Bias-Variance Tradeoff

Prepared by John Carr
DC-DSI-3
12/14/16

What is bias? What is variance?

- Definitions from Scott Fortman-Roe:
  - Error due to Bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict.
    - Bias measures how far off in general these models' predictions are from the correct value.
  - Error due to Variance is taken as the variability of a model prediction for a given data point.
    - Variance is how much the predictions for a given point vary between different realizations of the model.

Graphical Demonstration

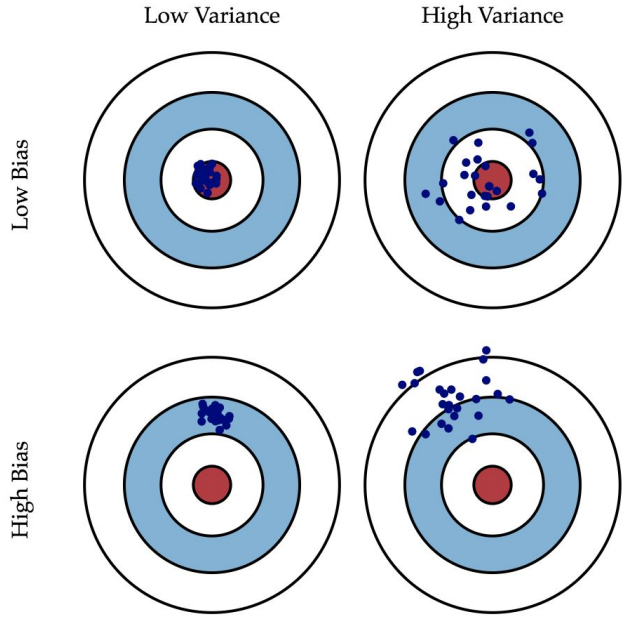


Fig. 1 Graphical illustration of bias and variance.

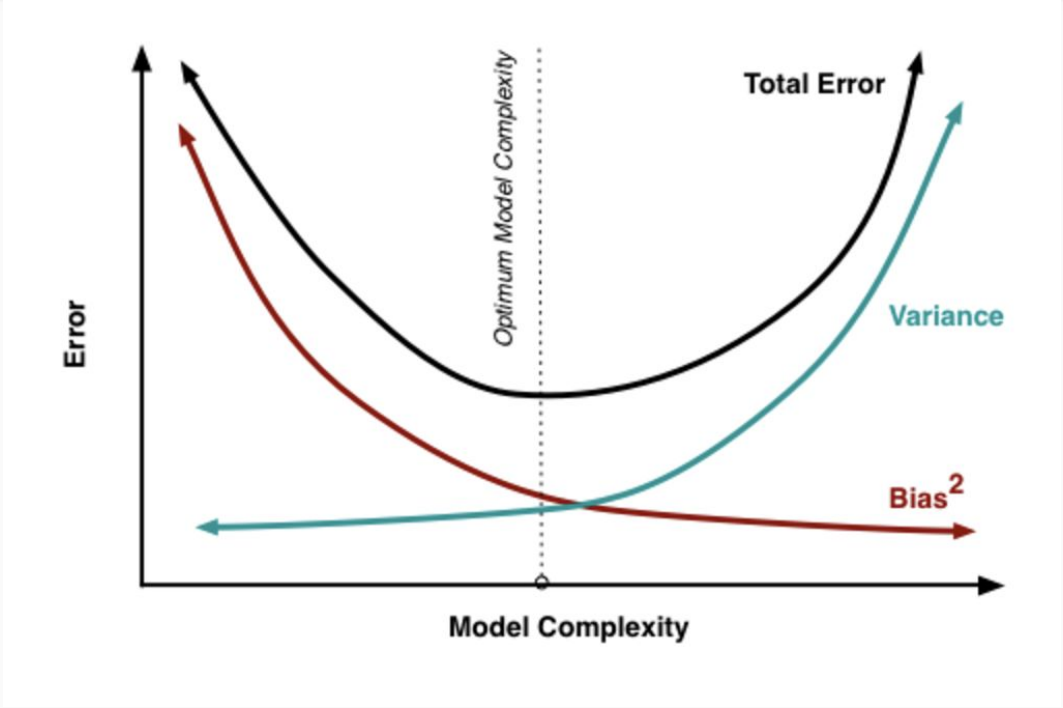
- Bullseye represents correctly predicted model - further away a point is, the less accurate the model's predictions

Source: Scott Fortman-Roe

Underfitting and Overfitting Models

- A highly biased model is underfit to the data.
  - A simpler model that does not as accurately predict data.
  - Predictions less accurate.
- A high variance model is overfit to the data.
  - Represent the training set more accurately, but are at risk of taking noisy or irrelevant data into account.
  - Predictions likely to vary greatly across testing sets.

What is the Ideal Model Complexity?



How to minimize?

- There is no analytical way to find the “sweet spot” that minimizes both bias and variance.
- The solution is to use an accurate measure of prediction error based on our model, explore different levels of complexity, then choose one that minimizes the overall error.
  - Use cross-validation when possible.
  - Use adjusted R^2 values, which account for increasing model complexity by penalizing a model for each additional variable.
  - Feature selection techniques.

Example 1 - Polling

- Goal - predict who voters will vote for
  - Sampling 50 people in a phonebook (Example from Scott Fortman-Roe)
- Potential sources of bias?
- Potential sources of variance?

Example 1 - Polling

- Goal - predict who voters will vote for
  - Sampling 50 people in a phonebook (Example from Scott Fortman-Roe)
- Potential sources of bias -
  - Limiting poll to members in a phonebook
  - Not following up with non-respondents
- Potential sources of variance -
  - Small sample size of all voters in an election

Example 2 - Health Insurance

- Goal - predict costs for enrolled members for upcoming year
  - Given data on all claims while enrolled
    - Procedures completed, diagnoses, costs, type of claim, place of service
- Potential sources of bias?
- Potential sources of variance?

Example 2 - Health Insurance

- Goal - predict costs for enrolled members for upcoming year
  - Given data on all claims while enrolled
    - Procedures completed, diagnoses, costs, type of claim, place of service
- Potential sources of bias -
  - Certain people more likely to go to the doctor/ER than others
    - Claims determine they are less healthy
  - Lab/biometric results are not legally usable
- Potential sources of variance -
  - High levels of turnover in enrollment
  - Questionable data

Sources:

Scott Fortman-Roe : Understanding the Bias-Variance Tradeoff

Wikipedia : Bias-variance tradeoff