General decomposition of the generalization error into Variance - Bias - Noise

Consider:

1.) Sample space $\mathcal{X}$
2.) target space $\mathcal{Y}$
3.) training data $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$
as a random variable
on $(\mathcal{X} \times \mathcal{Y})^N$
4.) test data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$

D and (x, y) sampled from "true" distribution P_0

5.) loss function:

$$L : \mathcal{Y} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R} \cup \{\pm\infty\}$$

↗ space of all probability distributions/densities on $\mathcal{Y}$

s.t. a) L is convex in 2nd argument.
b) if $y \sim p(y) \Rightarrow p \in \arg\min_q \mathbb{E}_p[L(y, q)]$

6.) Any statistical model class
$\mathcal{Q} = \{q(y|x)\}$

and learning algorithm ϕ
(e.g. MLE, MAP, full Bayesian)
together determine via learning on $\mathcal{D}$ a map $q_{\mathcal{D}}$:

$$q(y|x, \mathcal{D}) : (\mathcal{X} \times \mathcal{Y})^N \times \mathcal{X} \to \mathcal{P}(\mathcal{Y})$$

training data; test x data input; prediction for y

7.) Put : $\bar{q}(Y|X) := \mathbb{E}_{\mathcal{D}}[q(Y|X,\mathcal{D})]$

$= \int q(Y|X,\mathcal{D})\, p_0(\mathcal{D}|X)\, d\mathcal{D}$

the mean prediction probability averaged over all data sets.

(i.e. sampling several $\mathcal{D}_1, \mathcal{D}_2 \ldots \sim p_0$ running on each the learning algorithm A and afterwards averaging)

8.) Let $q_0(Y|X) \in \operatorname{argmin}_{q(Y|X)} \mathbb{E}_0[L(Y, q(Y|X))|X]$

the optimal prediction probability which usually is $p_0(Y|X)$ "true" (depending on L)

We then define:

9.) Generalization error:

$$\mathcal{E}_L(q_D|x) := \mathbb{E}_0[L(Y, q(Y|x,D))|X]$$

10.) Variance:

$$\mathrm{Var}_L(q_D|x) := \mathbb{E}_0[L(Y, q(Y|x,D))|X] - \mathbb{E}_0[L(Y, \bar{q}(Y|x))|X] \geq 0$$

Jensen

11.) Bias:

$$\mathrm{Bias}_L(q_D|x) := \mathbb{E}_0[L(Y, \bar{q}(Y|x)|X] - \mathbb{E}_0[L(Y, q_0(Y|x))|X] \geq 0$$

$p_0 \in \mathrm{argmin}$

12.) Noise:

$$\mathrm{Noise}_L(q_D|x) := \mathbb{E}_0[L(Y, q_0(Y|x))|X]$$

↳ independent of $\mathcal{Q}$ and $\mathcal{A}$.
↳ non-reducible noise.

Then we have the decomposition:

$$\mathcal{E}_L = Var_L + Bias_L + Noise_L$$

$$\mathcal{E}_L \geq Noise_L$$

Example : $L(y, q) := (y - \mathbb{E}_{y' \sim q}[y'])^2$

"square-loss"

$$\text{Var}_L = \mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{q(y|x,\mathcal{D})}(y|x) - \mathbb{E}_{\bar{q}(y|x)}(y|x) \right)^2 \middle| x \right]$$

$$\text{Bias}_L = \left(\mathbb{E}_{\bar{q}}(y|x) - \mathbb{E}_{\mathcal{D}}(y|x) \right)^2$$

$$\text{Noise}_L = \text{Var}_{\mathcal{D}}(y|x)$$

$$\mathcal{E}_L = \mathbb{E}_{\mathcal{D}}\left[\left(y - \mathbb{E}_{q(y|x,\mathcal{D})}(y|x) \right)^2 \middle| x \right]$$

$\Rightarrow$ Recover old formula

Check:

$$\mathrm{Var}_L = \mathbb{E}_0\left[L(Y, q(Y|X,\mathcal{D}))|X\right] - \mathbb{E}_0\left[L(Y, \bar{q}(Y|X))|X\right]$$

$$= \mathbb{E}_0\left[(Y - \mathbb{E}_q[Y|X,\mathcal{D}])^2|X\right] - \mathbb{E}_0\left[(Y - \mathbb{E}_{\bar{q}}[Y|X])^2|X\right]$$

$$= \mathbb{E}_0\Big[\cancel{Y^2} - 2\cdot Y\cdot \mathbb{E}_q[Y|X,\mathcal{D}] + \mathbb{E}_q[Y|X,\mathcal{D}]^2 \cancel{- Y^2} + 2\cdot Y\cdot \mathbb{E}_{\bar{q}}[Y|X] - \mathbb{E}_{\bar{q}}[Y|X]^2 \;|X\Big]$$

$$= \mathbb{E}_0\left[\mathbb{E}_q[Y|X,\mathcal{D}]^2 - 2\,Y\cdot \mathbb{E}_q[Y|X,\mathcal{D}]\,|X\right] + 2\,\mathbb{E}_0[Y|X]\cdot\mathbb{E}_{\bar{q}}[Y|X] - \mathbb{E}_{\bar{q}}[Y|X]^2$$

$Y \perp\!\!\!\perp \mathcal{D} \mid X$; $p_0(Y,\mathcal{D}|X) = p_0(Y|X)\cdot p_0(\mathcal{D}|X)$

$$= \mathbb{E}_0\left[\mathbb{E}_q[Y|X,\mathcal{D}]^2 - 2\,\mathbb{E}_0[Y|X]\cdot\mathbb{E}_{\bar{q}}[Y|X]\,|X\right] + 2\,\mathbb{E}_0[Y|X]\,\mathbb{E}_{\bar{q}}[Y|X] - \mathbb{E}_{\bar{q}}[Y|X]^2$$

$$= \mathbb{E}_0\left[\left(\mathbb{E}_q[Y|X,\mathcal{D}] - \mathbb{E}_{\bar{q}}[Y|X]\right)^2|X\right]$$

$$\text{Bias}_L = \mathbb{E}_0\left[\left(Y - \mathbb{E}_{\bar{q}}(Y|X)\right)^2 | X\right] - \mathbb{E}_0\left[\left(Y - \mathbb{E}_0(Y|X)\right)^2 | X\right]$$

$$= \mathbb{E}_0\Big[\cancel{Y^2} - 2Y \cdot \mathbb{E}_{\bar{q}}(Y|X) + \mathbb{E}_{\bar{q}}(Y|X)^2 \cancel{- Y^2} + 2Y \cdot \mathbb{E}_0(Y|X) + \mathbb{E}_0(Y|X)^2 \,|\, X\Big]$$

$$= \mathbb{E}_{\bar{q}}|Y|X|^2 - 2\,\mathbb{E}_0(Y|X) \cdot \mathbb{E}_{\bar{q}}(Y|X) + \mathbb{E}_0(Y|X)^2$$

$$= \left(\mathbb{E}_{\bar{q}}(Y|X) - \mathbb{E}_0(Y|X)\right)^2$$

$$\text{Noise}_L = \mathbb{E}_0\left[\left(Y - \mathbb{E}_0(Y|X)\right)^2 | X\right]$$

$$= \text{Var}_0(Y|X)$$

Example : $L(y, q) := -\log q(y)$

a log-loss a

$$\text{Var}_L = \mathbb{E}_0\left[\log \frac{\bar{q}(Y|X)}{q(Y|X,\mathcal{D})}\right] \geq 0$$

$$\text{Bias}_L = KL(p_0(Y|X) \,\|\, \bar{q}(Y|X)) \geq 0$$

$$\text{Noise}_L = H(p_0(Y|X))$$

$$\mathcal{E}_L = CE[p_0(Y|X) \,\|\, q(Y|X,\mathcal{D})]$$

or generalization error beyond noise:

$$KL(p_0(Y|X) \,\|\, q(Y|X,\mathcal{D}))$$

$$= \underset{y \sim p_0(Y|X)}{\mathbb{E}}\left[\log \frac{\bar{q}(Y|X)}{q(Y|X,\mathcal{D})}\right] + KL(p_0(Y|X) \,\|\, \bar{q}(Y|X))$$

$\text{Var}_L \geq 0$ $\quad$ $\text{Bias}_L \geq 0$