Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: Data Scientist.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

Bias Variance Tradeoff

Consider a training set $S = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ and the random variable y which we cant to model it.

We asumming there is a function $f(x)$ that:

$$\boxed{y = f(x) + \epsilon}$$

with $e \sim D(0, \sigma^2)$.

When we training a model we want to find a function $\hat{f}(x|S)$ which approximates $f(x)$. Note $f(x)$ is deterministic and $\hat{f}(x|S)$ depends of the model chosen and its hyperparameters among others and for that it is a random variable.

The error of the aproximation is:

$$\boxed{\mathbb{E}\left[(y - \hat{f}(x))^2|S\right]}$$

We will developed every term.

$$\mathbb{E}\left[(y - \hat{f}(x))^2|S\right] = \mathbb{E}\left[y^2 - 2 \cdot y \cdot \hat{f}(x) + (\hat{f}(x))^2|S\right] = \mathbb{E}\left[y^2|S\right] - 2 \cdot \mathbb{E}\left[y \cdot \hat{f}(x)|S\right] + \mathbb{E}\left[(\hat{f}(x))^2|S\right]$$

$\Rightarrow$

$$\boxed{\mathbb{E}\left[(y - \hat{f}(x))^2|S\right] = \mathbb{E}\left[y^2|S\right] - 2 \cdot \mathbb{E}\left[y \cdot \hat{f}(x)|S\right] + \mathbb{E}\left[(\hat{f}(x))^2|S\right]} \quad (1)$$

Also we have:

$$\mathbb{V}\left[\hat{f}(x)|S\right] = \mathbb{E}\left[(\hat{f}(x))^2|S\right] - \left(\mathbb{E}\left[\hat{f}(x)|S\right]\right)^2$$

$\Rightarrow$

$$\boxed{\mathbb{E}\left[(\hat{f}(x))^2|S\right] = \mathbb{V}\left[\hat{f}(x)|S\right] + \left(\mathbb{E}\left[\hat{f}(x)|S\right]\right)^2} \quad (2)$$

Combining (1) and (2) we have:

$$\boxed{\mathbb{E}\left[(y-\hat{f}(x))^2|S\right] = \mathbb{E}\left[y^2|S\right] - 2\cdot\mathbb{E}\left[y\cdot\hat{f}(x)|S\right] + \mathbb{V}\left[\hat{f}(x)|S\right] + \left(\mathbb{E}\left[\hat{f}(x)|S\right]\right)^2} \tag{3}$$

Now:

$$\mathbb{E}\left[y^2|S\right] = \mathbb{E}\left[(f(x)+\epsilon)^2|S\right] = \mathbb{E}\left[f(x)^2 + 2\cdot\epsilon\cdot f(x) + \epsilon^2|S\right]$$

$$= \mathbb{E}\left[f(x)^2|S\right] + \mathbb{E}\left[2\cdot\epsilon\cdot f(x)|S\right] + \mathbb{E}\left[\epsilon^2|S\right] = f(x)^2 + 2\cdot f(x)\cdot\mathbb{E}\left[\epsilon|S\right] + \mathbb{E}\left[\epsilon^2|S\right]$$

$$= f(x)^2 + 2\cdot f(x)\cdot 0 + \mathbb{V}\left[\epsilon|S\right] + \left(\mathbb{E}\left[\epsilon|S\right]\right)^2 = f(x)^2 + \sigma^2$$

$\Rightarrow$

$$\mathbb{E}\left[y^2|S\right] = f(x)^2 + \sigma^2$$

Replacing this in (3):

$$\boxed{\mathbb{E}\left[(y-\hat{f}(x))^2|S\right] = f(x)^2 + \sigma^2 - 2\cdot\mathbb{E}\left[y\cdot\hat{f}(x)|S\right] + \mathbb{V}\left[\hat{f}(x)|S\right] + \left(\mathbb{E}\left[\hat{f}(x)|S\right]\right)^2} \tag{4}$$

On the other hand:

$$\mathbb{E}\left[y\cdot\hat{f}(x)|S\right] = \mathbb{E}\left[(f(x)+\epsilon)\cdot\hat{f}(x)|S\right] = \mathbb{E}\left[f(x)\cdot\hat{f}(x)|S\right] + \mathbb{E}\left[\epsilon\cdot\hat{f}(x)|S\right]$$

$$\overset{\hat{f}(x)\text{ and }\epsilon\text{ are independent}}{=} f(x)\cdot\mathbb{E}\left[\hat{f}(x)|S\right] + \mathbb{E}\left[\epsilon|S\right]\cdot\mathbb{E}\left[\hat{f}(x)|S\right]$$

$$= f(x)\cdot\mathbb{E}\left[\hat{f}(x)|S\right] + 0\cdot\mathbb{E}\left[\hat{f}(x)|S\right] = f(x)\cdot\mathbb{E}\left[\hat{f}(x)|S\right]$$

$\Rightarrow$

$$\mathbb{E}\left[y\cdot\hat{f}(x)|S\right] = f(x)\cdot\mathbb{E}\left[\hat{f}(x)|S\right]$$

Replacing this in (4) we have that:

$$\mathbb{E}\left[(y-\hat{f}(x))^2|S\right] = f(x)^2 + \sigma^2 - 2\cdot f(x)\cdot\mathbb{E}\left[\hat{f}(x)|S\right] + \mathbb{V}\left[\hat{f}(x)|S\right] + \left(\mathbb{E}\left[\hat{f}(x)|S\right]\right)^2$$

$$= \left[f(x)^2 - 2\cdot f(x)\cdot\mathbb{E}\left[\hat{f}(x)|S\right] + \left(\mathbb{E}\left[\hat{f}(x)|S\right]\right)^2\right] + \mathbb{V}\left[\hat{f}(x)|S\right] + \sigma^2$$

$$= \left(f(x) - \mathbb{E}\left[\hat{f}(x)|S\right]\right)^2 + \mathbb{V}\left[\hat{f}(x)|S\right] + \sigma^2$$

$\Rightarrow$

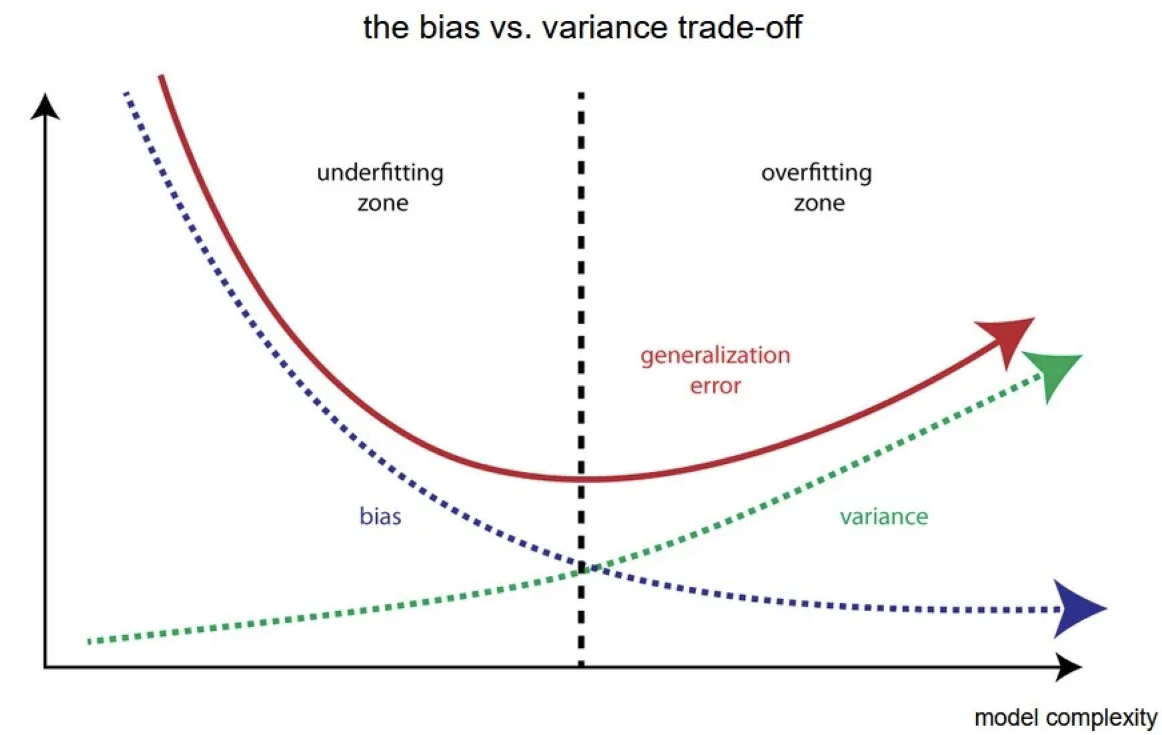
$$\mathbb{E}\left[(y-\hat{f}(x))^2|S\right] = \left(f(x) - \mathbb{E}\left[\hat{f}(x)|S\right]\right)^2 + \mathbb{V}\left[\hat{f}(x)|S\right] + \sigma^2$$

$\Rightarrow$

$$\mathbb{E}\left[(y-\hat{f}(x))^2|S\right] = bias^2 + variance + \sigma^2$$

We can consider:

- $bias$ = error in training set.
- $variance$ = error in test set.



How to solve underfitting?

- Try more complex models.
- Increase the number of parameters.
- Get more data. Because its possible that we don't have the enough data to fit a more complex model because if we do it we will get a overfitting model.

How to solve overfitting?

- Try more simple models.
- Reduce the number of parameters.
- Check if the problem is imbalanced (classification or regression).
- Get more data, because an option is the model is overfitting because the data size is small.

As you can see, the data size is FUNDAMENTAL both to underfitting and overfitting.
