Proof of the bias-variance decomposition

Assumptions

Labels y are given by a deterministic function f of the input variables $\mathbf{x}$ plus some random noise ϵ, such that

$$y(\mathbf{x}) = f(\mathbf{x}) + \epsilon \tag{1}$$

The noises ϵ are independent and identically distributed. We will assume a normal distribution with mean 0 and variance σ^2, such that

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

There exists a (generally unknown) distribution P from which the input variables $\mathbf{x}$ have been sampled:

$$\exists P \text{ s.t. } \mathbf{x} \sim P \tag{3}$$

We can then use this distribution as well as (1) and (2) to sample a dataset $\mathcal{D}$ of N elements

$$\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n \in [\![1,N]\!]} \tag{4}$$

We will estimate a prediction function $\hat{f}_\mathcal{D}$ from $\mathcal{D}$, from which we can predict labels $\hat{y}$ from new inputs $\mathbf{x}$:

$$\hat{y} = \hat{f}_\mathcal{D}(\mathbf{x}) \tag{5}$$

Finally, we assume that $\hat{f}_\mathcal{D}$ and ϵ are independent (this is not obvious, we will prove this result for the linear regression later as an exercise).

$$\hat{f}_\mathcal{D} \perp\!\!\!\perp \epsilon \tag{6}$$

Reminders

$$\operatorname{cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{7}$$

$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if and only if X and Y are independent, *i.e.*

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \iff X \perp\!\!\!\perp Y \tag{8}$$

If $X = Y$ then

$$\operatorname{cov}[X, X] = \operatorname{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{9}$$

Proof

We will prove

$$\underbrace{\mathbb{E}_{\mathcal{D},\mathbf{x}}[(y(\mathbf{x}) - \hat{y}_\mathcal{D}(\mathbf{x}))^2]}_{\text{expected error}} = \underbrace{\mathbb{E}_\mathbf{x}[f(\mathbf{x}) - \mathbb{E}_\mathcal{D}[\hat{f}]]^2}_{\text{bias}^2} + \underbrace{\operatorname{var}_\mathcal{D}[\hat{f}]}_{\text{variance}} + \underbrace{\sigma^2}_{\text{noise}} \tag{10}$$

To simplify notations we write $f = f(\mathbf{x})$ and $y = y(\mathbf{x})$

$$\mathbb{E}[(y-\hat{y})^2] = \mathbb{E}[(f+\epsilon-\hat{f})^2] \quad \text{from (1) and (5)} \tag{11}$$

$$= \mathbb{E}[\Big(f+\epsilon-\hat{f}+\mathbb{E}[\hat{f}]-\mathbb{E}[\hat{f}]\Big)^2] \tag{12}$$

$$= \mathbb{E}[\Big(\underbrace{(f-\mathbb{E}[\hat{f}])}_{=a \text{ (def.)}} + \underbrace{(\mathbb{E}[\hat{f}]-\hat{f}+\epsilon)}_{=b \text{ (def.)}}\Big)^2] \tag{13}$$

$$= \mathbb{E}[\underbrace{(f-\mathbb{E}[\hat{f}])^2}_{=a^2} + \underbrace{2(f-\mathbb{E}[\hat{f}])(\mathbb{E}[\hat{f}]-\hat{f}+\epsilon)}_{=2ab} + \underbrace{(\mathbb{E}[\hat{f}]-\hat{f}+\epsilon)^2}_{=b^2}] \tag{14}$$

From the linearity of the expectation $\mathbb{E}$,

$$\mathbb{E}[a^2+2ab+b^2] = \mathbb{E}[a^2]+2\mathbb{E}[ab]+\mathbb{E}[b^2] \tag{15}$$

We have

$$2ab = 2(f-\mathbb{E}[\hat{f}])(\mathbb{E}[\hat{f}]-\hat{f}+\epsilon) \tag{16}$$

$$= 2(f-\mathbb{E}[\hat{f}])(\mathbb{E}[\hat{f}]-\hat{f}) + 2(f-\mathbb{E}[\hat{f}])\epsilon \tag{17}$$

$$b^2 = (\mathbb{E}[\hat{f}]-\hat{f}+\epsilon)^2 \tag{18}$$

$$= (\mathbb{E}[\hat{f}]-\hat{f})^2 + 2(\mathbb{E}[\hat{f}]-\hat{f})\epsilon + \epsilon^2 \tag{19}$$

So

$$\mathbb{E}[(y-\hat{y})^2] = \underbrace{\mathbb{E}[(f-\mathbb{E}[\hat{f}])^2]}_{=\mathbb{E}[a^2]} + \underbrace{2\mathbb{E}[(f-\mathbb{E}[\hat{f}])(\mathbb{E}[\hat{f}]-\hat{f})]}_{=c \text{ (def.)}} + \underbrace{2\mathbb{E}[(f-\mathbb{E}[\hat{f}])\epsilon]}_{=d \text{ (def.)}} \tag{20}$$

$$+ \underbrace{\mathbb{E}[(\mathbb{E}[\hat{f}]-\hat{f})^2]}_{=e \text{ (def.)}} + \underbrace{2\mathbb{E}[(\mathbb{E}[\hat{f}]-\hat{f})\epsilon]}_{=g \text{ (def.)}} + \underbrace{\mathbb{E}[\epsilon^2]}_{=\mathbb{E}[\epsilon^2]} \tag{21}$$

Let's have a look at the different terms

$$\mathbb{E}[a^2] = \mathbb{E}[(f-\mathbb{E}[\hat{f}])^2] \tag{22}$$

$(f-\mathbb{E}[\hat{f}])$ does not depend on $\mathcal{D}$ so

$$\mathbb{E}[a^2] = (f-\mathbb{E}[\hat{f}])^2 \tag{23}$$

$$c = 2\mathbb{E}[(f-\mathbb{E}[\hat{f}])(\mathbb{E}[\hat{f}]-\hat{f})] \tag{24}$$

Again $(f-\mathbb{E}[\hat{f}])$ does not depend on $\mathcal{D}$ so $\mathbb{E}[f-\mathbb{E}[\hat{f}]] = f-\mathbb{E}[\hat{f}]$

Then

$$c = 2(f-\mathbb{E}[\hat{f}])\mathbb{E}[\mathbb{E}[\hat{f}]-\hat{f}] \tag{25}$$

and

$$\mathbb{E}[\mathbb{E}[\hat{f}]-\hat{f}] = \mathbb{E}[\mathbb{E}[\hat{f}]]-\mathbb{E}[\hat{f}] = \mathbb{E}[\hat{f}]-\mathbb{E}[\hat{f}] = 0 \tag{26}$$

so

$$c = 0 \tag{27}$$

Since $\hat{f}$ and ϵ are independent (assumption (6)), from (8) we have

$$d = 2\mathbb{E}[(f - \mathbb{E}[\hat{f}])\epsilon] = 2 \cdot \underbrace{\mathbb{E}[\epsilon]}_{=0 \text{ from (2)}} \cdot \mathbb{E}[f - \mathbb{E}[\hat{f}]] = 0 \tag{28}$$

Similarly

$$g = 2\mathbb{E}[(\mathbb{E}[\hat{f}] - \hat{f})\epsilon] = 0 \tag{29}$$

$$e = \mathbb{E}[(\mathbb{E}[\hat{f}] - \hat{f})^2] = \mathrm{var}[\hat{f}] \quad \text{from (9)} \tag{30}$$

$$\mathbb{E}[\epsilon^2] = \mathrm{var}[\epsilon] \quad \text{from (9) since } \mathbb{E}[\epsilon]^2 = 0^2 = 0 \tag{31}$$

$$= \sigma^2 \quad \text{from (2)} \tag{32}$$

And finally, from (20), (23), (27), (28), (30), (29), (28) and (31)

$$\mathbb{E}[(y - \hat{y})^2] = (f - \mathbb{E}[\hat{f}])^2 + \mathrm{var}[\hat{f}] + \sigma^2 \tag{33}$$

Q.E.D.