Ch. 2. Statistical Learning

Suppose that we observe a quantitative response Y and p different predictors X_1, …, X_p. We assume that there is some relationship between Y and X = (X_1, …, X_p), which can be written in the very general form Y = f(X) + ϵ.
Here f is some *fixed but unknown* function of X_1, …, X_p, and ϵ is a random *error term*, which is independent of X and has mean 0. In this formula, f represents the *systematic* information that X provides about Y.
In essence, statistical learning refers to a set of approaches for estimating f.

Why Estimate f?

There are two main reasons that we may wish to estimate f: *prediction* and *inference*.

Prediction
In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In this setting, since the error term averages to zero, we can predict Y using $\widehat{Y} = \widehat{f}(X)$ where $\widehat{f}$ represents our estimate for f, and $\widehat{Y}$ represents the resulting prediction for Y. In this setting, $\widehat{f}$ is often treated as a *black box*, in the sense that one is not typically concerned with the exact form of $\widehat{f}$, provided that it yields accurate predictions for Y.

The accuracy of $\widehat{Y}$ as a prediction for Y depends on two quantities, which we will call the *reducible error* and the *irreducible error*. In general, $\widehat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error.

This error is *reducible* because we can potentially improve the accuracy of $\widehat{f}$ by using the most appropriate statistical learning technique to estimate f. However, even if it were possible to form a perfect estimate for f, so that our estimated response took the form $\widehat{Y} = f(X)$, our prediction would still have some error in it. This is because Y is also a function of ϵ, which, by definition, cannot be predicted using X. Therefore, variability associated with ϵ also affects the accuracy of our predictors. This is known as the *irreducible* error, because no matter how well we estimate f, we cannot reduce the error introduced by ϵ.

Why is the irreducible error larger than zero? The quantity ϵ may contain unmeasured variables that are useful in predicting Y: since we don't measure them, f cannot use them for its prediction.
The quantity ϵ may also contain unmeasurable variation.

Consider a given estimate $\widehat{f}$ and a set of predictors X, which yields the prediction $\widehat{Y} = \widehat{f}(X)$. Assume for a moment that both $\widehat{f}$ and X are fixed. Then it is easy to show that:

$$E(Y - \hat{Y})^2 = E[f(X) + \varepsilon - \hat{f}(X)]^2$$
$$= [f(X) - \hat{f}(X)]^2 + Var(\varepsilon)$$
Reducible **Irreducible**

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

It is important to keep in mind that the irreducible error will always provide an upper bound on the accuracy of our prediction for Y. This bound is almost always unknown in practice.

Inference

We are often interested in understanding the way that Y is affected as X_1, …, X_p change. In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y. We instead want to understand the relationship between X and Y, or more specifically, to understand how Y changes as a function of $X_1,...,X_p$. Now $\hat{f}$ cannot be treated as a black box, because we need to know its exact form. In this setting one may be interested in answering the following questions:

- *Which predictors are associated with the response?* It is often the case that only a small fraction of the available predictors are substantially associated with Y. Identifying the few *important* predictors among a large set of possible variables can be extremely useful, depending on the application.
- *What is the relationship between the response and each predictor?* Some predictors may have a positive relationship with Y, in the sense that increasing the predictor is associated with increasing values of Y. Other predictors may have the opposite relationship. Depending on the complexity of f, the relationship between the response and a given predictor may also depend on the values of the other predictors.
- *Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?* Historically, most methods for estimating f have taken a linear form. In some situations, such an assumption is reasonable or even desirable. But often the true relationship is more complicated, in which case a linear model may not provide an accurate representation of the relationship between the input and output variables.

Depending on whether our ultimate goal is prediction, inference, or a combination of the two, different methods for estimating f may be appropriate.

For example, *linear models* allow for relatively simple and interpretable inference, but may not yield as accurate predictions as some other approaches.
In contrast, some of the highly non-linear approaches that we will discuss in the later chapters of this book can potentially provide quite accurate predictions for Y, but this comes at the expense of a less interpretable model for which inference is more challenging.

How Do We Estimate f?

Generally statistical learning methods share certain characteristics. We will always assume that we have observed a set of n different data points. These obs are called the *training data* because we will use these obs to train, or teach, our method how to estimate f.

Let x_{ij} represent the value of the jth predictor, or input, for obs i, where i = 1,...,n and j = 1,...,p. Correspondingly, let y_i represent the response variable for the ith obs. Then our training data consists of $\{(x_1,y_1),...,(x_n,y_n)\}$ where $x_i = (x_{i1},...,x_{ip})^T$.

Our goal is to apply a statistical learning method to the training data in order to estimate the unknown function f. We want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X,Y). Broadly speaking, most statistical learning methods for this task can be characterized as either *parametric* or *non-parametric*.

Parametric Methods

Parametric methods involve a two-step model-based approach

1. We make an assumption about the functional form, or shape, of f. E.g. f is linear. Once we have *assumed* that f is linear, the problem of estimating f is greatly simplified. Instead of having to estimate an entirely arbitrary p-dimensional function f(X), one only needs to estimate the p + 1 coefficients.
2. After a model has been selected, we need a procedure that uses the training data to *fit* or *train* the model. In the case of the linear model, the p + 1 coefficients

The *parametric* approach reduces the problem of estimating f down to one of estimating a set of parameters that define the f of the assumed family of functions.

A potential disadvantage of the parametric approach is that the model we choose will usually not match the true unknown form of f.
If the chosen model is too far from the true f, then our estimate will be poor.
We can try to address this problem by choosing flexible models that can fit many functional forms for f. But 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 follow the errors, or *noise*, too closely.

Non-parametric Methods

Non-parametric methods do not make explicit assumptions about the functional form of f. Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.
The major advantage of these methods over parametric approaches is that, 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. Essentially, no assumption about the form of f is made.
Non-parametric methods suffer from a major disadvantage: since they do not reduce the problem of estimating f to a small number of parameters, a very large number of obs (far

more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

The Trade-Off Between Prediction Accuracy and Model Interpretability

One might reasonably ask the following question: *why would we ever choose to use a more restrictive method instead of a very flexible approach?* There are several reasons that we might prefer a more restrictive model.

If we are mainly interested in inference, then restrictive models are much more interpretable. In some settings, however, we are only interested in prediction, and the interpretability of the predictive model is simply not of interest.
Surprisingly, we will often obtain more accurate predictions using a less flexible method, and this has to do with the potential for overfitting in highly flexible methods.

Supervised Versus Unsupervised Learning

Most statistical learning problems fall into one of two categories: *supervised* or *unsupervised*.

Supervised learning: for each observation of the predictor measurement(s), x_i, there is an associated response measurement y_i. We wish to fit a model that relates the response to the predictors with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference).

Unsupervised learning: a somewhat more challenging situation in which for every obs i = 1, …, n, we observe a vector of measurements x_i but no associated response y_i. It is not possible to fit a linear regression model since there is no response variable to predict. In this setting, we are in some sense working blind; the situation is referred to as *unsupervised* because we lack a response variable (label) that can supervise our analysis.

What analysis are possible in an *unsupervised* setting?
We can seek to understand the relationships between the variables or between the observations. One statistical learning tool that we may use in this setting is *cluster analysis*, or clustering. The goal of cluster analysis is to ascertain, on the basis of $x_1,...,x_n$, whether the observations fall into relatively distinct groups. Identifying such groups can be of interest because it might be that the groups differ with respect to some property of interest.

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 obs. For m of the obs, where m < n, we have both predictor measurements and a response measurement. For the remaining n - m we don’t have response measurements. Such 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. In this setting, we wish to use a statistical learning method that can incorporate the m obs for which the response measurements are available as well as the n - m obs for which they are not.

Regression Versus Classification Problems

Variables can be characterized as either *quantitative* or *qualitative* (aka *categorical*). Quantitative variables take on numerical values while *qualitative* variables take on values in one of K different *classes*, or categories.

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.

We tend to select statistical learning methods on the basis whether the response is quantitative or qualitative. However, whether the *predictors* are qualitative or quantitative is considered less important. Most of the statistical learning methods discussed in this book can be applied regardless of the predictor variable type, provided that any qualitative predictors are properly *coded* before the analysis is performed.

Assessing Model Accuracy

Why is it necessary to introduce so many different statistical learning approaches, rather than just a single *best* method? No one method dominates all others over all possible data sets. On a particular data set, one specific method may work best, but some other method may work better on a similar but different data set. Hence it is an important task to decide for any given set of data which method produces the best result. Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice.

Measuring the Quality of Fit

In order to evaluate the performance of a statistical learning method on a given data set, we need some way to measure how well its predictions actually match the observed data. That is, we need to quantify the extent to which the predicted response value for a given obs is close to the true response value for that obs.

In the regression setting the most commonly-used measure is the *mean squared error* (MSE) given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2$$

The MSE will be small if the predicted responses are very close to the true responses, and will be large if for some of the obs, the predicted and true responses differ substantially. This MSE is computed using the training data that was used to fit the model, ando so should more accurately be referred to as the *training MSE*. But in general, we do not really care how

well the method works on the training data. Rather, we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data.
We want to choose the method that gives the lowest *test MSE*.
One way to estimating test MSE is *cross-validation*.

The Bias-Variance Trade-Off

The test MSE is the result of two competing properties of statistical learning methods. It is possible to show that the expected test MSE, for a given value x_0, can always be decomposed into the sum of three fundamental quantities: the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$ and the variance of the error terms ϵ. That is:

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

Here the notation $E(y_0 - \hat{f}(x_0))^2$ defines the *expected test MSE*, and refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at x_0. The overall expected test MSE can be computed by averaging $E(y_0 - \hat{f}(x_0))^2$ over all possible values of x_0 in the test set.

This equation tells us that in order to minimize the expected test error, we need to select a statistical learning method that simultaneously achieves *low variance* and *low bias*.

Note that variance is inherently a nonnegative quantity, and squared bias is also nonnegative. Hence, we see that the expected test MSE can never lie below Var(ϵ), the irreducible error.

What do we mean by the *variance* and *bias* of a statistical learning method?

Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. Since the training data are used to fit the statistical learning method, different training data sets will result in a different $\hat{f}$. But ideally the estimate for f should not vary too much between training sets. However, if a method has high variance then small changes in the training data can result in large changes in $\hat{f}$.
A flexible method that follows the observations very closely has more variance than a more "rigid" one such as linear regression.

On the other hand, *bias* refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. For example, linear regression assumes that there is a linear relationship between Y and $X_1,...,X_p$. It is unlikely that any real-life problem truly has such a simple linear relationship, and so performing linear regression will undoubtedly result in some bias in the estimate of f.

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. As we increase the flexibility of a class of methods, the bias tends to initially decrease faster than the variance increases. Consequently, the expected test MSE declines. However, at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance. When this happens, the MSE increases.

The flexibility level corresponding to the optimal test MSE can differ considerably based on data sets, because squared bias and variance change depending on the data set.

Good test set performance of a statistical learning method requires low variance as well as low bias. This is referred as the *bias-variance trade-off* because it's easy to obtain a method with extremely low bias but with high variance or a method with very low variance but with high bias. The challenge lies in finding a method for which both the variance and the squared bias are low.

In a real-life situation in which f is unobserved, it is generally not possible to explicitly compute the test MSE, bias, or variance for a statistical learning method. Nevertheless, one should always keep the bias-variance trade-off in mind.