Chapter 2: What is Statistical Learning?

Esben Eickhardt

2023-05-10

Introduction

In this chapter we go through overarching statistical concepts that are relevant for all the models in the coming chapters.

Notation

Here we go over how we will be referring to different concepts in the coming chapters e.g. input variables are referred to as *predictors*, *independent variables*, *features* and *variables*, whereas outvariable are referred to as *response* or the *dependent variable*.

Y: Outputs
X: Inputs
n: The number of distinct data points
p: The number of variables
x_{ij}: Sample by variable in a matrix

We assume there is a relationship between X and Y, and that it can be written in the general form:

$Y = f(X) + \epsilon$

f is some unknown function and ϵ is the error term.

Reducible and Irreducible Errors

The accuracy of a prediction depends on two quantities, we are referred to as the *reducible error* and the *irreducible error*. The *reducible error* is called so as it can be reduced by choosing a better model, while the *irreducible* error cannot be reduced as it cannot be predicted by X.

Inference

Inference refers to the process of drawing conclusions about a population based on the analysis of a sample from that population.

Inference includes:

- Identifying which predictors are associated with the response
- Identifying the relationship between the response and each predictor
- Identifying if the model can adequately capture the relationship between X and Y

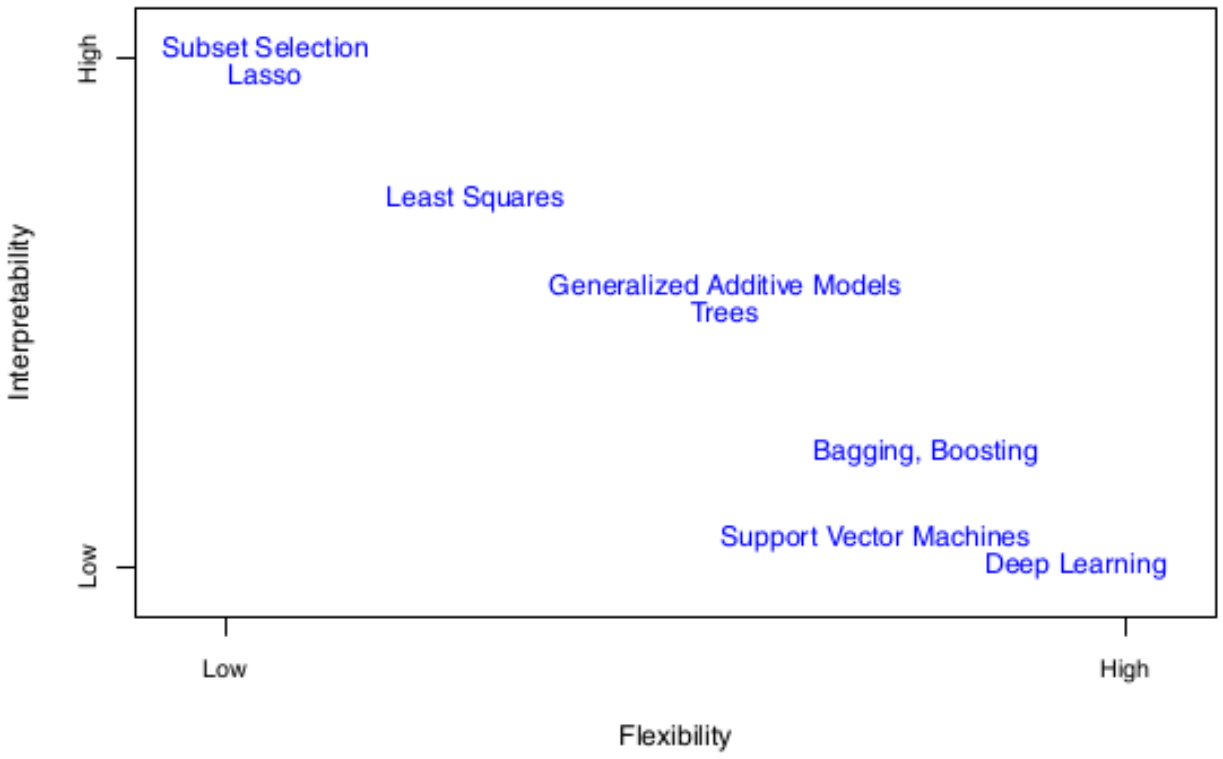
Often *inference* and *prediction accuracy* conflict as simple models are more interpretable, while complex model fit the data better and thus allow for more precise predictions.

Parametric vs Non-Parametric Methods

- Parametric
  - We reduce the problem of estimating f down to estimating a set of parameters
  - Examples: Linear model
  - Steps:
    1. We chose a model
    2. We use training data to *fit* or *train* the model e.g. using *least squares*
  - Disadvantage: It can be hard to chose the right model and one often underfits the data
- Non-Parametric
  - We do not make explicit assumptions about the functional form of f
  - Example: K-means, Splines
  - Disadvantages: Often requires a large number of observations and often overfits

Trade-Off between Prediction Accuracy and Model Interpretability

In the figure is shown the different models that are covered in the book and how flexible and how interpretable they are. We can see that the more flexible a model is the harder it is to interpret.



We have established that when inference is the goal, there are clear advantages to using simple and relatively inflexible statistical learning methods. In some settings, however, we are only interested in prediction, and the interpretability of the predictive model is simply not of interest.

Supervised vs Unsupervised Learning

- Supervised Machine Learning:

  - In supervised learning, the training data consists of input features and corresponding labeled outputs.The goal is to learn a mapping function from input to output by using the labeled data.
  - Examples of supervised learning methods include:
    * Linear regression
    * Logistic regression
    * Decision trees
    * Support vector machines
    * Neural networks

- Unsupervised Machine Learning:
  - In unsupervised learning, the training data does not have labeled outputs, and the algorithm explores the patterns and structures within the data on its own. The goal is to discover inherent structures, relationships, or clusters in the data.
  - Examples of unsupervised learning methods include:
    * K-means clustering
    * Hierarchical clustering
    * Dimensionality reduction techniques like Principal Component Analysis (PCA) and t-SNE

Regression vs Classification

Variables can be characterized as either quantitative (numerical values) or qualitative (categorical values). We refer to problems with a quantitative response as regression problems, while those involving a qualitative response are referred to as classification problems. The characterization is, however, not clear cut as e.g. logistic regression is often used for classification, as the probabilities the model outputs are translated to categories.

Assessing Model Accuracy (Regression)

A models accuracy should always be assessed on a previously unseen test data set, as it may overfit the data it was trained on! To the left in the figure is show some simulated data (black) and three fitted models: Linear (orange), spline 1 (blue), spline 2 (green). To the right the MSE is show of the training data (grey) and on the test data (red). While the most flexible model (spline 2) fits the training data the best, we can see it fits the test data poorly. When a given method yields a small training MSE but a large test MSE, we are said to be *overfitting* the data.

There is no measure of accuracy that fits all problems, but the most commonly used measure for regression is MSE.

- Measuring the Quality of Fit
  - In regression setting the most commonly used measure is *mean squared error*, which is just the average squared distance between the real values and the predicted values.

The Bias-Variance Trade-Off

The expected average *test MSE* can be calculated as the variance of the predictions plus the squared bias of the predictions plus the variance of the ϵ.

- Variance
  - The amount by which the model would change if we estimated it using a different training data set

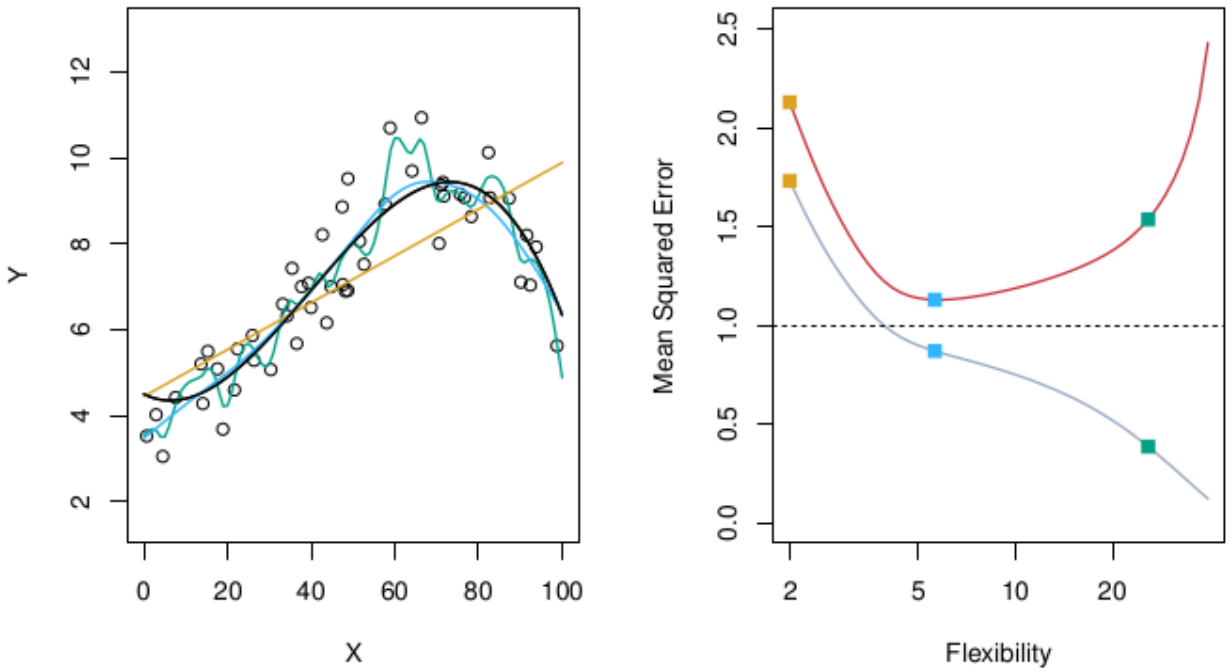


Figure 1: Training MSE vs Test MSE

  - If a method has a high variance then small changes in the training data can result in large changes in the model's parameters.

- Bias
  - The error that is introduced by a choosing a model that cannot capture the relationships in the data e.g. using a linear model to capture non-linear relationships.
  - A model has a high bias if it cannot produce accurate estimates of the data.

As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease. The concept can be seen in the figure from left to right for three datasets: 1. Simple non-linear relationship, 2. linear relationship, 3. Complicated non-linear relationship. The MSEs in the plots are for the test data.

The reason it is referred to as a trade-off is because it is easy to obtain a method with extremely low bias but high variance (for instance, by drawing a curve that passes through every single training observation).

Assessing Model Accuracy (Classification)

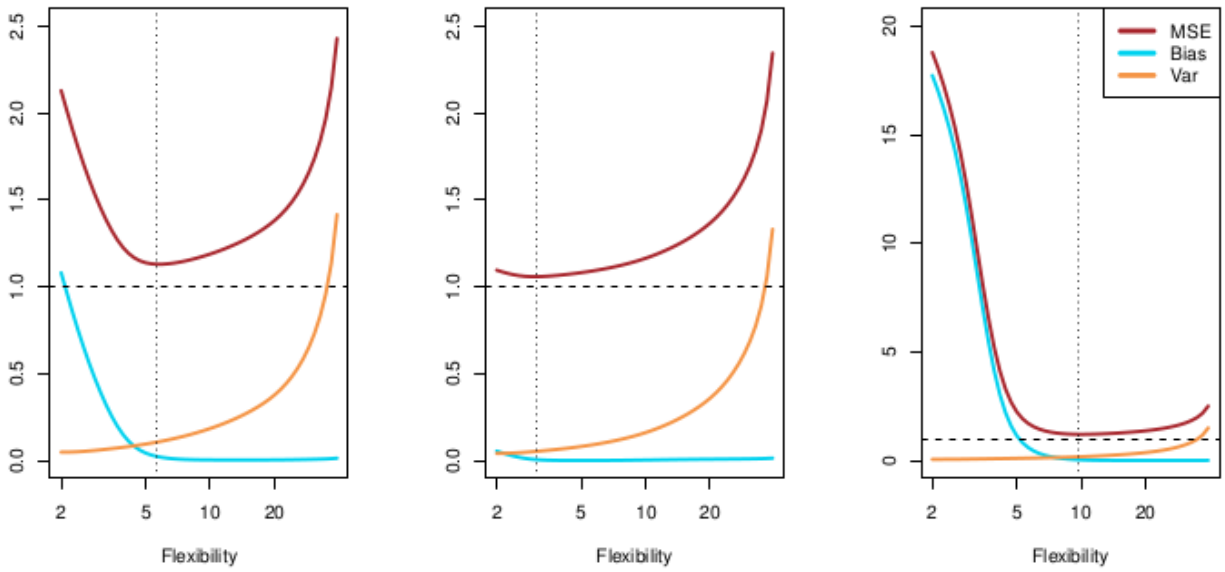


Figure 2: Variance vs Bias for three models