An Introduction to Statistical Learning

2 Chapter 2 Statistical Learning

2.1 What is Statistical Learning

2.1.1

x = input variables/predictors/independent variables/features/variables
y = output variables/response/depend variables

$$Y = f(X) + \epsilon \tag{1}$$

f is some fixed but unknown function, while ϵ is a random error term, independent from x and with mean 0.
Statistical learning refer to a set of approaches for estimating f.

We estimate f in order to predict and infer (how Y affects by X_i, what are the associations between X_i).

$$\hat{Y} = \hat{f}(X) \tag{2}$$

$\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y.

$$E(Y - \hat{Y})^2 = E(f(X) + \epsilon - \hat{f}(x))^2 = \underbrace{[f(x) - \hat{f}(x)]^2}_{Reducible} + \underbrace{Var(\epsilon)}_{Irreducible} \tag{3}$$

$E(Y - \hat{Y})^2$ represents the average, or expected value, of the squared difference between the predicted and actual value of Y , and $Var(\epsilon)$ represents the variance associated with the error term ϵ.

The accuracy of $\hat{Y}$ as a prediction for $\hat{Y}$ depends on reducible error and the irreducible error.
Y is also a function of ϵ, which, by definition, cannot be predicted using X.

2.1.2

Methods:
In general, fitting a more flexible model requires estimating a greater number of parameters. These more complex models can lead to a phenomenon known as overfitting the data, which essentially means they overfitting follow the errors, or noise, too closely.
Parametric methods:
First, we make an assumption about the functional form, or shape,of f.
After a model has been selected, we need a procedure that uses the training data to fit or train the model. **(most common used one is least squares)**

Disadvantage
The model we choose will usually not match the true unknown form of f. **Non-parametric**
Non-parametric methods do not make explicit assumptions about the functional form of f.
Advantage
By avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f.
Disadvantage
Since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

2.1.3
The Trade-Off Between Prediction Accuracy and Model Interpretability.

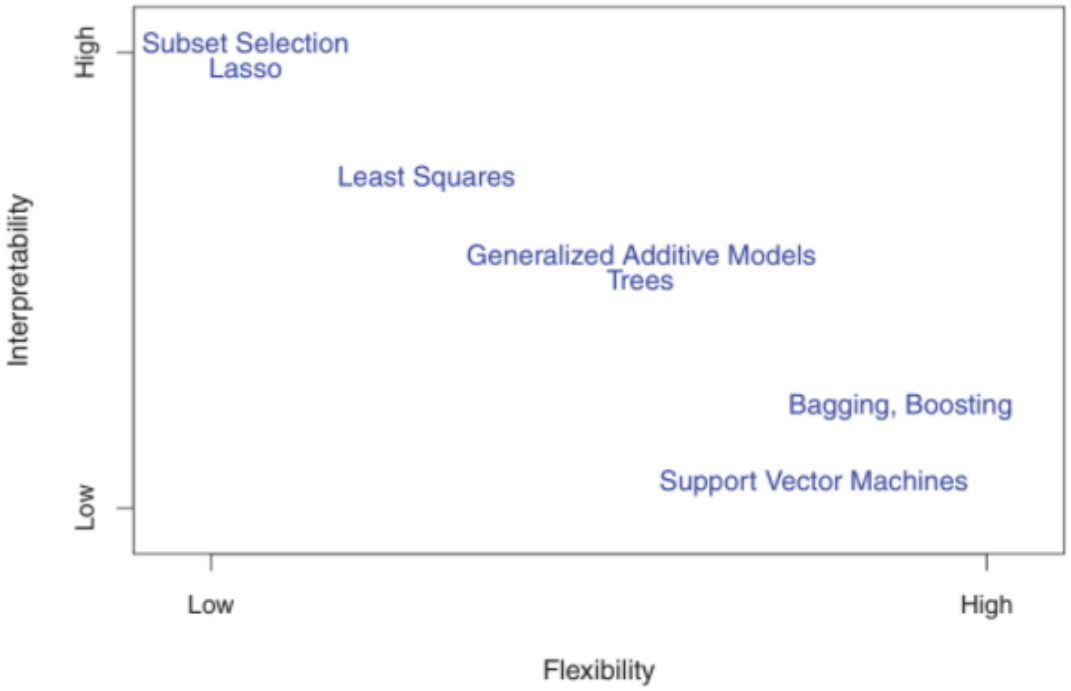

Figure 1: A representation of the trade-off between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.

2.1.4
Most statistical learning problems fall into one of two categories: **supervised supervised or unsupervised**. Unsupervised learning is because we lack a response variable that can supervise our analysis.

Many problems fall naturally into the supervised or unsupervised learning paradigms. However, sometimes the question of whether an analysis should be considered supervised or unsupervised

is less clear-cut. For instance, suppose that we have a set of n observations. For m of the observations, where $m < n$, we have both predictor measurements and a response measurement. For the remaining n m observations, we have predictor measurements but no response measurement. Such a scenario can arise if the predictors can be measured relatively cheaply but the corresponding responses are much more expensive to collect. We refer to this setting as a semi-supervised learning problem.

2.1.5
We tend to refer to problems with a quantitative response as **regression problems**, while those involving a qualitative response are often referred to as **classification problems**.(but not always) We tend to select statistical learning methods on the basis of whether the response is quantitative or qualitative.

2.2 Assessing Model Accuracy

There is no free lunch in statistics: no one method dominates all others over all possible data sets.

2.2.1 2.2.2
In **regression setting** the most common used method is mean squared error (MSE).

$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}(x_i))^2 \tag{4}$$

where $f(x_i)$ is the prediction that $\hat{f}$ gives for the ith observation.

We are really not interested in whether $f(x_i)y_i$; instead, we want to know whether $f(x_0)$ is approximately equal to y_0, where (x_0, y_0) is a previously unseen test observation not used to train the statistical learning method. We want to choose the method that gives the lowest test MSE, as opposed to the lowest training MSE.

There is no guarantee that one method which has the smallest training MSE will also generate smallest test MSE.

As model flexibility increases, training MSE will decrease, but the test MSE may not.

When a given method yields a small training MSE but a large test MSE, we are said to be **overfitting** the data.

$$E(y_0 - \hat{y}(x_0))^2 = Var(f(x_0)) + [Bias(f(x_0))]^2 + Var(\epsilon) \tag{5}$$

Here the notation $E(y_0 - f(x_0)^2$ defines the expected test MSE, and refers expected to the average test MSE that we would obtain if we repeatedly estimated test MSE f using a large number of

training sets, and tested each at x_0. The overall expected test MSE can be computed by averaging $E(y_0 - f(x_0)^2$ over all possible values of x0 in the test set.

Variance refers to the amount by which f would change if we estimated it using a different training data set.
Different training data sets will result in a different f.
In general, more flexible statistical methods have higher variance.

Flexible method has high variance because changing any one of these data points may cause the estimate f to change considerably. In contrast, inflexible method has low variance, because moving any single observation will likely cause only a small shift in the estimation.

Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.

As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease. The relative rate of change of these two quantities determines whether the test MSE increases or decreases. However, at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance.
The challenge lies in finding a method for which both the variance and the squared bias are low. This trade-off is one of the most important recurring themes in this book.

2.2.3
In **classification setting** most common approach for quantifying the accuracy is the error rate.

$$\frac{1}{n} = \sum_{n}^{1} I(y_i \neq \hat{y}_i) \tag{6}$$

Here y_i is the predicted class label for the ith observation using f. And $I(y_i = y_i)$ is an indicator variable that equals 1 if $y_i = y_i$ and zero if $y_i = y_i$. If $I(y_i = y_i) = 0$ then the ith observation was classified correctly by our classification method; otherwise it was misclassified.

The Bayes Classifier
Bayes classifier serves as an unattainable gold standard against which to compare other methods.

We should simply assign a test observation with predictor vector x_0 to the class j for which

$$Pr(Y = j|X = x0) \tag{7}$$

is largest. Note that the above notation is a conditional probability: it is the probability conditional that $Y = j$, given the observed predictor vector x_0. This very simple classifier is called the Bayes classifier.

The Bayes Bayes classifier's prediction is determined by the **Bayes decision boundary**.
The Bayes classifier produces the lowest possible test error rate, called the **Bayes error rate**. The Bayes error rate is analogous to the irreducible error.

In general, the overall Bayes error rate is given by

$$1 - E(max_j Pr(Y = j|X)) \tag{8}$$

where the expectation averages the probability over all possible values of X.

K-nearest neighbors
Given a positive integer K and a test observation x_0, the KNN classifier first identifies the neighbors K points in the training data that are closest to x_0, represented by N_0. It then estimates the conditional probability for class j as the fraction of points in N0 whose response values equal j:

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j). \tag{9}$$

KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.
When K = 1, the decision boundary is overly flexible. This corresponds to a classifier that has low bias but very high variance. As K grows, the method becomes less flexible and produces a decision boundary that is close to linear. This corresponds to a low-variance but high-bias classifier.