Lecture #10: Bias and Variance Decomposition

Bias and Variance Analysis: Why?

- In this lecture, we will look into some theoretical analysis of learning
- Such analysis will help us build stronger intuition and develop rules of thumb about how to best apply learning algorithms in different settings

Intuition

- We want the learned classifier to have good generalization performance
- Problem:
  - Models with too few parameters can perform poorly (under-fitting)
  - Models with too many parameters can perform poorly (over-fitting)
- Solution:
  - Need to optimize the complexity of the model to achieve the best performance

Intuition (contd.)

- One way to get insight into this tradeoff is the decomposition of generalization error into squared bias and variance
  - A model which is too simple (inflexible) will have large bias
  - A model which is too complex (flexible) will have high variance

Intuition (contd.)

- Bias
  - Measures the accuracy or quality of the algorithm
  - High bias means a poor match
- Variance
  - Measures the precision or specificity of the match
  - High variance means a weak match
- We would like to minimize each of these
- Unfortunately, we can’t do this independently, there is a trade-off

Classical Statistical Analysis

- Suppose we are given a training sample S drawn from some population of possible training samples according to the distribution P(S) to learn a classifier h
- Compute $E_P\ [y\ != h(x)]$
- Decompose this into
  - Bias
  - Variance
  - Noise

Bias, Variance, and Noise

- Variance
  - describes how much h(x) varies from one training set to another
- Bias
  - describes the average error of h(x)
- Noise
  - describes the average noise in the labels

Squared Bias

- Low bias
  - Linear regression applied to linear data
  - 2nd degree polynomial applied to quadratic data
  - Neural network with many hidden units trained to completion
- High bias
  - Constant function
  - Linear regression applied to non-linear data
  - Neural network with few hidden units applied to non-linear data

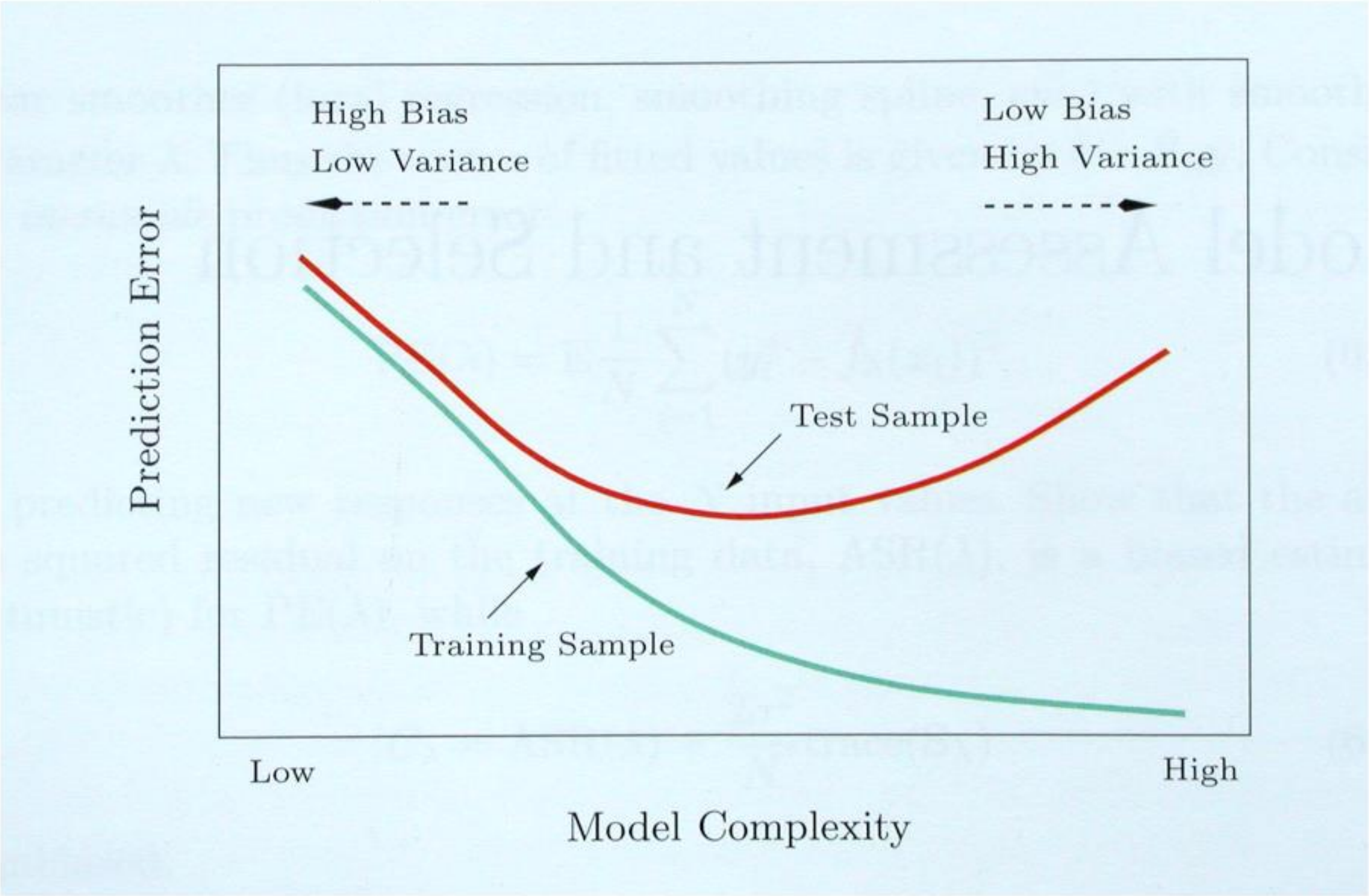
Variance

- Low variance
  - Constant function
  - Model independent of training data
- High variance
  - High degree polynomial
  - Neural network with many hidden units trained to completion

Bias / Variance Tradeoff

- (Squared bias + variance) is what counts for prediction
- Often
  - Low bias => high variance
  - Low variance => high bias
- Tradeoff
  - Squared bias vs. Variance

Bias / Variance Tradeoff



Hastie, Tibshirani, Friedman "Elements of Statistical Learning" 2001

Reduce Variance without Increasing Bias

- Averaging reduces variance

- Average models to reduce model variance
  - Bagging does exactly this!

When will Bagging improve Accuracy

- Depends on the stability of the base classifier
- A learner is unstable if a small change to the training set D causes a large change in the output hypothesis h
- Bagging helps unstable learners, but could hurt the performance of stable procedures
- Neural networks and decision trees are unstable
- K-NN and Naïve Bayes are stable

Reduce squared bias and decrease variance?

- Bagging reduces variance by averaging
- Bagging has little effect on bias
- Can we average and reduce bias?
  - Yes, Boosting!

Summary

- Expected error can be decomposed into three components: squared bias, variance, and noise
- Complex models lead to lower bias but higher variance
- Simple models lead to high bias but low variance
- Model selection is needed to trade-off the bias and variance