An Introduction to Statistical Learning

(Summary by *Pablo Vivas*)

Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani

January 13, 2020

1 Introduction

Statistical learning refers to a vast set of tools for **understanding data**. These tools can be classified as:

- supervised
- unsupervised

Broadly speaking, supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs. Problems of this nature occur in fields as diverse as business, medicine, astrophysics, and public policy. With unsupervised statistical learning, there are inputs but no supervising output; nevertheless we can learn relationships and structure from such data. The datasets used in this books are:

- Wage
- Stock Market
- Gene Expression

Some history

Legendre and Gauss published papers on the method of *least squares*.
Fisher proposed *linear discriminant analysis* in 1936.
In the 1940s, various authors put forth an alternative approach, *logistic regression*.
In the early 1970s, Nelder and Wedderburn coined the term *generalized linear models* for an entire class of statistical learning methods that include both linear and logistic regression as special cases.
In mid 1980s Breiman, Friedman, Olshen and Stone introduced *classification and regression trees*.
Hastie and Tibshirani coined the term *generalized additive models* in 1986 for a class of non-linear extensions to generalized linear models.

2 Statistical Learning

Suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, ..., X_p$. **We assume** that there is some relationship between Y and $X = (X_1, X_2, ..., X_p)$, which can be written in the very general form:

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

Here f is some fixed but unknown function of $X_1, X_2, ..., X_p$, and ϵ is a random error term, which is independent of X and has mean zero. In this formulation, f represents the systematic information that X provides about Y.

Why Estimate f ?

- *Prediction*: In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. Since the error term averages to zero, we can predict Y using

  $$\boxed{\hat{Y} = \hat{f}(X)}$$

  Where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y . **In this setting,$\hat{f}$ is often treated as a black box**, in the sense that one is not typically concerned with the exact form of $\hat{f}$, provided that it yields accurate predictions for Y.

  The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities, which we call the *reducible error* and the *irreducible error*.$\hat{f}$ 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 $\hat{f}$ 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 $\hat{Y} = f(x)$, 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 predictions. This is known as the irreducible error, because no matter how well we estimate f, we cannot reduce the error introduced by ϵ.

- *Inference*: We are often interested in understanding the way that Y is affected as $X_1, X_2, ..., X_p$ 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 $X_1, X_2, ..., X_p$. **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?
  - What is the relationship between the response and each predictor?
  - Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

How Do We Estimate f?

- *Parametric Models*: Parametric methods involve a two-step model-based approach.

    1. First, we make an assumption about the functional form, or shape, of f. For example, one very simple assumption is that f is linear in X:

    $$\boxed{f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, ..., \beta_p X_p}$$

    This is a *linear model*. 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 $\beta_0, \beta_1, .., \beta_p$.

    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, we need to estimate the parameters $\beta_0, \beta_1, .., \beta_p$. That is, we want to find values of these parameters such that

    $$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2, ..., \beta_p X_p$$

    This approach reduces the problem of estimating f down to one of estimating a set of parameters. Assuming a parametric form for f simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters than it is to fit an entirely arbitrary function f. The potential disadvantage of a 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 different possible 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 Models*: 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. Such approaches can have a major advantage over parametric approaches: 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. Any parametric approach brings with it the possibility that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well. In contrast, non-parametric approaches completely avoid this danger, since essentially no assumption about the form of f is made. But non-parametric approaches do 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 observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

Prediction Accuracy and Model Interpretability: The Trade-Off

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. For instance, when inference is the goal, the linear model may be a good choice since it will be quite easy to understand the relationship between Y and $X_1, X_2, ..., X_p$. In contrast, very flexible approaches can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response.

In some settings, however, we are only interested in **prediction**, and the interpretability of the predictive model is simply not of interest. For instance, if we seek to develop an algorithm to predict the price of a stock, our sole requirement for the algorithm is that it predict accurately— interpretability is not a concern. In this setting, we might expect that it will be best to use the most flexible model available. Surprisingly, this is not always the case! We will often obtain more accurate predictions using a less flexible method. This phenomenon, which may seem counterintuitive at first glance, has to do with the potential for overfitting in highly flexible methods.

Figure 1 provides an illustration of the trade-off between flexibility and interpretability for some methods.

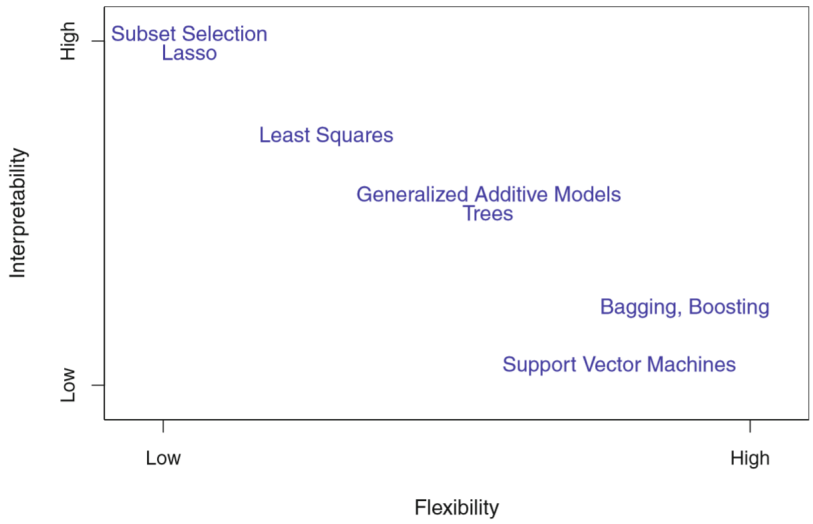

Figure 1: A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods

Supervised vs Unsupervised Learning

Regression vs Classification

Quality of fit

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

Bias vs Variance

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

3 Linear Regression

4 Classification

5 Resampling Methods

6 Linear Model Selection and Regularization

7 Moving Beyond Linearity

8 Tree-Based Methods

9 Support Vector Machines

10 Unsupervised Learning