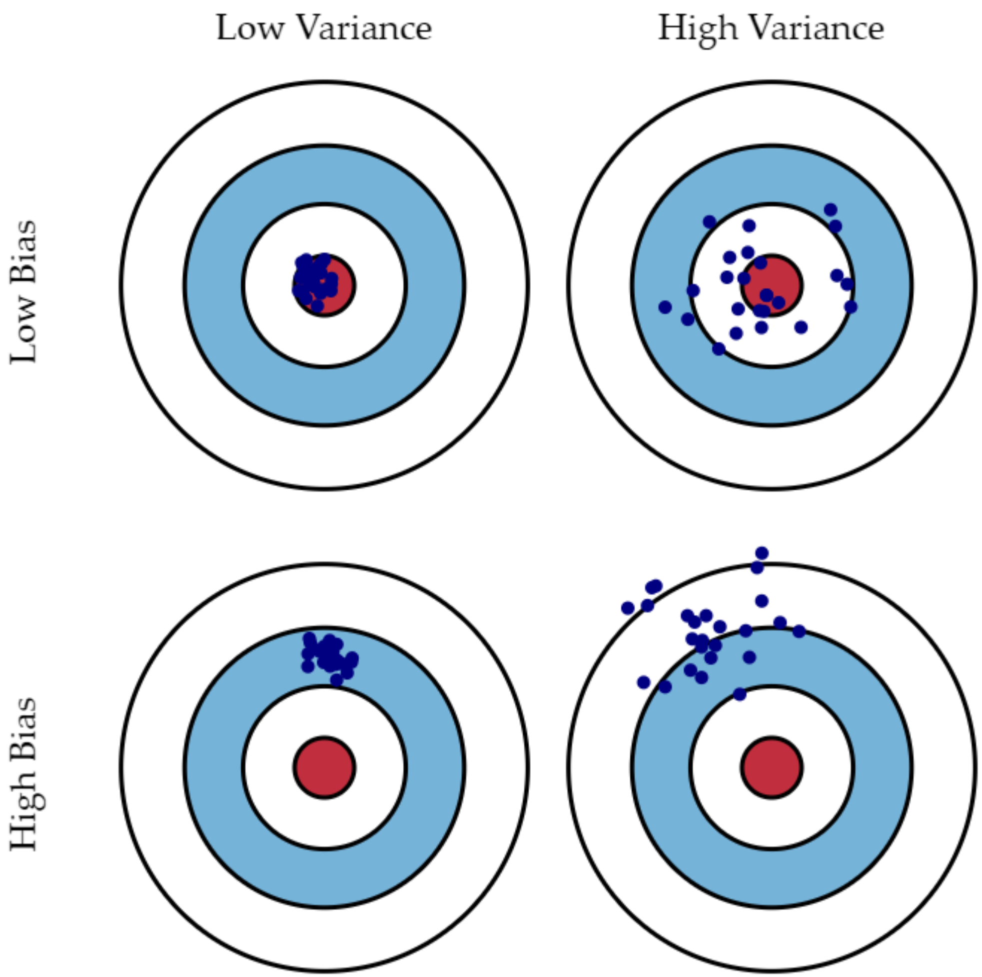
CPE/EE 695: Applied Machine Learning

Lecture 5-1: Bias-Variance Trade-offs

Dr. Shucheng Yu, Associate Professor
Department of Electrical and Computer Engineering
Stevens Institute of Technology

Prediction Errors: Bias-Variance (BV) Trade-off



Assuming the center (red circle) is the perfect model

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

Prediction Errors: Bias-Variance (BV) Trade-off

Statistically, define:

Y: target value (variable we want to predict)

X: input data (covariates)

$$Y = f(X) + \varepsilon$$

where $\varepsilon \sim N(0, \sigma_\varepsilon)$ is the error term.

$\hat{\boldsymbol{f}}(\boldsymbol{X})$: trained model of $f(X)$

The expected squared error at point x:

$$\boldsymbol{Err}(\boldsymbol{x}) = \boldsymbol{E}\left[\left(\boldsymbol{Y} - \hat{\boldsymbol{f}}(\boldsymbol{x})\right)^2\right] = \left(\boldsymbol{E}[\hat{\boldsymbol{f}}(\boldsymbol{x})] - \boldsymbol{f}(\boldsymbol{X})\right)^2 + \boldsymbol{E}\left[\left(\hat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{E}[\hat{\boldsymbol{f}}(\boldsymbol{x})]\right)^2\right] + \boldsymbol{\sigma_\varepsilon}^2$$

Prediction Errors: Bias-Variance (BV) Trade-off

Statistically, define:

Y: target value (variable we want to predict)

X: input data (covariates)

$$Y = f(X) + \varepsilon$$

where $\varepsilon \sim N(0, \sigma_\varepsilon)$ is the error term.

$\hat{\boldsymbol{f}}(\boldsymbol{X})$: trained model of $f(X)$

The expected squared error at point x:

$$\boldsymbol{Err}(\boldsymbol{x}) = \boldsymbol{E}\left[\left(\boldsymbol{Y} - \hat{\boldsymbol{f}}(\boldsymbol{x})\right)^2\right] = \boxed{(\boldsymbol{E}[\hat{\boldsymbol{f}}(\boldsymbol{x})] - \boldsymbol{f}(\boldsymbol{X}))^2} + \boldsymbol{E}\left[\left(\hat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{E}[\hat{\boldsymbol{f}}(\boldsymbol{x})]\right)^2\right] + \boldsymbol{\sigma_\varepsilon}^2$$

Bias

Prediction Errors: Bias-Variance (BV) Trade-off

Statistically, define:

Y: target value (variable we want to predict)

X: input data (covariates)

$$Y = f(X) + \varepsilon$$

where $\varepsilon \sim N(0, \sigma_\varepsilon)$ is the error term.

$\hat{\boldsymbol{f}}(\boldsymbol{X})$: trained model of $f(X)$

The expected squared error at point x:

$$\boldsymbol{Err}(\boldsymbol{x}) = \boldsymbol{E}\left[\left(\boldsymbol{Y} - \hat{\boldsymbol{f}}(\boldsymbol{x})\right)^2\right] = \underbrace{\boxed{\left(\boldsymbol{E}[\hat{\boldsymbol{f}}(\boldsymbol{x})] - \boldsymbol{f}(\boldsymbol{X})\right)^2}}_{\textbf{Bias}} + \underbrace{\boxed{\boldsymbol{E}\left[\left(\hat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{E}[\hat{\boldsymbol{f}}(\boldsymbol{x})]\right)^2\right]}}_{\textbf{Variance}} + \boldsymbol{\sigma_\varepsilon}^2$$

Prediction Errors: Bias-Variance (BV) Trade-off

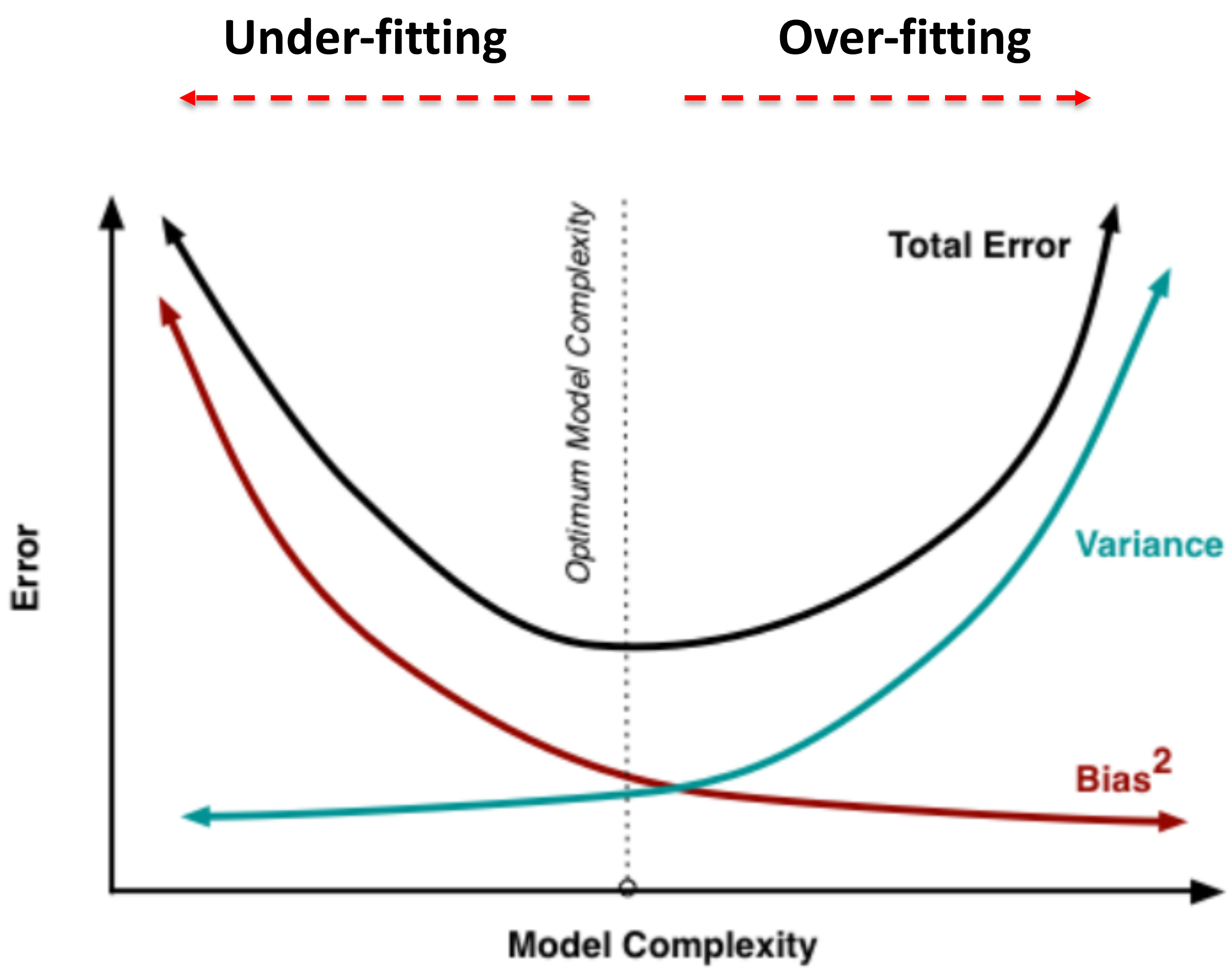
Bias Problem

The hypothesis space made available by a particular classification method does **not** include **sufficient** hypotheses

Variance Problem

The hypothesis space made available is **too large** for the training data, and the selected hypothesis may not be accurate on unseen data

Prediction Errors: Bias-Variance (BV) Trade-off



Prediction Errors: Bias-Variance (BV) Trade-off

- Practical techniques to reduce bias:

  Increase hypothesis space (i.e., model complexity)

- Practical techniques to reduce variance:

  - resampling (e.g., bagging models like random forest)

    * bias of each tree is the same as full model but higher variance

    * averaging many trees decreases variance without increasing bias

    * theoretically the more trees, the less variance (if no computation limit)

- No analytical methods to find optimal bias-variance trade-off

    * try different model complexity (needing accurate error measurement)