Jim Lambers
MAT 610
Spring Semester 2018-19
Lecture 22 Notes

These notes correspond to Chapter 2 of *An Introduction to Statistical Learning* by James, et al.

What is Statistical Learning?

Suppose that an *output variable* Y depends on an *input variable* X in the following manner:

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

where the function f is unknown and ϵ is an *error term*, which is a random variable assumed to have zero mean; that is, $E[\epsilon] = 0$. Statistical learning is the task of estimating f from given data, so that we can better understand the dependence of Y on X. That is, f conveys the *systematic information* that X provides about Y.

Often X is a vector-valued quantity of the form $X = (X_1, X_2, \ldots, X_p)$. These components are often referred to as *predictors* or *independent variables*, while Y is also known as a *dependent variable*, or *response*.

Example Suppose our data set consists of sales of a product in 200 different markets, and the advertising budget for each market in three different media: TV, radio and newspaper. How can we determine the relationship between each advertising budget and sales? The three budgets are predictors, and sales is the response. □

Why Estimate f?

Estimation of f is useful for both *prediction* and *inference*.

Prediction

If, in practice, the output Y is not easily obtained but the input X is, then it is desirable to be able to *predict* what the output Y will be for a given value of X. Such a prediction has the form

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

where $\hat{f}$ is an estimate of the unknown function f.

Example Suppose the predictors $X_1, X_2, \ldots, X_p$ are characteristics of a patient's blood sample that can easily be measured in a lab, and Y represents the patient's risk for a severe reaction to some drug. If we can assess a patient's risk before giving them the drug, then we can avoid doing so if their risk turns out to be high. □

The error in our prediction has two components: *reducible error* and *irreducible error*. The reducible error is a measure of the deviation of $\hat{f}$ from f, which can be reduced by choosing the right statistical learning technique or tuning it correctly. In some cases, we may be able to reduce this error to zero, and obtain $\hat{f} = f$. The irreducible error stems from the error term ϵ, which cannot be predicted using X. For example, it may depend on unmeasured variables not among the predictors in X.

For a fixed value of X, our overall error can be decomposed as follows:

$$\begin{aligned} E[(Y-\hat{Y})^2] &= E[(f(X)+\epsilon-\hat{f}(X))^2] \\ &= E[(f(X)-\hat{f}(X))]^2 + 2\epsilon(f(X)-\hat{f}(X)) + E[\epsilon^2] \\ &= (f(X)-\hat{f}(X))^2 + \mathrm{Var}(\epsilon) \end{aligned} \tag{3}$$

The first term is the reducible error, and the second term is the irreducible error. We will focus on minimizing the reducible error, since the irreducible error is beyond our control once we have chosen our predictors.

Inference

Prediction is about estimating the value of the unknown function f at values of X outside of the given data. For such a task, f can be treated as a "black box", meaning the the form of f does not need to be known. Inference, on the other hand, is about better understanding the relationship between the each predictor and the response, which means at least approximate knowledge of the form of f is of paramount importance.

Inference is often concerned with answering these questions:

- Which predictors most influence the response? In some cases, certain predictors may have little to no impact on the response, so that we can exclude them from consideration, resulting in a simpler mathematical model.
- Is the response an increasing or decreasing function of each predictor? Knowing this is a helpful first step toward determining an approximate form for f.
- Is the relationship between the response and each predictor linear or nonlinear? There are well-established techniques for modeling f in the case where the response depends linearly on the predictors, so it is worthwhile to ascertain whether such a model is feasible.

Example From the example of measuring sales based on advertising budgets, inference can involve questions such as:

- Which media contribute to sales? It may be discovered, for example, that radio advertising has little impact on sales.
- Which media generate the largest boost in sales? It is worth knowing if sales are more heavily influenced by TV or newspaper advertising, for example.
- What increase in sales can be attributed to an increase in a particular form of advertising, such as TV? Answering this question helps us understand the effect of modifying advertising budgets, and what is needed to reach sales targets.

For the case of a *linear model*, in which the response Y (sales) is an *affine* function of X_1, X_2, X_3 (advertising budgets), meaning that

$$Y \approx Y_0 + m_1X_1 + m_2X_2 + m_3X_3,$$

all of these questions can be rephrased mathematically in terms of the slopes m_1, m_2 and m_3:

- Which of m_1, m_2, m_3 are not negligibly small?

- Which is the largest of m_1, m_2 or m_3?
- What is the value of m_i for a given i, where i an index referring to television, radio or newspaper?

□

How Do We Estimate f?

There are many methods for estimating f, but they generally work with a set of observations, which are predictor-response pairs (x_i, y_i), $i = 1, 2, \dots, n$. These observations are called the *training data*, as they are used to "train', or "teach", our statistical learning method of choice how to estimate f.

Estimation methods generally fall into two categories: *parametric* and *non-parametric*. We now examine each of these categories.

Parametric Methods

In a parametric method, we first assume f has a particular form. For example, we may assume that f is a *linear*, or more precisely, *affine* function of X:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{4}$$

Thus what remains is to estimate the coefficients $\beta_0, \dots, \beta_p$.

Next, we use the training data to *fit*, or *train*, the model. In the case of a linear model, that means using some algorithm to obtain values for the coefficients $\beta_0, \dots, \beta_p$ such that

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

for each predictor-response pair (x_i, y_i) in the training set. For example, if we wish to minimize

$$\sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right]^2,$$

we can do so by solving an appropriate *least squares problem.* Other criteria for determining the best fit lead to other methods of obtaining the coefficients.

A parametric method has the advantage of greatly simplifying the task of obtaining an estimate $\hat{f}$ of f, because it reduces the task to one of computing certain coefficients. However, if the chosen form for f is not representative of f, then the estimate will be far less useful. This can be remedied by choosing a more flexible form for f, such as a nonlinear function instead of a linear model, but then, not only is the task of obtaining the specifics of that form more complicated, but the estimate $\hat{f}$ might *overfit* the data, meaning that it captures errors, or *noise*, in the data, which throws off the estimate.

Non-parametric Methods

In non-parametric methods, it is not assumed that the function f has a particular form, at least not over the entire domain of the training data. One example of a non-parametric method is to construct a *thin-plate spline*, which is linear combination of *radial basis functions.* Another, for one-dimensional data sets, is a *cubic spline*, which is a piecewise polynomial.

The term "non-parametric", as well as its informal definition of not assuming a particular functional form, is somewhat misleading because, mathematically, some form *is* assumed and values of coefficients, or parameters, need to be obtained. The practical difference is that for non-parametric methods, a much larger number of parameters is required. In fact, the number of parameters is often proportional to n, the number of observations, which is not the case for parametric methods such as least-squares fitting.

Prediction Accuracy vs. Model Interpretability

When choosing a method for estimating f, one must consider the competing criteria of *prediction accuracy* and *model interpretability.* For example, by choosing a linear model, one is sacrificing prediction accuracy, because the error in the resulting estimate $\hat{f}$ may be poor if the true function f does not behave like a linear function of the predictors, even in an approximate sense. However, such a model is very easy to interpret, because the relationship between the response and each predictor is described very simply in terms of a slope.

More generally, the more flexible the model, the more accurate the prediction will be, unless there is overfitting that includes noise in the estimate. However, a more flexible model will also be more difficult to interpret, because the relationship between the response and each predictor will be a more complicated function.

An example of an even more interpretable, and therefore less flexible, model than a linear model is a *lasso*, in which a linear form is still assumed, but only certain coefficients among β_i, $i = 0, 1, \ldots, p$ are allowed to be nonzero. While this results in a less accurate fit than a standard linear model, it facilitates describing the response in terms of only a few predictors.

By contrast, a more flexible but less interpretable method is the use of a *generalized additive model*, in which Y is not assumed to be a linear combination of a constant function β_0 and the X_i, but rather chosen *functions* of the X_i. For example, a function of the form $X_1X_2^2$ may be included. However, such a function makes it difficult to understand the effect on the response of a change in a particular predictor, unlike for a linear model.

Supervised vs. Unsupervised Learning

The various methods discussed so far for estimating f are all examples of *supervised learning*, in which predictor-response pairs are used to obtain the estimate $\hat{f}$. By contrast, in *unsupervised learning*, there are no response values available; we only have a set of predictor values x_i, $i = 1, 2, \ldots, n$. This greatly limits our ability to estimate f, but some analysis is still possible. For example, *cluster analysis* can be used to group the observations into categories that may correspond to different responses.

Regression vs. Classification

Statistical learning works with both *quantitative* variables, which have numerical values, and *qualitative*, or *categorical*, values, which do not. For example, temperature is a quantitative variable, while gender or day of the week is not. When a response is quantitative, it is often estimated by solving a *regression* problem, such as least-squares fitting. When it is qualitative, the problem of estimating the response in terms of its predictors is called a *classification* problem.

For purposes of classifying a problem as a regression or classification problem, it generally does not matter if the *predictors* are quantitative or qualitative; in the latter case, one can use *dummy variables* to represent qualitative predictors.

Example Suppose we wish to determine whether males or females have higher credit card balances, using a training data set consisting of balances for a large number of customers identified only by gender. Then, we can use a linear model of the form

$$Y = \beta_0 + \beta_1 X$$

where X denotes gender, and has the values

$$x_i = \begin{cases} 1 & \text{female,} \\ 0 & \text{male.} \end{cases}$$

Then, we can obtain values for β_0 and β_1 by solving a regression problem, as the response in this case, being a credit card balance, is quantitative. To interpret the model, we can substitute $X = 0$ or $X = 1$ to obtain β_0, the credit card balance for males, and $\beta_0 + \beta_1$, the balance for females. □