Jérémie Cabessa Laboratoire DAVID, UVSQ

► Soit une relation fonctionnelle

$$Y = f(\boldsymbol{X}) + \epsilon$$

entre des variables d'input  $\boldsymbol{X}=(X_1,\ldots,X_p)$  et d'output Y. Le bruit  $\epsilon$  satisfait  $\mathrm{E}(\epsilon)=0$ .

On cherche à obtenir un modèle

$$\hat{Y} = \hat{f}(\boldsymbol{X})$$

qui soit le plus performant possible sur le test set!

► Soit une relation fonctionnelle

$$Y = f(\boldsymbol{X}) + \epsilon$$

entre des variables d'input  $\boldsymbol{X}=(X_1,\ldots,X_p)$  et d'output Y. Le bruit  $\epsilon$  satisfait  $\mathrm{E}(\epsilon)=0$ .

On cherche à obtenir un modèle

$$\hat{Y} = \hat{f}(\boldsymbol{X})$$

qui soit le plus performant possible sur le test set!

Soit un dataset

$$S = \{(\boldsymbol{x_i}, y_i) \in \mathbb{R}^M \times \mathbb{R} : i = 1, \dots, N\}$$

et soit  $(x,y) \in S$  un point du dataset.

- $lackbox{ On note } \mathrm{E}_{S,\epsilon}\left[\ldots\right] := \mathrm{E}\left[\ldots\right].$
- Par hypothèse, on a:  $E[\epsilon] = 0$ . Et puisque la "vraie" relation fonctionnelle f est déterministe, on a: E[f] = f.
- On rappelle que

$$Var[X] := E[(X - E[X])^2] = E[X^2] - E[X]^2$$

Soit un dataset

$$S = \{(\boldsymbol{x_i}, y_i) \in \mathbb{R}^M \times \mathbb{R} : i = 1, \dots, N\}$$

et soit  $(x, y) \in S$  un point du dataset.

- ightharpoonup On note  $\mathrm{E}_{S,\epsilon}\left[\dots\right]:=\mathrm{E}\left[\dots\right].$
- Par hypothèse, on a:  $E[\epsilon] = 0$ . Et puisque la "vraie" relation fonctionnelle f est déterministe, on a: E[f] = f.
- On rappelle que

$$Var[X] := E[(X - E[X])^2] = E[X^2] - E[X]^2$$

Soit un dataset

$$S = \{(\boldsymbol{x_i}, y_i) \in \mathbb{R}^M \times \mathbb{R} : i = 1, \dots, N\}$$

et soit  $(x,y) \in S$  un point du dataset.

- ightharpoonup On note  $\mathrm{E}_{S,\epsilon}\left[\ldots\right]:=\mathrm{E}\left[\ldots\right].$
- Par hypothèse, on a:  $E[\epsilon] = 0$ . Et puisque la "vraie" relation fonctionnelle f est déterministe, on a: E[f] = f.
- On rappelle que

$$Var[X] := E[(X - E[X])^2] = E[X^2] - E[X]^2$$

Soit un dataset

$$S = \{(\boldsymbol{x_i}, y_i) \in \mathbb{R}^M \times \mathbb{R} : i = 1, \dots, N\}$$

et soit  $(x,y) \in S$  un point du dataset.

- ightharpoonup On note  $\mathrm{E}_{S,\epsilon}\left[\ldots\right]:=\mathrm{E}\left[\ldots\right].$
- Par hypothèse, on a:  $E[\epsilon] = 0$ . Et puisque la "vraie" relation fonctionnelle f est déterministe, on a: E[f] = f.
- On rappelle que

$$Var[X] := E[(X - E[X])^2] = E[X^2] - E[X]^2.$$

#### On a:

$$\mathbf{E}\left[\left]y - \hat{f}(x)\right]^{2} \stackrel{\mathbf{i}}{=} \mathbf{E}\left[\left]f(x) + \epsilon - \hat{f}(x)\right]^{2} \stackrel{\mathbf{i}}{=} \mathbf{E}\left[\left]f(x) - \mathbf{E}\left[\hat{f}(x)\right] + \underbrace{\epsilon}_{=B} + \mathbf{E}\left[\hat{f}(x)\right] - \hat{f}(x)\right]^{2} \stackrel{\mathbf{i}}{=} \mathbf{E}\left[\left[f(x)\right] - f(x)\right]^{2} \stackrel{\mathbf{i}}{=} \mathbf{$$

$$\operatorname{Biais}[\hat{f}(\boldsymbol{x})] := \operatorname{E}[\hat{f}(\boldsymbol{x})] - f(\boldsymbol{x}) \quad \operatorname{Var}[\hat{f}(\boldsymbol{x})] := \operatorname{E}[]\hat{f}(\boldsymbol{x}) - \operatorname{E}[\hat{f}(\boldsymbol{x})]^{2^{1}}$$

On a:

$$\mathbf{E}\left[\left]y - \hat{f}(\boldsymbol{x})\right]^{2} \stackrel{\mathbf{i}}{=} \mathbf{E}\left[\left]f(\boldsymbol{x}) + \epsilon - \hat{f}(\boldsymbol{x})\right]^{2} \stackrel{\mathbf{i}}{=} \mathbf{E}\left[\left]\underbrace{f(\boldsymbol{x}) - \mathbf{E}\left[\hat{f}(\boldsymbol{x})\right]}_{=A} + \underbrace{\epsilon} + \underbrace{\mathbf{E}\left[\hat{f}(\boldsymbol{x})\right] - \hat{f}(\boldsymbol{x})}_{=C}\right)^{2} \stackrel{\mathbf{i}}{=} \mathbf{E}\left[\left[\frac{\hat{f}(\boldsymbol{x}) - \mathbf{E}\left[\hat{f}(\boldsymbol{x})\right]}{\mathbf{E}(B)}\right]^{2} \stackrel{\mathbf{i}}{=} \mathbf{E}\left[\left[\frac{\hat{f}(\boldsymbol{x}) - \mathbf{E}\left[\hat{f}(\boldsymbol{x})\right]}{\mathbf{E}(B)}\right]^{2} + \mathbf{E}\left[\left[\frac{\hat{f}(\boldsymbol{x}) - \mathbf{E}\left[\hat{f}(\boldsymbol{x})\right]}{\mathbf{E}(B)}\right]^{2} \stackrel{\mathbf{i}}{=} \mathbf{E}\left[\left[\frac{\hat{f}(\boldsymbol{x}) - \mathbf{E}\left[\hat{f}(\boldsymbol{x})\right]}{\mathbf{E}(B)}\right] \stackrel{\mathbf{i}}{=} \mathbf{E}\left[\left[\frac{\hat{f}(\boldsymbol{x}) - \mathbf{E}\left$$

οù

$$\operatorname{Biais}[\hat{f}(\boldsymbol{x})] := \operatorname{E}[\hat{f}(\boldsymbol{x})] - f(\boldsymbol{x}) \quad \operatorname{Var}[\hat{f}(\boldsymbol{x})] := \operatorname{E}[\hat{f}(\boldsymbol{x}) - \operatorname{E}[\hat{f}(\boldsymbol{x})])^{2^{1}}$$

On a:

$$\mathbf{E}\left[\left]y - \hat{f}(\boldsymbol{x})\right]^{2} \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left]f(\boldsymbol{x}) + \epsilon - \hat{f}(\boldsymbol{x})\right]^{2} \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left]\underbrace{f(\boldsymbol{x}) - \mathbf{E}\left[\hat{f}(\boldsymbol{x})\right]}_{=A} + \underbrace{\epsilon} + \underbrace{\mathbf{E}\left[\hat{f}(\boldsymbol{x})\right] - \hat{f}(\boldsymbol{x})}_{=C}\right)^{2} \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[\hat{f}(\boldsymbol{x})\right] - \hat{f}(\boldsymbol{x})\right]^{2} \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[\hat{f}(\boldsymbol{x})\right] - \hat{f}(\boldsymbol{x})\right] \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[\hat{f}(\boldsymbol{x})\right] - \hat{f}(\boldsymbol{x})\right] \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[\hat{f}(\boldsymbol{x})\right] - \hat{f}(\boldsymbol{x})\right] \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[\hat{f}(\boldsymbol{x})\right] - \hat{f}(\boldsymbol{x})\right] \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[\hat{f}(\boldsymbol{x})\right] \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[\hat{f}(\boldsymbol{x})\right] - \hat{f}(\boldsymbol{x})\right] \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[\hat{f}(\boldsymbol{x})\right] - \hat{f}(\boldsymbol{x})\right] \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[\hat{f}(\boldsymbol{x})\right] \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[\hat{f}(\boldsymbol{x})\right] - \hat{f}(\boldsymbol{x})\right] \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[\hat{f}(\boldsymbol{x})\right] \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[\hat{f}(\boldsymbol{x})\right] \stackrel{\boldsymbol{i}}{=} \mathbf{E}\left[\left[$$

οù

$$\operatorname{Biais}\big[\hat{f}(\boldsymbol{x})\big] := \operatorname{E}\big[\hat{f}(\boldsymbol{x})\big] - f(\boldsymbol{x}) \quad \operatorname{Var}\big[\hat{f}(\boldsymbol{x})\big] := \operatorname{E}\Big[\big]\hat{f}(\boldsymbol{x}) - \operatorname{E}[\hat{f}(\boldsymbol{x})]\Big)^{2^{b}}$$

On a:

$$\begin{split} \mathbf{E}\Big[\Big]y - \hat{f}(\boldsymbol{x})\Big)^{2} &= \mathbf{E}\Big[\Big]f(\boldsymbol{x}) + \epsilon - \hat{f}(\boldsymbol{x})\Big)^{2} \\ &= \mathbf{E}\Big[\Big]\underbrace{f(\boldsymbol{x}) - \mathbf{E}[\hat{f}(\boldsymbol{x})]}_{=A} + \underbrace{\epsilon}_{=B} + \underbrace{\mathbf{E}[\hat{f}(\boldsymbol{x})] - \hat{f}(\boldsymbol{x})}_{=C}\Big)^{2} \\ &= \mathbf{E}\left[A^{2} + B^{2} + C^{2} + 2AB + 2BC + 2CA\right] \\ (\mathbf{E}(B) = \mathbf{E}(C) = \mathbf{0}) &= \mathbf{E}\left[A^{2}\right] + \mathbf{E}\left[B^{2}\right] + \mathbf{E}\left[C^{2}\right] \\ (\mathbf{E}(\epsilon) = \mathbf{0}) &= \mathbf{E}\Big[\int f(\boldsymbol{x}) - \mathbf{E}[\hat{f}(\boldsymbol{x})]\Big)^{2} + \mathbf{E}\Big[\int \epsilon - \mathbf{E}(\epsilon)\Big)^{2} + \mathbf{E}\Big[\int \hat{f}(\boldsymbol{x}) - \mathbf{E}[\hat{f}(\boldsymbol{x})]\Big)^{2} \\ (\mathbf{E}[f(\boldsymbol{x})] = f(\boldsymbol{x})) &= \left(\mathbf{E}[\hat{f}(\boldsymbol{x})] - f(\boldsymbol{x})\right)^{2} + \mathbf{Var}[\hat{f}(\boldsymbol{x})] + \mathbf{Var}(\epsilon) \\ &= \mathbf{E}[\mathbf{ais}[\hat{f}(\boldsymbol{x})]^{2} + \mathbf{Var}[\hat{f}(\boldsymbol{x})] + \mathbf{Var}(\epsilon) \end{split}$$

$$\operatorname{Biais}[\hat{f}(\boldsymbol{x})] := \operatorname{E}[\hat{f}(\boldsymbol{x})] - f(\boldsymbol{x}) \quad \operatorname{Var}[\hat{f}(\boldsymbol{x})] := \operatorname{E}[]\hat{f}(\boldsymbol{x}) - \operatorname{E}[\hat{f}(\boldsymbol{x})])^{2}$$

On a:

$$E[]y - \hat{f}(x))^{2} = E[]f(x) + \epsilon - \hat{f}(x)]^{2}$$

$$= E[]\underbrace{f(x) - E[\hat{f}(x)]}_{=A} + \underbrace{\epsilon}_{=B} + \underbrace{E[\hat{f}(x)] - \hat{f}(x)}_{=C})^{2}$$

$$= E[A^{2} + B^{2} + C^{2} + 2AB + 2BC + 2CA]$$

$$(E(B) = E(C) = 0) = E[A^{2}] + E[B^{2}] + E[C^{2}]$$

$$(E(\epsilon) = 0) = E[]f(x) - E[\hat{f}(x)])^{2} + E[]\epsilon - E(\epsilon)^{2} + E[]\hat{f}(x) - E[\hat{f}(x)])^{2}^{i}$$

$$(E[f(x)] = f(x)) = (E[\hat{f}(x)] - f(x))^{2} + Var[\hat{f}(x)] + Var(\epsilon)$$

$$= Biais[\hat{f}(x)]^{2} + Var[\hat{f}(x)] + Var(\epsilon)$$

οi

$$\operatorname{Biais}[\hat{f}(\boldsymbol{x})] := \operatorname{E}[\hat{f}(\boldsymbol{x})] - f(\boldsymbol{x}) \quad \operatorname{Var}[\hat{f}(\boldsymbol{x})] := \operatorname{E}[\hat{f}(\boldsymbol{x}) - \operatorname{E}[\hat{f}(\boldsymbol{x})])^{2}$$

On a:

$$E[]y - \hat{f}(x))^{2} \stackrel{i}{=} E[]f(x) + \epsilon - \hat{f}(x)]^{2}$$

$$= E[]\underbrace{f(x) - E[\hat{f}(x)]}_{=A} + \underbrace{\epsilon}_{=B} + \underbrace{E[\hat{f}(x)] - \hat{f}(x)}_{=C})^{2}$$

$$= E[A^{2} + B^{2} + C^{2} + 2AB + 2BC + 2CA]$$

$$(E(B) = E(C) = 0) = E[A^{2}] + E[B^{2}] + E[C^{2}]$$

$$(E(\epsilon) = 0) = E[]f(x) - E[\hat{f}(x)])^{2} + E[]\epsilon - E(\epsilon)^{2} + E[]\hat{f}(x) - E[\hat{f}(x)])^{2}$$

$$(E[f(x)] = f(x)) = (E[\hat{f}(x)] - f(x))^{2} + Var[\hat{f}(x)] + Var(\epsilon)$$

$$= Biais[\hat{f}(x)]^{2} + Var[\hat{f}(x)] + Var(\epsilon)$$

οù

$$\operatorname{Biais}[\hat{f}(\boldsymbol{x})] := \operatorname{E}[\hat{f}(\boldsymbol{x})] - f(\boldsymbol{x}) \quad \operatorname{Var}[\hat{f}(\boldsymbol{x})] := \operatorname{E}[\hat{f}(\boldsymbol{x}) - \operatorname{E}[\hat{f}(\boldsymbol{x})])^{2}$$

On a:

$$\begin{split} \mathbf{E}\Big[\Big]y - \hat{f}(\boldsymbol{x})\Big)^{2} &= \mathbf{E}\Big[\Big]f(\boldsymbol{x}) + \epsilon - \hat{f}(\boldsymbol{x})\Big)^{2} \\ &= \mathbf{E}\Big[\Big]\underbrace{f(\boldsymbol{x}) - \mathbf{E}\big[\hat{f}(\boldsymbol{x})\big]}_{=A} + \underbrace{\mathbf{E}\big[\hat{f}(\boldsymbol{x})\big] - \hat{f}(\boldsymbol{x})}_{=C}\Big)^{2} \\ &= \mathbf{E}\left[A^{2} + B^{2} + C^{2} + 2AB + 2BC + 2CA\right] \\ (\mathbf{E}(B) = \mathbf{E}(C) = 0) &= \mathbf{E}\left[A^{2}\right] + \mathbf{E}\left[B^{2}\right] + \mathbf{E}\left[C^{2}\right] \\ (\mathbf{E}(\epsilon) = 0) &= \mathbf{E}\Big[f(\boldsymbol{x}) - \mathbf{E}\big[\hat{f}(\boldsymbol{x})\big]\Big)^{2} + \mathbf{E}\Big[f(\epsilon) - \mathbf{E}(\epsilon)\Big]^{2} + \mathbf{E}\Big[f(\epsilon) - \mathbf{E}(\epsilon)\Big]^{2} \\ (\mathbf{E}[f(\boldsymbol{x})] = f(\boldsymbol{x})) &= \left(\mathbf{E}\big[\hat{f}(\boldsymbol{x})\big] - f(\boldsymbol{x})\right)^{2} + \mathbf{Var}\big[\hat{f}(\boldsymbol{x}) + \mathbf{Var}(\epsilon) \\ &= \mathbf{Biais}\big[\hat{f}(\boldsymbol{x})\big]^{2} + \mathbf{Var}\big[\hat{f}(\boldsymbol{x})\big] + \mathbf{Var}(\epsilon) \end{split}$$

$$\operatorname{Biais}[\hat{f}(\boldsymbol{x})] := \operatorname{E}[\hat{f}(\boldsymbol{x})] - f(\boldsymbol{x}) \quad \operatorname{Var}[\hat{f}(\boldsymbol{x})] := \operatorname{E}[\hat{f}(\boldsymbol{x}) - \operatorname{E}[\hat{f}(\boldsymbol{x})])^{2}$$

On a:

$$E[]y - \hat{f}(\boldsymbol{x}))^{2} = E[]f(\boldsymbol{x}) + \epsilon - \hat{f}(\boldsymbol{x})]^{2}$$

$$= E[]\underbrace{f(\boldsymbol{x}) - E[\hat{f}(\boldsymbol{x})]}_{=A} + \underbrace{\epsilon}_{=B} + \underbrace{E[\hat{f}(\boldsymbol{x})] - \hat{f}(\boldsymbol{x})}_{=C})^{2}$$

$$= E[A^{2} + B^{2} + C^{2} + 2AB + 2BC + 2CA]$$

$$(E(B) = E(C) = 0) = E[A^{2}] + E[B^{2}] + E[C^{2}]$$

$$(E(\epsilon) = 0) = E[]f(\boldsymbol{x}) - E[\hat{f}(\boldsymbol{x})])^{2} + E[]\epsilon - E(\epsilon)^{2} + E[]\hat{f}(\boldsymbol{x}) - E[\hat{f}(\boldsymbol{x})])^{2}^{i}$$

$$(E[f(\boldsymbol{x})] = f(\boldsymbol{x})) = (E[\hat{f}(\boldsymbol{x})] - f(\boldsymbol{x}))^{2} + Var[\hat{f}(\boldsymbol{x})] + Var(\epsilon)$$

$$= Biais[\hat{f}(\boldsymbol{x})]^{2} + Var[\hat{f}(\boldsymbol{x})] + Var(\epsilon)$$

οù

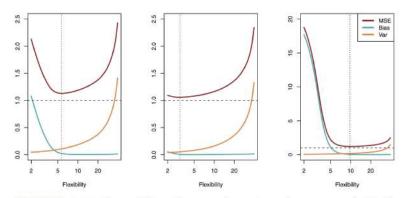
$$\operatorname{Biais} \left[ \hat{f}(\boldsymbol{x}) \right] := \operatorname{E} \left[ \hat{f}(\boldsymbol{x}) \right] - f(\boldsymbol{x}) \quad \operatorname{Var} \left[ \hat{f}(\boldsymbol{x}) \right] := \operatorname{E} \left[ \left| \hat{f}(\boldsymbol{x}) - \operatorname{E} \left[ \hat{f}(\boldsymbol{x}) \right] \right|^{2} \right]$$

- Le biais  $\operatorname{Biais}[\hat{f}(x)]$  représente l'erreur due à la complexité du modèle  $\hat{f}$ .
- La variance  $\operatorname{Var}[\hat{f}(\boldsymbol{x})]$  représente la sensibilité du modèle  $\hat{f}$  par rapport à son training set, i.e., la variation moyenne de  $\hat{f}(\boldsymbol{x})$  autour de sa moyenne  $\operatorname{E}[\hat{f}(\boldsymbol{x})]$  si le modèle  $\hat{f}$  était estimé à partir de différents training sets.
- $\operatorname{Var}(\epsilon)$  représente l'erreur irréductible liée au bruit inhérent à la vraie relation fonctionnelle f.
- Dilemme biais-variance (bias-variance trade-off): plus le modèle f est complexe, plus le biais sera faible, mais plus la variance sera élevée.

- Le biais  $\operatorname{Biais}[\hat{f}(x)]$  représente l'erreur due à la complexité du modèle  $\hat{f}$ .
- La variance  $\mathrm{Var} \big[ \hat{f}(\boldsymbol{x}) \big]$  représente la sensibilité du modèle  $\hat{f}$  par rapport à son training set, i.e., la variation moyenne de  $\hat{f}(\boldsymbol{x})$  autour de sa moyenne  $\mathrm{E}[\hat{f}(\boldsymbol{x})]$  si le modèle  $\hat{f}$  était estimé à partir de différents training sets.
- $ightharpoonup Var(\epsilon)$  représente l'erreur irréductible liée au bruit inhérent à la vraie relation fonctionnelle f.
- ▶ Dilemme biais-variance (bias-variance trade-off): plus le modèle  $\hat{f}$  est complexe, plus le biais sera faible, mais plus la variance sera élevée.

- Le biais  $\operatorname{Biais}[\hat{f}(x)]$  représente l'erreur due à la complexité du modèle  $\hat{f}$ .
- La variance  $\mathrm{Var} \big[ \hat{f}(\boldsymbol{x}) \big]$  représente la sensibilité du modèle  $\hat{f}$  par rapport à son training set, i.e., la variation moyenne de  $\hat{f}(\boldsymbol{x})$  autour de sa moyenne  $\mathrm{E}[\hat{f}(\boldsymbol{x})]$  si le modèle  $\hat{f}$  était estimé à partir de différents training sets.
- ▶  $Var(\epsilon)$  représente **l'erreur irréductible** liée au bruit inhérent à la vraie relation fonctionnelle f.
- ▶ Dilemme biais-variance (bias-variance trade-off): plus le modèle  $\hat{f}$  est complexe, plus le biais sera faible, mais plus la variance sera élevée.

- Le biais  $\operatorname{Biais}[\hat{f}(x)]$  représente l'erreur due à la complexité du modèle  $\hat{f}$ .
- La variance  $\mathrm{Var} \big[ \hat{f}(\boldsymbol{x}) \big]$  représente la sensibilité du modèle  $\hat{f}$  par rapport à son training set, i.e., la variation moyenne de  $\hat{f}(\boldsymbol{x})$  autour de sa moyenne  $\mathrm{E}[\hat{f}(\boldsymbol{x})]$  si le modèle  $\hat{f}$  était estimé à partir de différents training sets.
- ▶  $Var(\epsilon)$  représente **l'erreur irréductible** liée au bruit inhérent à la vraie relation fonctionnelle f.
- **Dilemme biais-variance (bias-variance trade-off):** plus le modèle  $\hat{f}$  est complexe, plus le biais sera faible, mais plus la variance sera élevée.



**FIGURE 2.12.** Squared bias (blue curve), variance (orange curve),  $Var(\epsilon)$  (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

Figure taken from [James et al., 2013]

#### **BIBLIOGRAPHIE**



Fleuret, F. (2022). Deep Learning Course.



James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013).

An Introduction to Statistical Learning: with Applications in R, volume 103 of Springer Texts in Statistics.

Springer, New York.



Wikipedia contributors (2022).

 ${\sf Bias-variance\ tradeoff-Wikipedia,\ the\ free\ encyclopedia}.$