Statistic Learning-Introduction

1. What is statistic learning

1.1 Introduction

Let's get started with a simple case. Suppose that you're a statistic consultant and hired by a client to provide suggestion on how to improve sales of a particular product. You have the `advertising` data consist of sales of three types of products: `TV` , `radio` , and `newspaper` . Generally, the client want to know whether there is an association between advertising and sales. So your goal is to develop an accurate model that can predict sales on the basis of existing data.

To be more mathematically , above can be written in a formula :

$$\text{Sales} \approx f(\text{TV}, \text{Radio}, \text{Newspaper})$$

In this setting, the advertising budgets are `input variables` while sales input is an `output variable` . The inputs `TV` , `Radio` and `Newspaper` go by different names, such as `predictors` , `independent variables` , `features` or `variables` . The output variable is often called `reponse` or `dependent variable` .

More generally, suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$ .The relationship can be marked in a very general form.

$$Y = f(X) + \epsilon.$$

Here ϵ is a random ***error term*** , which is independent of X and has mean zero, and $X = (X_1, X_2, \ldots, X_p)$. In essence, statistic learning refers to a set of approaches for estimating f . We know that the function f that connects input and output is but it general unknown. So how to estimate f based on the observed points is the key in statistic learning.

1.2 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. ***We want to know what is the output $Y=f(X)$ from the a new given input X.*** We assume $\hat{Y}$ is the prediction of Y :

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

The accuracy of $\hat{Y}$ depends on two quantities: the reducible error $\hat{f}$, which we can use techniques for improving the estimate of f; and the second is error term ϵ, called irreducible error, beacuse it cannot be predicting by using X. Statistics learning focus on techniques for estimating f with the aim of minimizing the reducible error.

Inference

We are often insterested in ***how Y changes as a function of $X_1, X_2, \ldots, X_p$.*** The following questions may be attracted our attention:

- Which features are associated with the response?
- What is the relationship between the response and each feature?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

1.3 How Do We Estimate f

Our goal is to apply a statistical learning method to the `training data` in order to estimate the unknown function f. In other words, we want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y). Broadly speaking, most of statistics 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. Make an assumption about the form of f. For example, one very simple assumption is that f is linear in X: $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$.
2. Use the training data to fit or train the model. In the case of the linear model, we need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$.

The advantages of this method is that it reduces the problem of estimating f down to one of ***estimating a set of parameters***, which is generally much easier to estimate.

The potential disadvantage of a parametric approach is that the model we choose will usually ***not match the true unknown form of f.*** To address this problem, we can choose ***flexible model*** that requires estimating a greater number of parameters which can lead to a phenomenon known as ***overfitting***.

Non-parametric methods

Non-parametric methods ***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 that has ***the potential to accurately fit a wider range of possible shapes for f.*** But non-parametric approaches do suffer from a major disadvantage: ***a very large number of observations*** is required in order to obtain an accurate estimate of f.

1.4 The Trade-Off Between Prediction Accuracy and Model Interpretability

There are lost of types of approaches for application, but why would we ever choose to use a more restrictive method instead of a very flexible approach? **That's depends on our goal: prediction or inference**.

- Prediction Accuracy V.S. Interpretability. In general, as the flexibility of a method increases, its interpretability decreases.
- Good Ft V.S Over-fit or Under-fit. A important question is how do we know when the fit is just right?
- Parsimony V.S Black-box. Sometimes, we prefer a simpler model involving fewer variables but with greatest explanatory power.

1.5 Supervised Versus Unsupervised Learning

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

- Supervised. For each observation of the predictor x_i, there is an associated response 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. We observe a vector of measurements x_i ,but no associated response y_i . We seek to understand the relationships between the variables or between the observations.
- Semi-supervised. In a training data set, we have some predictor measurements and a response measurement, and the left data only have predictor measurements but no response measurement.

1.6 Regression Versus Classification Problems

Variables can be characterized as either ***quantitative*** or ***qualitative*** (also known as ***categorical***).

- Quantitative variables take on numerical values, and the model refers to quantitative problems named ***regression problems*** .
- Qualitative variables take on values in one of K different classes, or categories, while those involving a qualitative response are often referred to as ***classification problems*** .

2. Assessing Model Accuracy

It is an important task to decide for any given set of data which method produces the best results.

2.1 Measuring the Quality of Fit

Suppose we fit a model $\hat{f}(x)$ to some training data Tr = $\{(x_i, y_i)\}_1^n$, we wish to see how well the model $\hat{f}$ performs. We could compute the ***mean square error(MSE)*** over Tr:

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

The MSE will be ***small*** if the predicted responses are very ***close*** to the true responses.

2.2 The Bias-Variance Trade-Off

The ***expected test MSE*** , for a given value x_0 , can 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:

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

From this equation, to minimize the expected test error we need to simultaneously achieve ***low variance*** and ***low bias*** . As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease.

2.3 The Classification Setting

Suppose that we seek to estimate f on the basis of training observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where now $y_1, \ldots, y_n$ are qualitative. The most common approach for quantifying the accuracy of our estimate $\hat{f}$ is the ***training error rate*** :

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

A good classifier is one for which ***the test error is smallest*** . Here are some examples: The Bayes Classifier and K-Nearest Neighbors.