Assessing-Model-Accuracy

The most important thing in Statistical Learning is choosing the right method for you data set, Its depends on these concepts:

Measuring the quality of fit

- A way to ensure how well its predictions actually match the Observed data
- How close is the predicted Response value for a given Observation to the real true Response
  In Linear Regression most commonly used :

$$\text{Mean squard error} = \text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

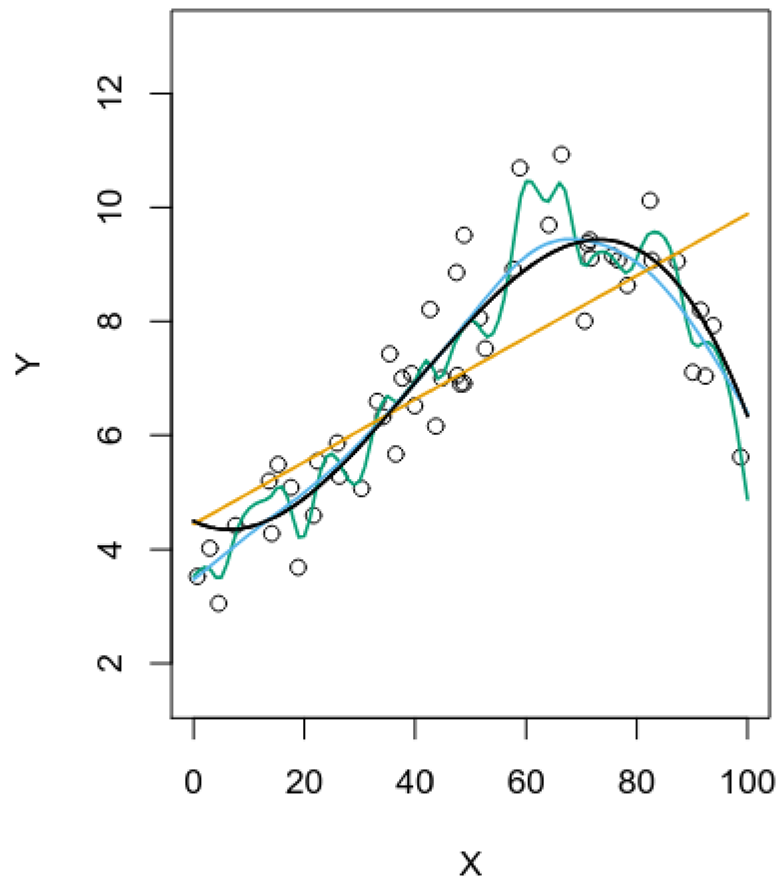
- $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the ith Observation
- $y_i - \hat{f}(x_i)$ how far is the prediction from the real Response we seek

MSE is computed using Training Data --> referred to Training MSE

- How well the model perform on it isn't important but the model accuracy on new **unseen data**
  After computing $\hat{f}(x_1)\ldots\hat{f}(x_n)$ we can about how $\hat{f}(x_i) \approx y_0$
- y_0 is the prediction for the **unseen data** (x_0, y_0)
- we want the method that gives the lowest $test\ MSE$
- $\text{Avg}(y_0 - \hat{f}(x_0))^2$ as small as possible (degrees of freedom)

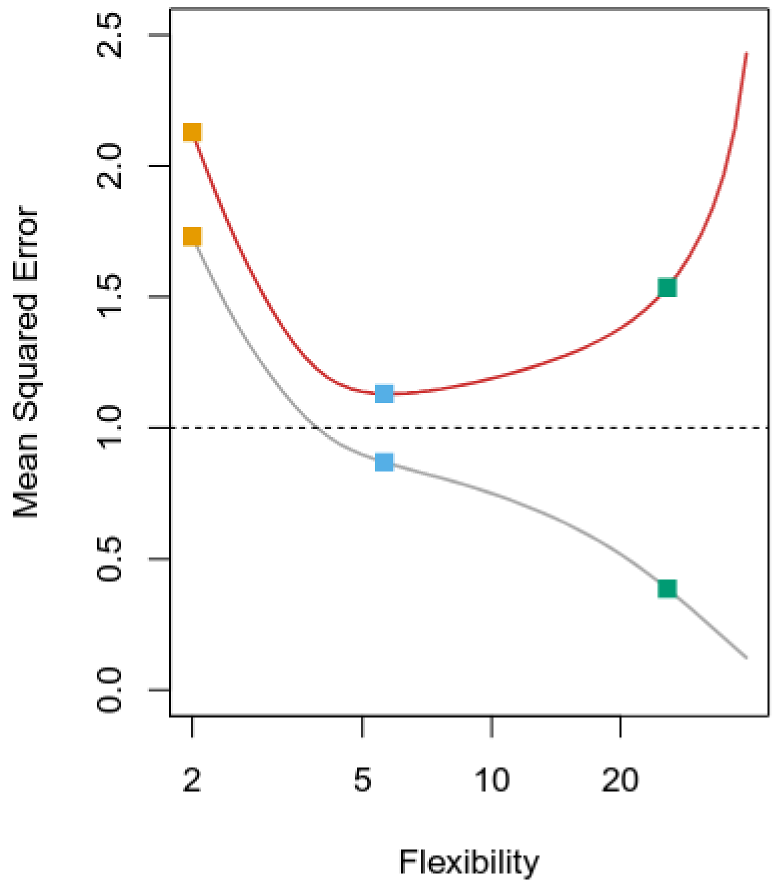
Note : $\text{Training MSE} < \text{Test MSE}$

Data from f



- Each circle is an Observation
- Black curve -> f real form
- All other curves are estimate of $f \approx \hat{f}$

Training vs Test MSE:



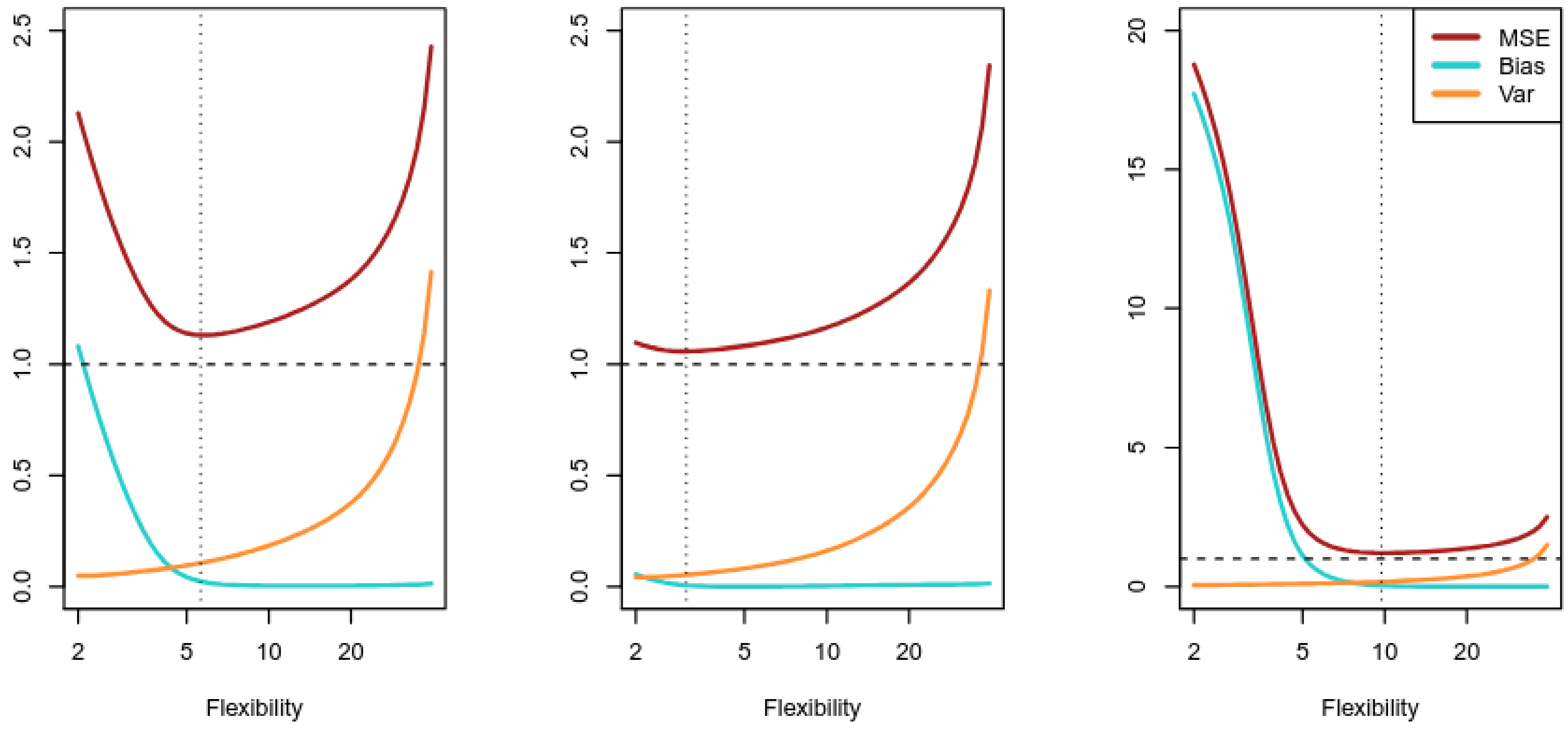
- Training MSE -> grey
- Test MSE -> Red
- Dashed line is the irreducible error $Var(\varepsilon$
  We notice that Test MSE is always bigger than the Training MSE, and that the more Flexible the model is better Test MSE results are till it reach a point increasing the model **Flexibility** will result spike on the Test MSE even tho it performed the best on the Training data, what happened is [Overfitting]
- We mostly estimate Test MSE and its more difficult most of the time no [Test data] is available, One important method is [cross validation] to estimate the Test MSE

The Bias-Variance Trade-off

- The U-shape curve on the Test MSE is the result two properties
  1. Variance of $\hat{f}(x_0)$
  2. Squared bias of $\hat{f}(x_0)$
  3. Variance of error $Var(\varepsilon)$ -> irreducible

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

- $E(y_0 - \hat{f}(x_0))^2$ expected Test MSE at x_0
- x_0 unseen [Observation]
  To minimize the expected Test MSE -> we need a method/model that minimize the $Variance$ and the Bias of $\hat{f}$
- The $Variance$ of $\hat{f}$ refers the amount of change in $\hat{f}$ if we estimated by different data set
  - different data sets will give different $\hat{f}$
  - Higher $Variance$ in $\hat{f} \implies$ The smallest change in the data set will result in a very different estimate of $\hat{f}$
  - $\hat{f}$ shouldn't vary much between different data sets
  - More flexible methods $\implies$ Higher $Variance$
- The **Bias** of $\hat{f}$ refers to the error introduced by approximating a real life problem which is very **complex** into $\rightarrow$ Simple Models
  - For example Linear regression will assume there is a Linear relationship between Y and X, its unlikely any real life problem is like that
  - More flexible methods $\implies$ Lower Bias



- As we increase the flexibility of the model the **Bias** tends to initially decrease
  - Faster than the $Variance$
  - As a result of that The Test MSE decline on
- At some point increasing the flexibility got no effect on the **Bias**
- But the $Variance$ will increase significantly (Left and Center Figures)
- In Center Figure f we estimating is Linear in nature so increasing the flexibility got no effect on the **Bias**
- In the Right Figure f is very non-Linear that's why we notice a huge drop off in the **Bias** as the flexibility increase

In real life Computing The $\text{Test MSE}, \text{Bais}, \text{Variance}$ is impossible due to not knowing the real form of f so most of the time we estimate them.

The Classification setting

- When it comes to classification problems $y_1, \ldots y_n$ is a qualitative
- To quantify the accuracy of estimating $\hat{f} \rightarrow \text{Training error rate}$

$$\frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- $\hat{y}_i$ is the predicted category, class label for the ith Observation using $\hat{f}$
- $I(y_i \neq \hat{y}_i)$ indicator variable $\rightarrow 1$ if the prediction is wrong $\implies y_i \neq \hat{y}_i$
- and $\rightarrow 0$ if the prediction is correct $\implies y_i = \hat{y}_i$
- its only for the Training Data

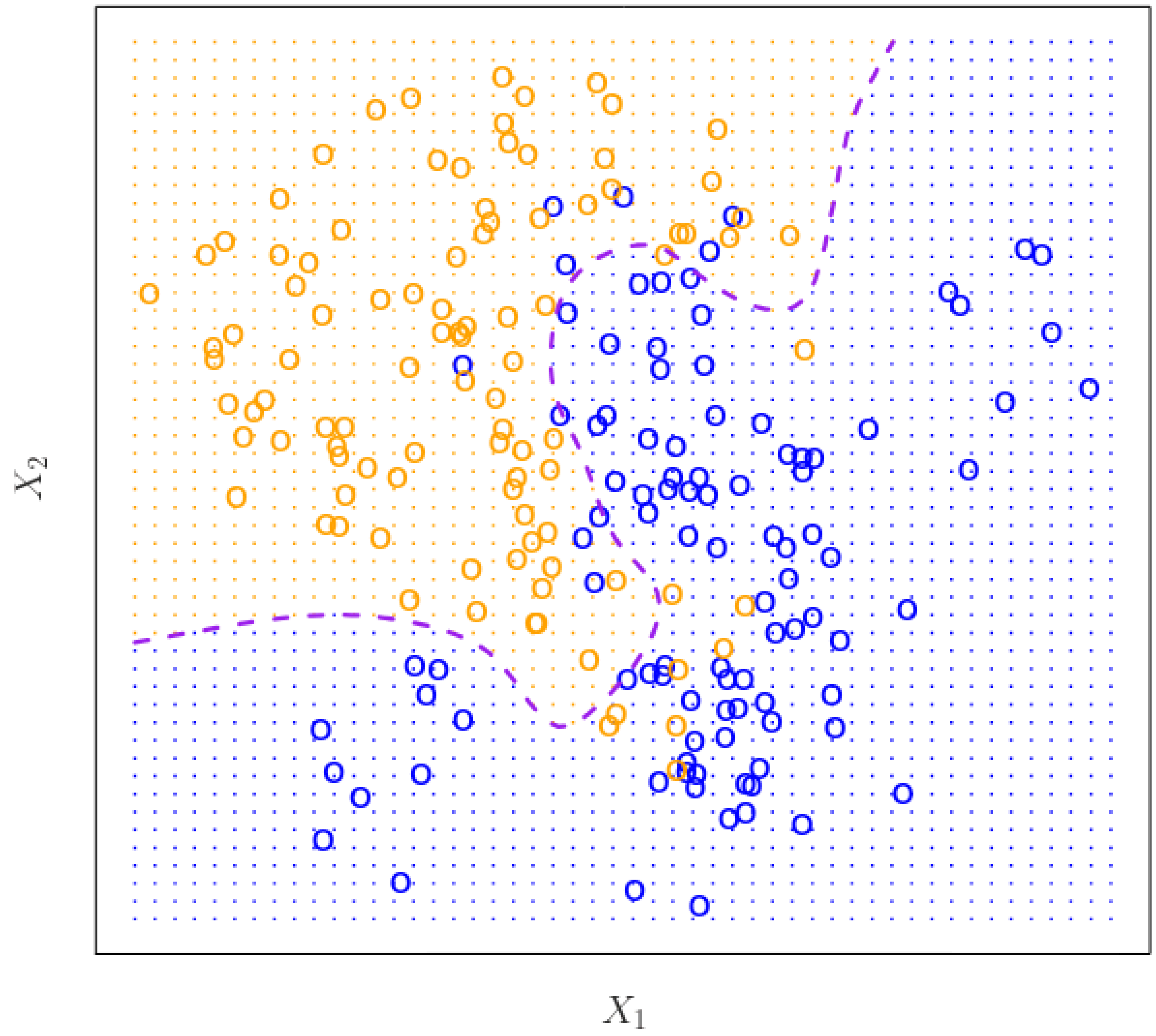
We interested in the **unseen data** error rate on the Observation (x_0, y_0)
and its called $test\ error$.

The Bayes Classifier

- Derived from the Bayes' theorem Bayes Classifier
- Assigns each Observation to the most likely class given the Predictor value $X = x_0$

$$Pr(Y = j | X = x_0)$$

- Bayes Classifier Predicts a class 1 for example if $Pr(Y=1|X=x_0) > 0.5$



- The orange shaded area the Probability $Pr(Y = orange|X) > 50\%$
- The Blue shaded area the Probability $Pr(Y = blue|X) > 50\%$
- The Purple line represent the points where Probability is exactly $50\% \to$ Bayes decision boundary
- The Bayes Classifier produces the lowest possible test error rate
- And always Pick the class with the highest value

The error rate will always be :

$$1 - E(max_j \Pr(Y = j|X))$$

- The Bayes Classifier is Used in K-Nearest Neighbors To predict the most likely class k for a given Observation