In the lecture, we intuitively defined the structural and estimation mistakes. Here we provide a more formal definition of these ideas for a regression problem.

Suppose we want to learn the relationship between random variables $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{d}}$ and $\boldsymbol{y} \in \mathbb{R}$, where the ground truth relationship is $\boldsymbol{y}=\boldsymbol{f}(\boldsymbol{x})$. Of course $\boldsymbol{f}$ is unknown to us, but we observe a training set $\left(\boldsymbol{x}_{\mathbf{1}}, \boldsymbol{y}_{\mathbf{1}}\right)$, $\left(\boldsymbol{x}_{\mathbf{2}}, \boldsymbol{y}_{\mathbf{2}}\right)$, ..., $\left(\boldsymbol{x}_{\boldsymbol{n}}, \boldsymbol{y}_{\boldsymbol{n}}\right)$ and we hope to find a function $\hat{\boldsymbol{f}}$ to approximate the true funtion $\boldsymbol{f}$.

However, our observed data might not be $\mathbf{100 \%}$ accurate as there can be many kinds of noise and uncertainty containing in the data. Thus, we further assume a random noise variable $\boldsymbol{\epsilon}$ is added on top of $\boldsymbol{y}$:

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

where $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\sigma}^{\mathbf{2}}\right)$

We learned from the lecture that we can find $\hat{\boldsymbol{f}}$ by minimizing the empirical risk on the training set. As we know the training set is a random observation of the underlying relationship and contains noise, different training set will give us different estimator $\hat{\boldsymbol{f}}\ (\boldsymbol{x})$. Hence, we can define $\mathbb{E}\left[\hat{\boldsymbol{f}}\ (\boldsymbol{x})\right]$ to be the expected estimator over all possible training sets.

Now let's look at when we have a new $\boldsymbol{x}$ with unknown $\boldsymbol{y}$, what is the expected prediction error looks like for our estimator given all possible training sets:

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

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

We skip some derivations of this result, can you get this result on your own?

As we can see, there are three terms in this error decomposition:

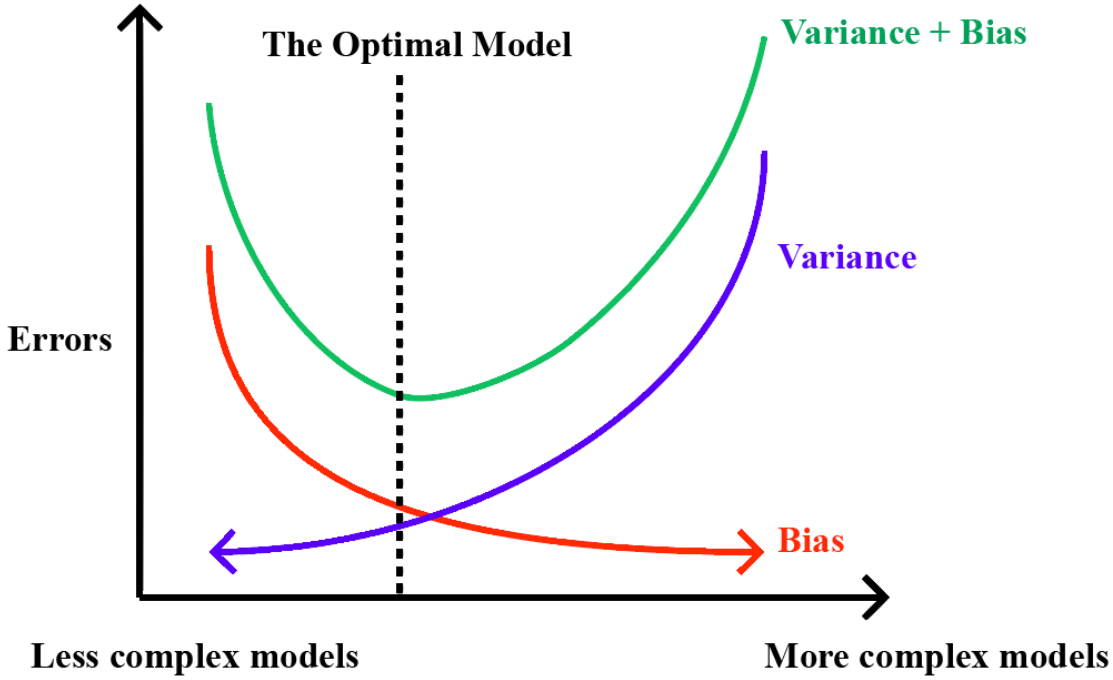
The first term is the square of the difference between the true $f\left(x\right)$ and the expected estimation over all possible training sets. This term is usually called **bias** and it describes how much the average estimator fitted over all datasets deviates from the underlying true $f\left(x\right)$. This corresponds to the structural mistake in the lecture.

The second term is the variance of the estimator (recall variance definition in statistics). It describes on average how much a single estimator deviates from the expected estimator over all data sets. This corresponds to the estimation error in the lecture.

The third term $\mathbb{E}\left[\epsilon^2\right] = \sigma^2$ is the error from the inherent noise of the data and we can do nothing to minimize it, thus it is sometimes called irreducible error. The irreducible error gives a lower bound on the expected prediction error.

Question for you to think: Apparently, the empricial risk is just an approximation of the true risk (i.e. $\mathbb{E}\left[\left(f\left(x\right) - \hat{f}\left(x\right)\right)^2\right]$). If we were able to learn the model by minimizing the true risk, which part of the error decomposition will become 0?

The task of supervised learning is to reduce the bias and variance at the same time, but because of the noise in the training data, it is not possible to simultaneously minimize these two sources of errors. This is known as the bias-variance trade-off.



To reduce bias, we can assume a more complex hypothesis space and fit a more powerful model. The model will be able to fit even the noise in the training set. However, this increases the error from variance because given another training set, the randomness of the noise will results in a very different model. This situation is often called 'overfitting'. On the other hand, we can have a more simple model to reduce the variance, but this can make the bias very large. For example, in the extreme case, let our model be $\hat{f}(x) = c$, where c is a constant. This will give us 0 variance but you can imagine it can hardly make any correct predictions. This situation is known as 'underfitting'.

In a few pages, you will see how regularization can be used to restrict the complexity for linear regression so that we will be able to search for a sweet spot where the total error from variance and bias is the minimum.

Hide