607 Meetup #4
Feb 25, 2016

Tonight:

Brief Discussion: Generalized Solutions

-- training vs. test datasets

-- underfitting and overfitting

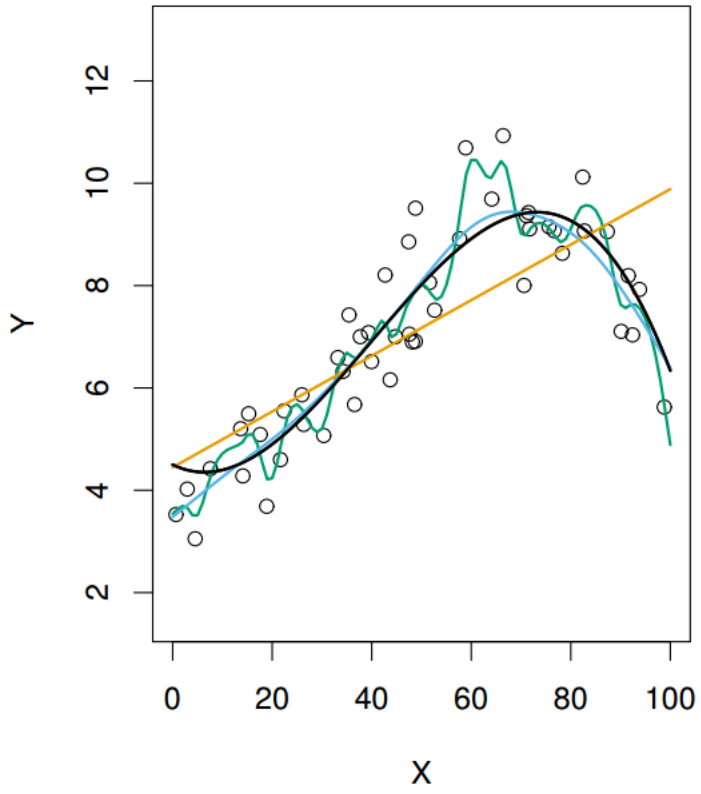
-- bias / variance tradeoff

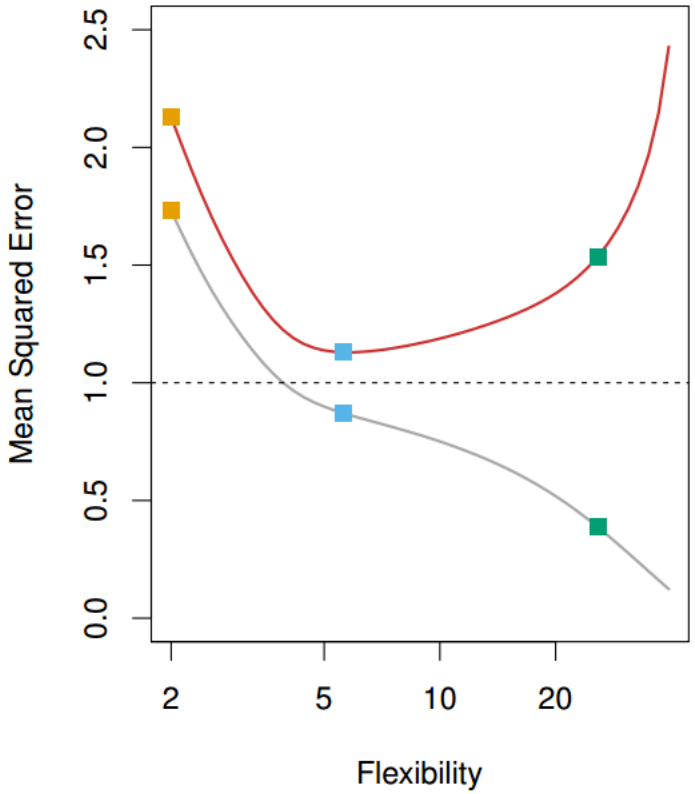
Reviewing your Week 4 assignments

Consider:

Suppose you took the code that you wrote for the six provided Simpsons character names in assignment four, problem 3 <training data set>, and ran the same code against an unseen set of 3 (or 3 million) names <test data set>.

- How accurate would your model be?
- How should you best measure your model's accuracy?





http://www-bcf.usc.edu/~gareth/ISL/Chapter2/2.9.pdf

As the flexibility of f-hat increases, its variance increases and its bias decreases.

Supplementary [Source: ISLR, pp. 34-35.]

- **Variance** refers to the amount by which f-hat (our estimate for y) would change if we estimated it using a different training data set.
- **Bias** refers to the error that is introduced by approximating a real life problem, which may be extremely complicated, by a much simpler model.

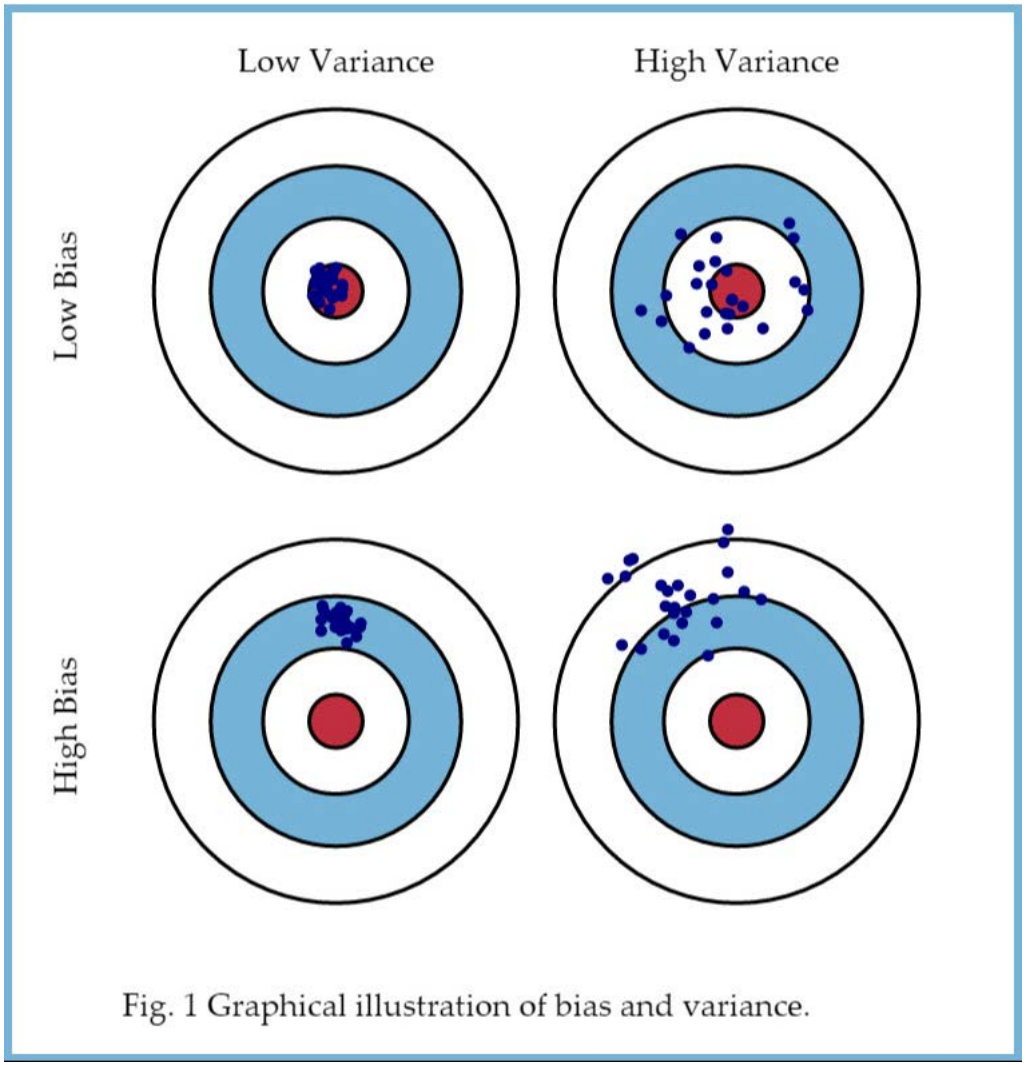
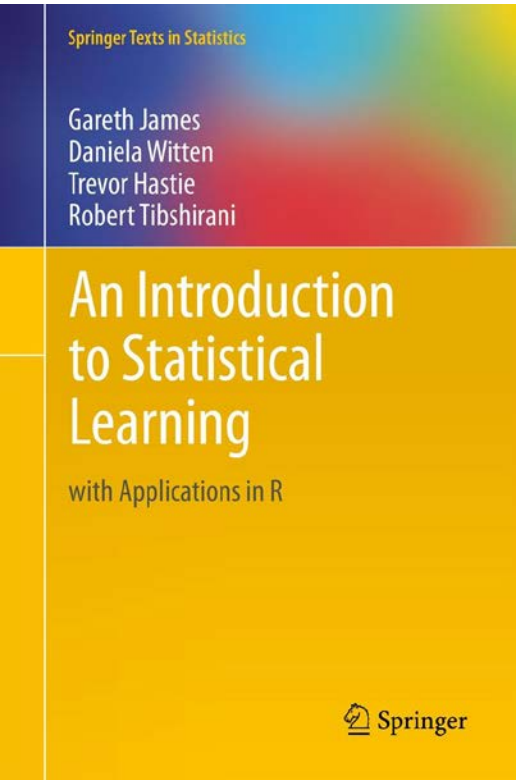


Fig. 1 Graphical illustration of bias and variance.

http:/source: Bias variance tradeoff. http://scott.fortmann-roe.com/docs/BiasVariance.html/

Best Summer Reading Project Ever?



Freely downloadable copy of book:
http://www-bcf.usc.edu/~gareth/ISL/

15 hours of video by two authors:
http://www.dataschool.io/15-hours-of-expert-machine-learning-videos/