Given a variety of different models, for example : linear models, nearest neighbors, thin plates splines, we need to know to how to decide among these models. So, we need a way for assess a model's accuracy i.e. when is the chosen model accurate?

1 Assessing Model Accuracy

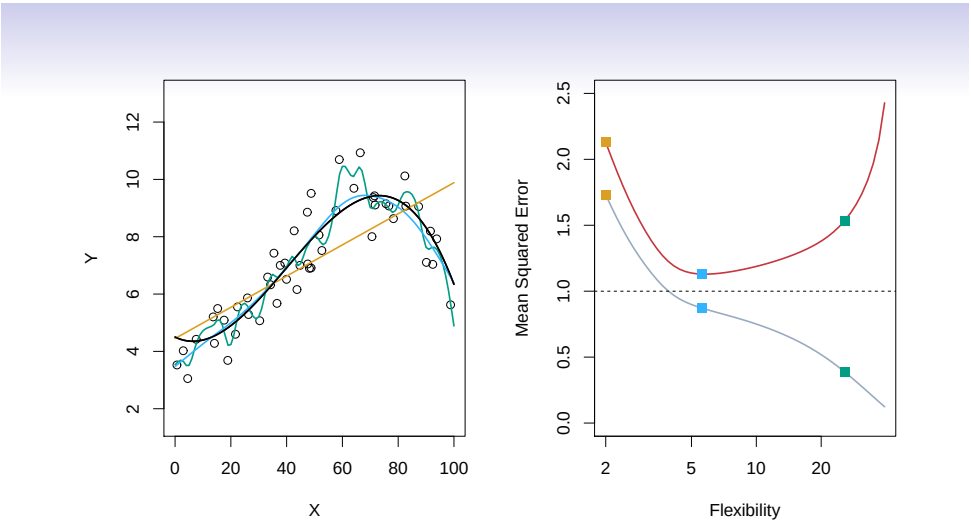
Suppose we have a model $\hat{f}_{(x)}$ that's been fit to some training data and we will denote the training data by $Tr = \{x_i, y_i\}_1^N$ this contains N data pairs of x_i, y_i. Remember that the x_i, y_i means the ith observation, where x_i may be a vector with different components (which we call features) and y_i are single label (usually a scalar). We want to see how well this model performs.

We could compute the average squared prediction error over our training dataset, Tr.

$$MSE_{Tr} = Average_{i \in Tr}[y_i - \hat{f}_{(x_i)}]^2$$

This means that we take our observed value i.e. y_i, we subtract from modeled value i.e. $\hat{f}_{(x_i)}$, we square the difference to get rid of the sign and finally we average them over the training data. As you may imagine, this may be biased towards more over-fit models (We can make the above error exactly equal to zero by fitting a highly flexible model). Instead, if possible, we should have computed the same using the fresh test dataset, $Te = \{x_i, y_i\}_1^M$ with additional M data pairs of x_i and y_i different from the training set. Then we compute the following, this might be a better refelection of the performance of our chosen model.

$$MSE_{Te} = Average_{i \in Te}[y_i - \hat{f}_{(x_i)}]^2$$



Black curve is truth. Red curve on right is MSE_{Te}, grey curve is MSE_{Tr}. Orange, blue and green curves/squares correspond to fits of different flexibility.

The above plot (left) shows three different models fitted to a given one dimensional data set. The orange model is a linear model (low complexity), the blue model is more flexible model some kind of spline (medium complexity) and the green is highly flexible function (high complexity) that fits the one dimensional data very closely. Since this is a simulated example, we can calculate the mean squared error on very large population of test data (represented in the right side of the above plot).

The red curve (testing error) value for a very rigid curve is high. The value drops down and becomes quite low for the blue model and rises up again for the green model. Where as the mean square error on the gray curve just keeps on decreasing because more flexible the model, the more close it gets to the data points. Therefore, there exists an optimal point that minimizes the MES_{Te}.

NOTE : The above is the simulated model. The dotted line is the error that the true function makes from the data from this population. This is the irreducible error, which we call the $Variance(\epsilon)$.

2 Bias - Variance Trade-off

Suppose we have fit a model $\hat{f}_{(x)}$ to some training data Tr, and let (x_0, y_0) be a test observation drawn from the population. If the true model is $Y = f_{(x)} + \epsilon$ (with $f_{(x)} = E(Y|X = x)$), then the expected prediction error between $\hat{f}_{(x_0)}$ and y_0 is given by

$$E(y_0 - \hat{f}_{(x_0)})^2 = Var(\hat{f}_{(x_0)}) + [Bias(\hat{f}_{(x_0)})]^2 + Var(\epsilon)$$

$$Bias(\hat{f}_{(x_0)}) = E[\hat{f}_{(x_0)}] - f_{(x_0)}$$

So, we can break this equation into three terms. $Var(\epsilon)$ is the irreducible error that comes from the random variation around the true function $f_{(x)}$. The other two pieces consists of reducible part of the error. $Var(\hat{f}_{(x_0)})$ is the variance that comes from having different training sets. For example, if I have a different set of training data then I would end up having different function $\hat{f}_{(x_0)}$. So, this accounts for the variability in the prediction with x_0 if I am looking at different sets of training sets. $Bias(\hat{f}_{(x_0)})$ is the difference between average prediction at x_0 (averaged over all the prediction with different training sets) and the true function value of x_0.

Typically as the ***flexibility*** of $\hat{f}$ increases, its variance increases, and its bias decreases. So, choosing the flexibility based on average test error amounts to a ***bias-variance trade-off***.