2695 Introduction to Machine Learning
Masters Program in Economics, Finance and Management

NOVA SCHOOL OF BUSINESS & ECONOMICS

3d

ERROR IN MODELS

Cause of error

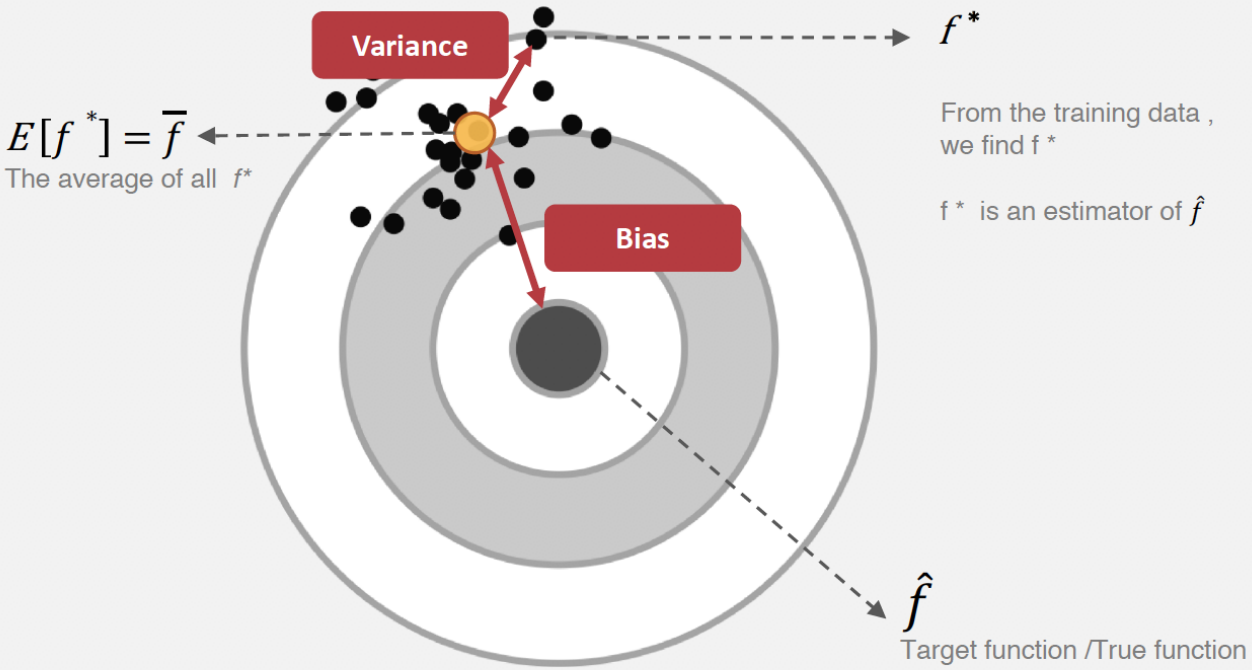
The errors a model makes can be characterized by three factors:

- **Inherent randomness (irreducible error):** due to this inherent probabilistic nature of the problem (e.g., we simply do not always get the same value for the target variable every time we see the same set of features).

- **Bias:** due to a systematic error that occurs in the machine learning model itself due to incorrect assumptions in the ML process (e.g., using simple linear model for quadratic data).

- **Variance:** due to limited training data (e.g., training a model with different subsets of data may lead to slightly different models).

Decompose the error



- **The variance** of the learning method represents how much the learning method will move around its mean; The variance is error from sensitivity to small fluctuations in the training set[3].
- **The Bias** of learning method represents the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict.

Expected predicted error = **Bias²** + **Variance** + Irreducible error

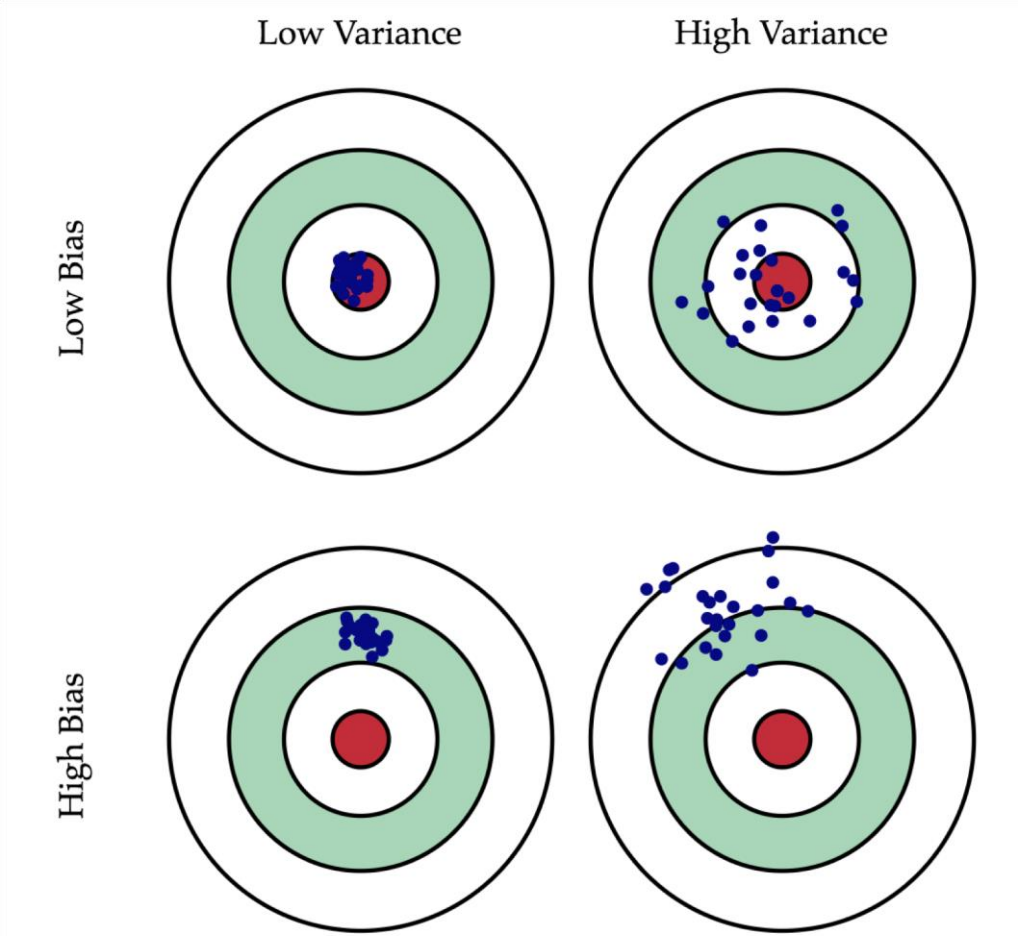
The prediction errors can be decomposed into two main subcomponents we care about:

- error due to "bias"
- error due to "variance" [1]

Irreducible error, is the noise term that cannot fundamentally be reduced by any model. Given the true model and infinite data to calibrate it, we should be able to reduce both the bias and variance terms to 0. However, in a world with imperfect models and finite data, there is a tradeoff between minimizing the bias and minimizing the variance[2].

http://scott.fortmann-roe.com/docs/BiasVariance.html

Bias vs Variance



- Low Bias & Low Variance -> Good case
- Low Bias & High Variance
- High Bias & Low Variance
- High Bias & High Variance -> Bad case

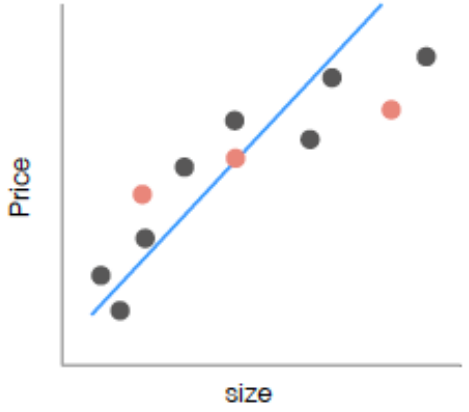
Model Selection

- There is usually a trade-off between Bias and Variance.
- Select a model that balances two kinds of error to minimize the total error.

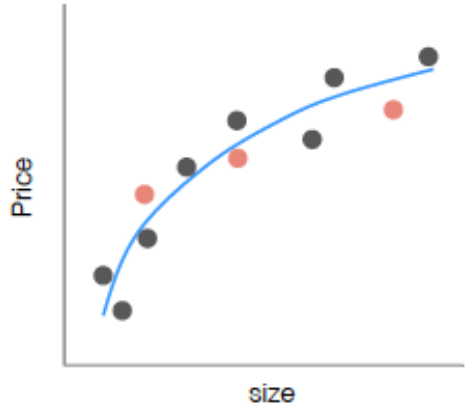
http://scott.fortmann-roe.com/docs/BiasVariance.html

Examples: Underfitting vs Overfitting

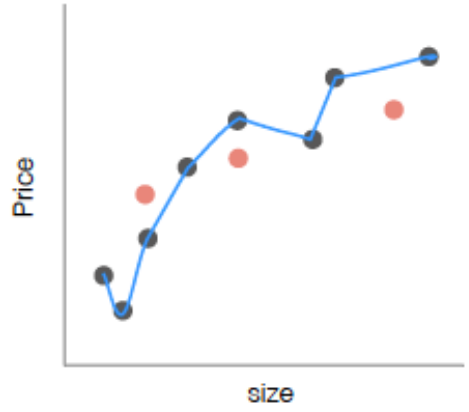
Regression example



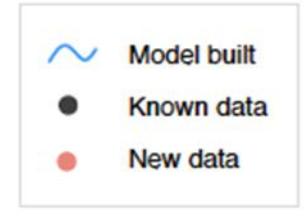
Under fitting model
(high training error, high test error)



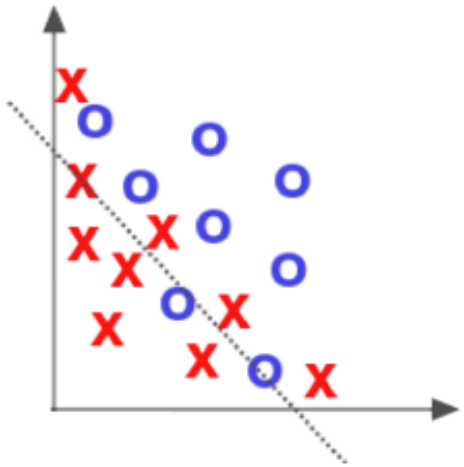
Good Fitting
(low training error, low test error)



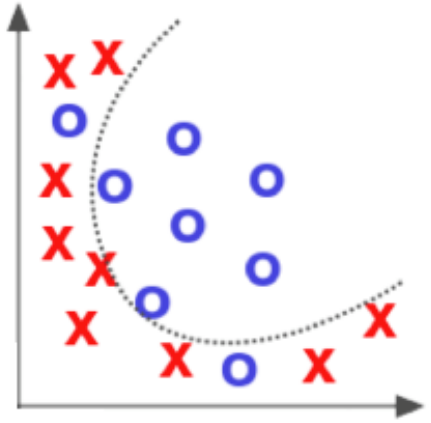
Overfitting Model
(no training error, high test error)



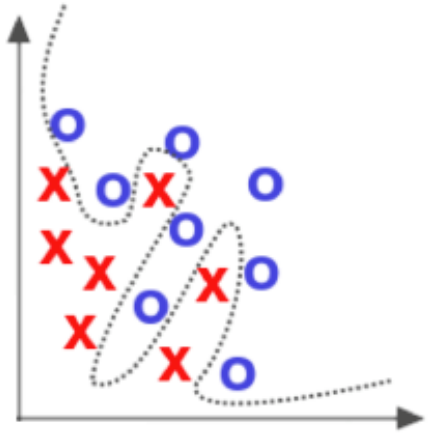
Classification example



Under fitting model
(high training error, high test error)



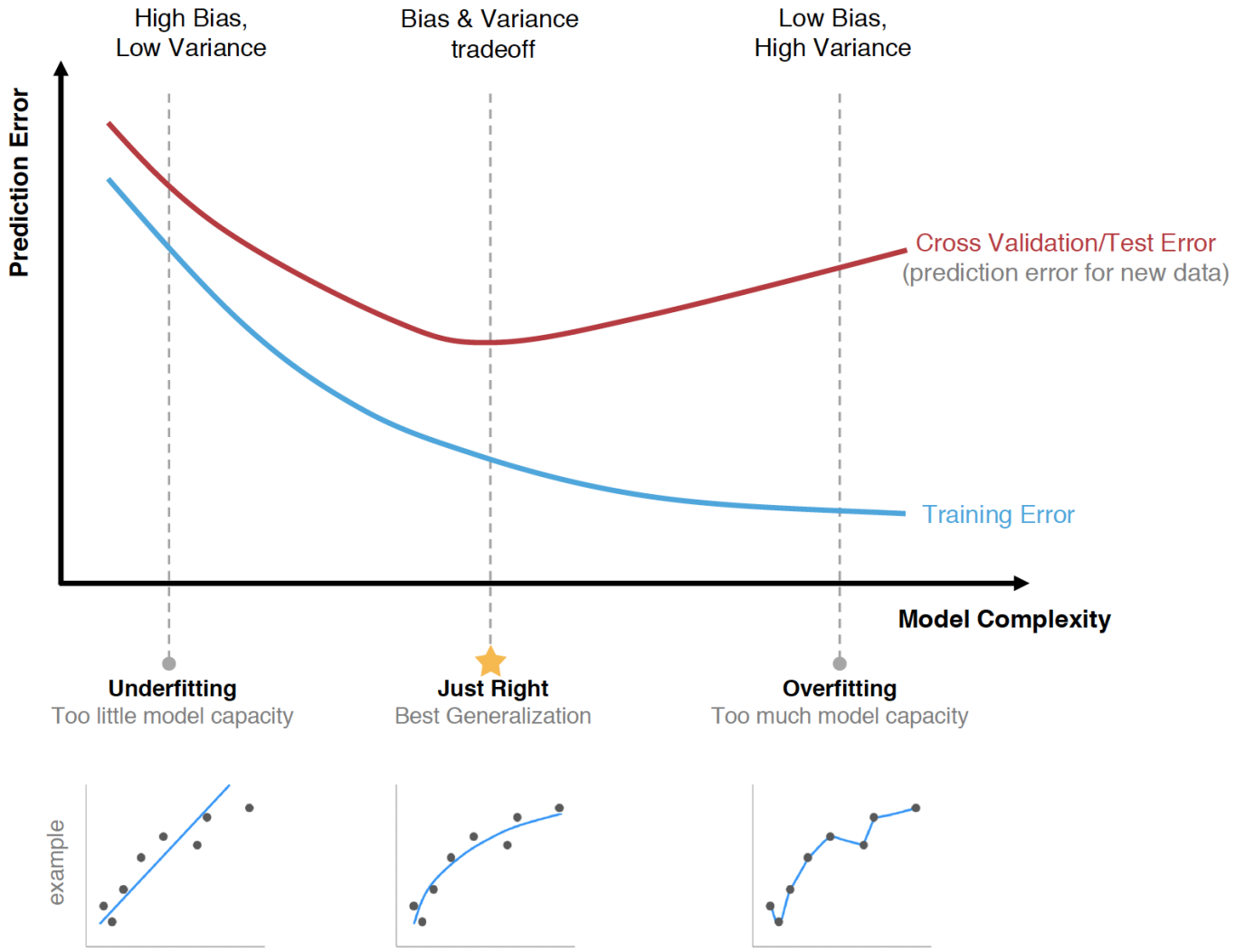
Good Fitting
(low training error, low test error)



Overfitting Model
(no training error, high test error)

Fitting graph: Bias and Variance Tradeoff

Fitting graph shows how the training and validation errors change as the model complexity increases



Underfitting = Bias problem
Overfitting = Variance problem

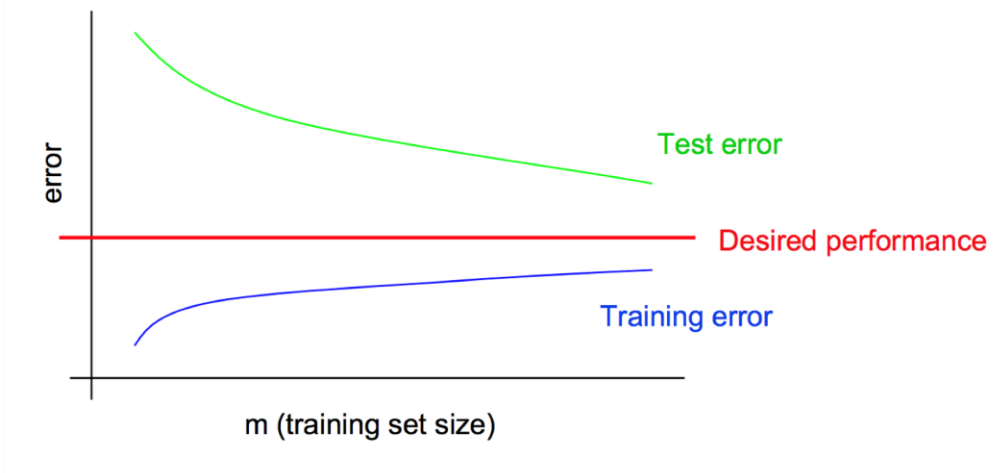
- **High bias**: Model is performing poorly on both training and test data sets
- **High variance**: Model is performing really well on the training, but not on the test set

Validation curve: shows how the training and validation performance change with the values of the model hyperparameter.

Learning curve

Learning curve: compares the performance of a model on training and testing data as the amount of training data increases. It helps determine whether adding more data will improve the model.

Typical learning curve for high variance



- Test error still decreasing as m increases. Suggests larger training set will help.
- Large gap between training and test error

Typical learning curve for high bias

- Even training error is unacceptably high.
- Poor generalization
- Small gap between training and test error