x}) - y)^2 \,\middle|\, \boldsymbol{x}\right] = \mathrm{N}(\boldsymbol{x})$$

- El término (0) al promediar en y desaparece:

$$\underbrace{2\,(h_T(\boldsymbol{x}) - \mathrm{reg}(\boldsymbol{x}))}_{\text{No depende de } y}\,\underbrace{\underset{y}{\boldsymbol{E}}\left[(\mathrm{reg}(\boldsymbol{x}) - y) \mid \boldsymbol{x}\right]}_{=0 \text{ por def. de } \mathrm{reg}(\boldsymbol{x})} = 0$$

Opcional: Demostración de la descomposición del error

- Al término $(B+V)$ le sumamos y restamos la predicción promedio $\bar{h}(\boldsymbol{x})$:

$$\left(h_T(\boldsymbol{x}) - \mathrm{reg}(\boldsymbol{x})\right)^2 = \left(h_T(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}) + \bar{h}(\boldsymbol{x}) - \mathrm{reg}(\boldsymbol{x})\right)^2$$

$$= \begin{array}{cr} \left(h_T(\boldsymbol{x}) - \bar{h}(\boldsymbol{x})\right)^2 & (V) \\ + & \\ \left(\bar{h}(\boldsymbol{x}) - \mathrm{reg}(\boldsymbol{x})\right)^2 & (B) \\ + & \\ 2\left(h_T(\boldsymbol{x}) - \bar{h}(\boldsymbol{x})\right)\left(\bar{h}(\boldsymbol{x}) - \mathrm{reg}(\boldsymbol{x})\right) & (0) \end{array}$$

Opcional: Demostración de la descomposición del error

- El término (V) al promediar en T da lugar a:

$$\underset{T\sim D^N}{\boldsymbol{E}}\left[\left(h_T(\boldsymbol{x})-\bar{h}(\boldsymbol{x})\right)^2 \,\middle|\, \boldsymbol{x}\right]=\mathrm{V}(\boldsymbol{x})$$

- El término (0) al promediar en T desaparece:

$$\underbrace{2\left(\bar{h}(\boldsymbol{x})-\mathrm{reg}(\boldsymbol{x})\right)}_{\text{No depende de } T}\underbrace{\underset{y}{\boldsymbol{E}}\left[\left(h_T(\boldsymbol{x})-\bar{h}(\boldsymbol{x})\right) \,\middle|\, \boldsymbol{x}\right]}_{=0 \text{ por def. de } \bar{h}(\boldsymbol{x})}=0$$

- El término (B) es $\mathrm{B}(\boldsymbol{x})$.

Bibliografía

- An introduction to statistical learning with applications in Python. Capítulo 5.1.
- Machine Learning - A First Course for Engineers and Scientists. Capítulo 4.