General overview of the game we play in machine learning.

- Nature generates data according to some function $f: x \to y$ we don't know
- In machine learning we want a good approximation of f on out of sample data.
- We start with set of possible functions(Hypothesis space H) and try to pick a function $\hat{f}$ best approximating true f(which we don't know)
  - Complex H means good chance of approximating f but hard to find approximation?
  - Simple H means good generalization on out of sample data.

Bias variance analysis effectively asks following questions.

- How well H can approximate f overall.
- Given the sample how good is our approximation.

Let's use square error analysis on linear regression Let's say f is true function and $\hat(f)$ is our choice of target function based on sample dataset D in hypothesis space H Then

$E_{out}(\hat{f}_{\mathrm{D}}) = \mathrm{E}_x(\hat{f}_{\mathrm{D}}(x) - f(x))^2$

This depends on a particular sample D. If we take another same number of sample this will change. To do error anlysis we need to get rid of a particular sample D choice.

Let's take expectation with respect to distriution over D

$$\begin{aligned} \mathrm{E}_{\mathrm{D}}(E_{out}(\hat{f}_{\mathrm{D}})) &= \mathrm{E}_{\mathrm{D}}\mathrm{E}_x(\hat{f}_{\mathrm{D}}(x) - f(x))^2 \\ &= \mathrm{E}_x\mathrm{E}_{\mathrm{D}}(\hat{f}_{\mathrm{D}}(x) - f(x))^2 \end{aligned}$$

Define average hypothesis

$$\bar{f} = \mathrm{E}_{\mathrm{D}}(\hat{f}_{\mathrm{D}}(x))$$

$$\begin{aligned} \mathrm{E}_{\mathrm{D}}(\hat{f}_{\mathrm{D}}(x) - f(x))^2 &= \mathrm{E}_{\mathrm{D}}[\hat{f}_{\mathrm{D}}(x) - \bar{f}(x) + \bar{f}(x) - f(x)]^2 \\ &= \mathrm{E}_{\mathrm{D}}[(\hat{f}_{\mathrm{D}}(x) - \bar{f}(x))^2 + (\bar{f}(x) - f(x))^2) + 2(\hat{f}_{\mathrm{D}}(x) - \bar{f}(x))(\bar{f}(x) - f(x))] \\ &= \mathrm{E}_{\mathrm{D}}[(\hat{f}_{\mathrm{D}}(x) - \bar{f}(x))^2] + [\bar{f}(x) - f(x)]^2 \end{aligned}$$

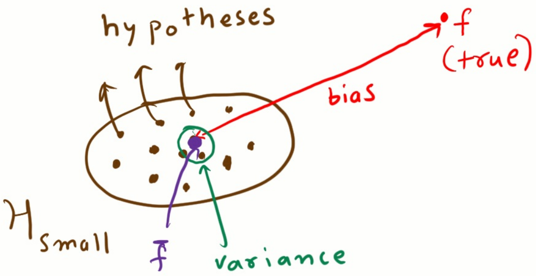
$$\underbrace{\mathrm{E}_{\mathrm{D}}}_{\text{variance(x)}}[(\hat{f}_{\mathrm{D}}(x) - \bar{f}(x))^2] + [\underbrace{\bar{f}}_{\text{bias(x)}}(x) - f(x)]^2$$

bias = a measure of best we can do in our hypothesis set

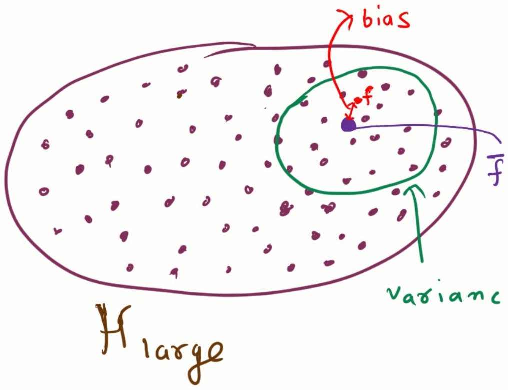
variance = how far our hypothesis based on D is away from average

$$\begin{aligned} \mathrm{E}_x\mathrm{E}_{\mathrm{D}}(\hat{f}_{\mathrm{D}}(x) - f(x))^2 &= \mathrm{E}_x[\mathrm{E}_{\mathrm{D}}[(\hat{f}_{\mathrm{D}}(x) - \bar{f}(x))^2] + [\bar{f}(x) - f(x)]^2] \\ &= \mathrm{E}_x[bias^2(x) + variance(x)] \\ &= bias^2 + variance \end{aligned}$$

BIAS VARIANCE TRADEOFF

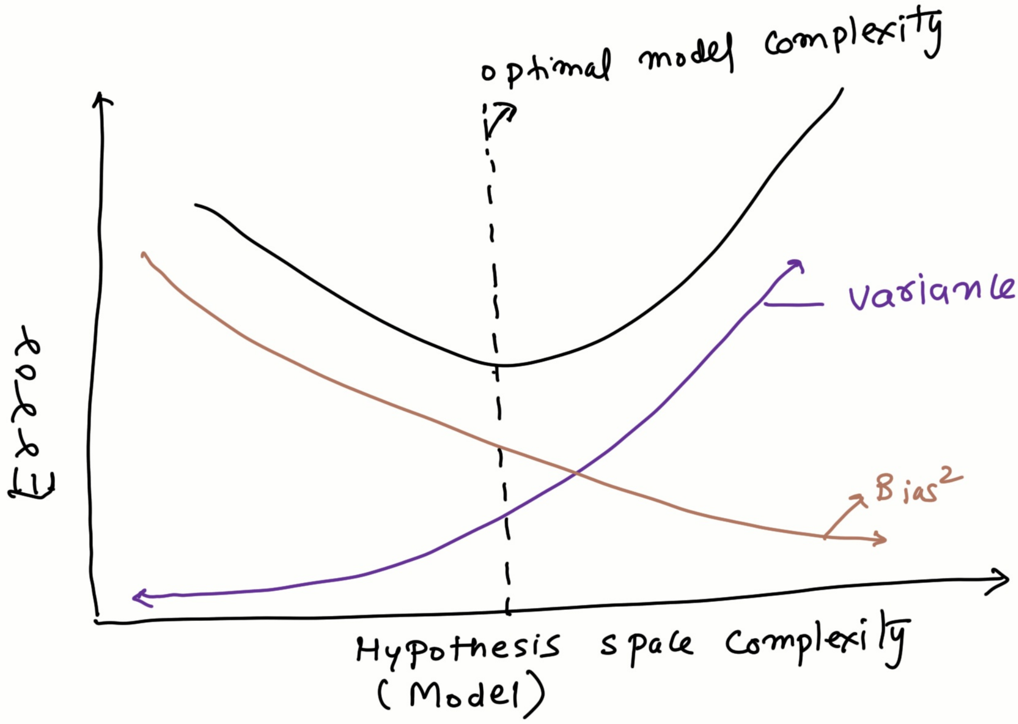


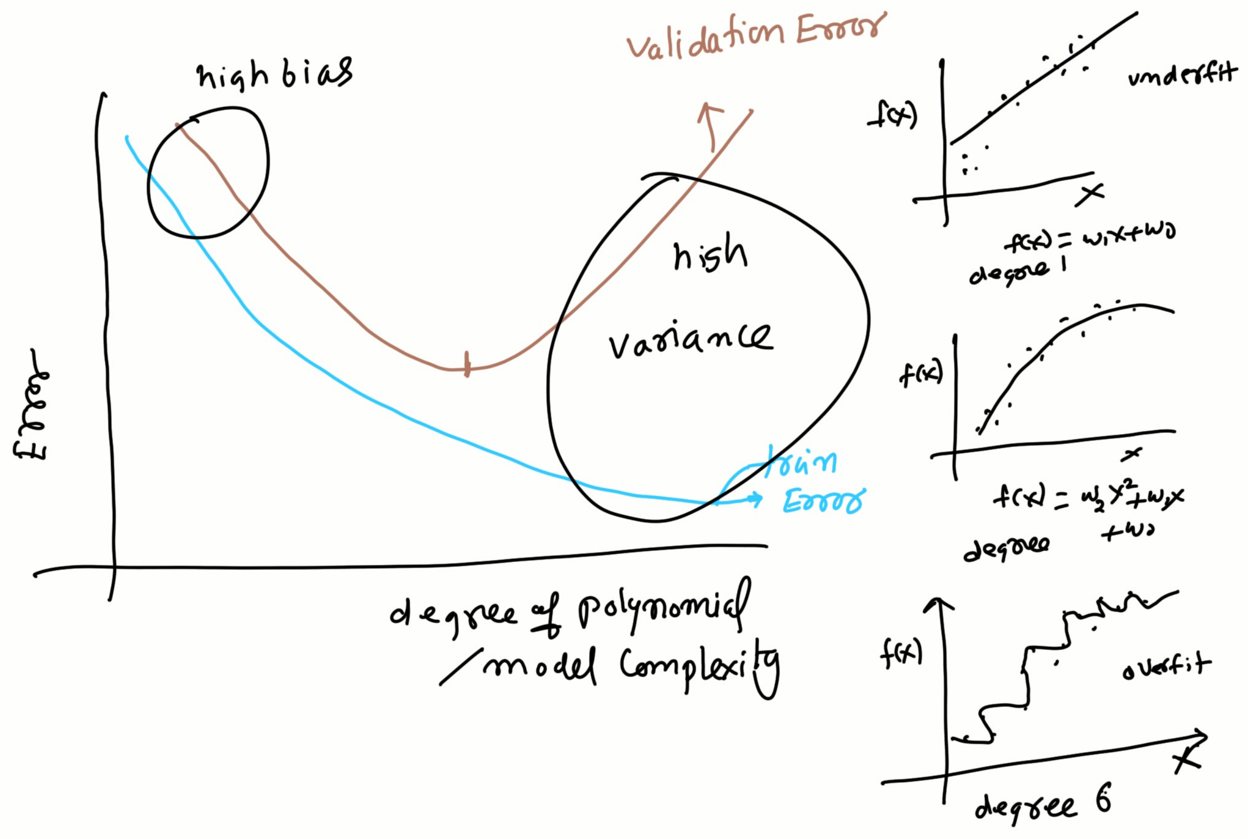
- High bias
- Low variance



- Low bias
- High variance

Bias Variance Trade off





Evaluation of Learning algorithms

We need a way to measure how well learning algorithm perfroms. Till now we have seen follwoing measures

- Accuracy
- Precision, Recall, ROC curve

let's make it more formal

- Given an Hypothesis space H, using training data D we select a hypothesis $\hat{f} \in \mathrm{H}$
- Given a feature vector x we predict using h as $\hat{y} = \hat{f}(x)$. Let say true value is y.

For Regresssion

- MAE (Mean absolute error) $\frac{\sum_i^N |\hat{f}(x_i) - y_i|}{N}$
- MSE (Mean square error) $\frac{\sum_i^N \|\hat{f}(x_i) - y_i\|^2}{N}$ ## For Classification
  - Miss classification error = $\frac{\sum_i^N \mathrm{I}(\hat{f}(x_i) \neq y_i)}{N}$
  - Build Confusion matrix, Remember TP, TN, FP, FN for binary calssification. You can build confusion matrix for K class classification too.
  - ROC

Confusion Matrix for a particular value of threshold τ

	$y = 1$	$y = 0$	
$\hat{y}$=1	TP	FP	$\hat{N}_+$ = TP + FN
$\hat{y}$=0	FN	TN	$\hat{N}_-$ = TN + FN
-------------	:-------------:	-----:	:-----
	N_+ = TP + FN	N_- = TN + FP	N= TP+FN +TN+FP

precision = $\frac{TP}{TP+FP}$

recall = True positive rate(**TPR**) = $\frac{TP}{N_+} \approx p(\hat{y} = 1 \mid y = 1)$

false positive rate (**FPR**)(type I error rate) = $\frac{FP}{N_-} \approx p(\hat{y} = 1 \mid y = 0)$

Sample error vs True Error(In classification setting)

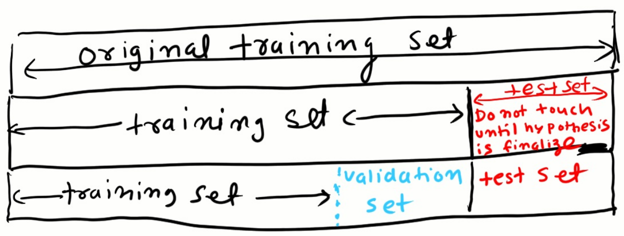
Sample error : The sample error of a hypothesis $\hat{f}(x)$ with respect to true function(target) f on a sample $\mathrm{D} = \{x_i, y_i = f(x_i)\}_{i=1}^{i=N}$ is miss classification error $\frac{\sum_i^N \mathrm{I}(\hat{f}(x_i) \neq f(x_i))}{N}$

True Error The true error of a hypothesis $\hat{f}(x)$ with respect to true function(target) f and distribution P, is the probability of mis classifying a random sample from distrubtion P, $Pr_{x \sim P}[\hat{f}(x) \neq f(x)]$

We need good **True Error**

Evaluation of hypothesis under limited sample data

- Split the samples into train and test



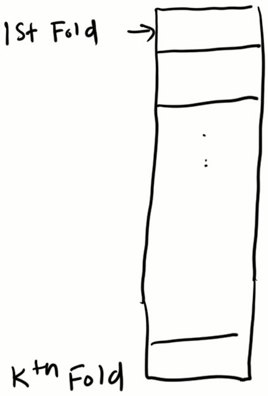
What if train set is too small

- we have small samples size.
- Can't touch test set
- Can we use data in validation for training too?

yes. Procedure is calls k -fold cross validation

K - Fold Cross Validation

- partition data into K subset



- For i in $\{1, 2 \cdots K\}$
  + use Fold i for validation
  + use rest of the Folds for training
  + calculate $Error_i$ on Fold i

$$\text{Final validation Error} = \sum_{i=1}^{K} \frac{Error_i}{K}$$