Sesgo y varianza en regresión

Matías Carrasco

11 de septiembre de 2023

Consideremos el siguiente setting:

- Espacio de atributos $\mathcal{X} \subset \mathbb{R}$.
- Espacio de etiquetas $\mathcal{Y} \subset \mathbb{R}$.
- Espacio de observaciones $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.
- Distribución desconocida $\mathcal{D}$ en $\mathcal{Z}$.
- Conjunto de datos $S = \{(x_i, y_i)\}_{i=1}^{N} \sim \mathcal{D}^N$.
- Función de pérdida cuadrática: $\text{Loss}(y, y') = (y - y')^2$

Definición (Función de regresión)**.** Fijado un $x \in \mathcal{X}$, la predicción **óptima** en x es

$$r(x) = \arg\min_{\widehat{y} \in \mathcal{Y}} \Big\{ \mathbb{E}\left[\text{Loss}(\widehat{y}, y) \mid x\right] \Big\}$$

Como la pérdida es cuadrática, se puede probar que $r(x) = \mathbb{E}\left[y \mid x\right]$. La función $r : \mathcal{X} \to \mathcal{Y}$ se llama **función de regresión**.

Considerando la función de regresión podemos escribir la relación estocástica entre x e y de la siguiente forma:

$$y = r(x) + \epsilon$$

en donde ϵ es una variable aleatoria que cumple $\mathbb{E}[\epsilon \mid x] = 0$. De esta forma podemos escribir también $y_i = r(x_i) + \epsilon_i$ para el conjunto de datos S.

Consideremos ahora un **algoritmo** que dado un conjunto de datos S nos devuelve un **modelo** $\widehat{y} = f_S(x)$ que de algún modo intenta aproximar $r(x)$.

Definición (Error cuadrático medio)**.** Fijado un $x \in \mathcal{X}$, el **MSE** en x se define como

$$\text{MSE}(x) = \mathbb{E}_{S \sim D^N}\left\{\mathbb{E}\left[\text{Loss}\left(f_S(x), y\right) \mid x\right]\right\} = \mathbb{E}_{S \sim D^N}\left\{\mathbb{E}\left[\left(f_S(x) - y\right)^2 \mid x\right]\right\}$$

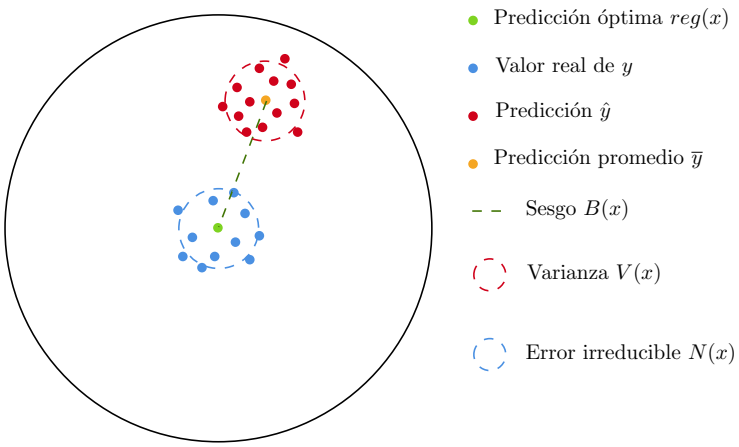

Figura 1: Descomposición del error en sesgo, varianza y error irreducible.

Vamos a definir el sesgo y la varianza para dicho algoritmo, ver Fig. 1.

Definición (Sesgo)**.** Fijado un $x \in \mathcal{X}$, sea $\bar{f}(x) = \mathbb{E}_{S\sim D^N}[f_S(x)]$ la **predicción** promedio del algoritmo en x. El **sesgo** en x se define como

$$\mathrm{B}(x) = \mathrm{Loss}\left(\bar{f}(x), r(x)\right) = \left(\bar{f}(x) - r(x)\right)^2$$

Definición (Varianza)**.** Fijado un $x \in \mathcal{X}$, la **varianza** en x se define como

$$\mathrm{V}(x) = \mathbb{E}_{S\sim D^N}\left[\mathrm{Loss}\left(f_S(x), \bar{f}(x)\right)\right] = \mathbb{E}_{S\sim D^N}\left[\left(f_S(x) - \bar{f}(x)\right)^2\right]$$

Definición (Error irreducible)**.** Fijado un $x \in \mathcal{X}$, el **error irreducible** (o ruido) en x se define como

$$\mathrm{N}(x) = \mathbb{E}\left[\mathrm{Loss}\left(y, r(x)\right) \mid x\right] = \mathbb{E}\left[\left(y - r(x)\right)^2 \mid x\right] = \mathrm{var}(y \mid x)$$

Teorema (Descomposición del error)**.** Fijado $x \in \mathcal{X}$:

$$\mathrm{MSE}(x) = \mathrm{B}(x) + \mathrm{V}(x) + \mathrm{N}(x)$$

Demostración. Para calcular el MSE debemos promediar las diferencias cuadráticas $(f_S(x) - y)^2$:

$$\begin{aligned}(f_S(x) - y)^2 &= (f_S(x) - (r(x) + \epsilon))^2 \\ &= (f_S(x) - r(x))^2 - 2(f_S(x) - r(x))\epsilon + \epsilon^2\end{aligned}$$

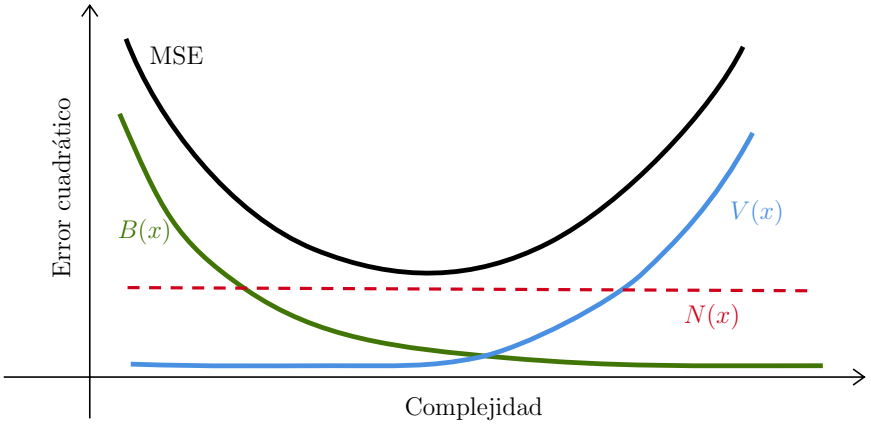

Figura 2: Descomposición del error en en función de la complejidad del modelo.

Al tomar esperanza $\mathbb{E}(\cdot \mid x)$, y observando que $\mathbb{E}(\epsilon \mid x) = 0$ y $\mathbb{E}(\epsilon^2 \mid x) = \mathrm{N}(x)$ obtenemos

$$\mathbb{E}\left[(f_S(x) - y)^2 \mid x\right] = (f_S(x) - r(x))^2 + \mathrm{N}(x)$$

Ahora, sumamos y restamos $\bar{f}(x)$:

$$\begin{aligned}\mathbb{E}\left[(f_S(x) - y)^2 \mid x\right] &= (f_S(x) - \bar{f}(x) + \bar{f}(x) - r(x))^2 + \mathrm{N}(x) \\ &= (f_S(x) - \bar{f}(x))^2 \\ &\quad - 2(f_S(x) - \bar{f}(x))(\bar{f}(x) - r(x)) \\ &\quad + (\bar{f}(x) - r(x))^2 \\ &\quad + \mathrm{N}(x)\end{aligned}$$

Notar que $\mathrm{B}(x) = (\bar{f}(x) - r(x))^2$ y al tomar esperanza en $S \sim \mathcal{D}^N$ el segundo término se anula. Entonces:

$$\begin{aligned}\mathrm{MSE}(x) &= \mathbb{E}_{S\sim\mathcal{D}^N}\left[(f_S(x) - \bar{f}(x))^2\right] + \mathrm{B}(x) + \mathrm{N}(x) \\ &= \mathrm{V}(x) + \mathrm{B}(x) + \mathrm{N}(x)\end{aligned}$$

Esto termina la demostración. □

En la Fig. 2 se muestra la típica curva en forma de U de la MSE y su descomposición en función de la complejidad de un modelo.