Statistical Learning

Problem Setup

Chapter 2: Statistical Learning

Let's dive into statistical learning by framing a typical problem setup:

Input variables (X1, X2, X3, … , XP); P predictors

Output variables (Y); response or dependent variable

We can mathematically define the relationship between Y and (X1, X2,…, XP) in the following general form:

Y = f(X) + e, where e is the random error term unable to estimate. The function f(X) represents the systematic information of the predictors about Y.

Throughout this course, we will study statistical learning approaches for estimating f.

We have two reasons for estimating f: prediction and inference.

Prediction of Y is the art of utilizing the information of the input variables to produce an accurate prediction of Y. The accuracy of the prediction depends on two quantities: reducible error and irreducible error.

The reducible error is the amount of error which can be reduced by using the most appropriate statistical techniques to estimate f.

But even if we find the perfect statistical model for f, we would still need to deal with the unknown quantity of the random error term. This is the irreducible error term. It is non-zero since it may contain unmeasured variables that are useful in predicting Y.

Focus of this course is to use statistical techniques to minimize the reducible error.

Inference about Y: In this scenario, we are interested in understanding the way that Y is affected as the predictors change values. Which predictors are more influential? Do we need to keep all the input variables in the model? Are the predictors correlated to each other?

We still need to estimate f, but a prediction might not be necessary.

Here are some key questions around inference of Y:

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?

- How complicated of a model do we need to capture accurately the relationship between the response and the predictors? Is linear model appropriate? Or a more complicated system of equations?

Depending on whether our ultimate goal is prediction, inference and/or a combination of the two, different methods will be considered and discussed.

Linear models are relatively simple to implement and easy to be utilized for inference, but may not yield satisfactory results in terms of accuracy.

Non-linear models tend to have the potential of providing very accurate predictions of Y, but this comes at the expense of a less interpretable model for which inference is more challenging.

To balance linear and non-linear models, parametric and non-parametric methods derived from a training data set (observed data points used for model construction) are considered.

Parametric Methods

Let's study the following two-step model-based approach:

Step 1: Need to define the shape of f. A typical example is a linear model:

$$f(x) = bo + b1X1 + b2X2 + ..+ bpXp$$

Looking at this equation, and since the predictors X1,…,Xp are observed in the training data set, one only needs to estimate the coefficients bo,…..bp.

Step 2: After the model is selected, utilize different approaches like the least squares method for fitting the model.

We can clearly see how parametric methods simplify our efforts to derive the best possible estimation for f. But one big disadvantage in many real-life examples, is that parametric models are not very accurate. Analysts sometimes try to introduce more parameters with the unfortunate result of overfitting the data and producing inaccurate predictions.

But, we do have non-parametric methods. These are techniques that avoid a specific functional form for f and allow different shapes for f to be considered for predictions. A typical example of a non-parametric technique and one to be covered in this course, is the

KNN method, or the K nearest neighbor. Non-parametric regression models which allow for thin-plate splines in a multidimensional space are also very popular. A key requirement for employing non-parametric techniques is the large data volume to avoid inaccurate estimates for f.

Statistical learning is an art! Advantages and disadvantages need to be considered for both approaches and only an experienced analyst will be able to make a final determination on the most appropriate model. It is also critical to use the perfect training data for the model construction.

What is the trade-off between prediction accuracy and model interpretability? The best way to make a decision is to utilize a two-dimensional plot, where Y-axis is interpretability and the x-axis is flexibility. The analysts can input the models under consideration in the plot before making the final conclusions on the model selections. Figure 2.7 in the textbook illustrates some of the methods studied in this course.

Assessing Model Accuracy

How do we determine the accuracy of the models? Are there different ways to measure the quality of fit between regression (Y is a quantitative variable) and classification (Y is a qualitative variable) methods?

In the regression setting, the most commonly used measure is the mean squared error (MSE), the average of the differences between the observed and predicted Y values for the training data set.

The MSE is small when the predicted responses are very close to the true/observed responses. The analyst should always aim for the smallest possible MSE.

In general, however, we are also interested in how well the statistical method works in data that is newly introduced and not part of the training data; these are the test data points.

Consider predicting a stock's price based on previous returns. The modeler cares about how well the model predicts a future price by comparing it to the observed/true value. We can calculate the test MSE and select the model with the smallest possible value.

Test data very often are hard to obtain; we will be discussing creative ways of deriving test data if true test data are not available. One important method is cross-validation.

We can also breakdown the expected test MSE as the sum of three well-known statistics: the variance of the estimated f, which is the amount by which it changes if we use different training datasets for the same models; the bias, which is the reducible error that is introduced by the model; and the variance of the random error term that is unknown.

A good analyst will be reviewing the variance and bias of each one of the models under consideration and determine the optimal environment under which the most appropriate

model will be the best solution for a given problem. The challenge is always to find the model with the smallest amount of bias and variance.

In the classification setting, the output variable Y is no longer numerical. It is a categorical variable with qualitative values, but we still seek to estimate f based on training data.

We evaluate the accuracy of our estimate by calculating the training error rate, the proportion of mistakes that are made if we apply the predicted f to the training observations.

As in the regression setting, however, analysts are more interested in applying their prediction models to test data and very similarly calculate the test error rate. The goal is always to minimize the amount of test error.

In theory, the Bayes Classifier is a very simple classifier that assigns each observation to the most likely class given its predictor values. It calculates conditional probabilities that Y falls into a specific group given an observed vector of predictor values. In a two-class problem, where there are only two possible response values, the Bayes classifier corresponds to predicting class one if the conditional probability is greater than 50% and class two otherwise.

The Bayes classifier produces the smallest possible test error rate, called the Bayes error rate.

In practice, however, we don't know the conditional distributions of Y given X. We would estimate these probabilities and then classify the values of Y into the appropriate groups.

There are methods to get conditional probabilities. A very popular one is the K-nearest neighbor or KNN classifier method with very successful results in different settings.

Chapter 3: Linear Regression Models

The simple linear regression model is the simplest model and it is used to predict Y on the basis of a single predictor X. The key assumption is that there is an approximately linear relationship between X and Y.

Some key terms that have been studied in our Statistical Foundations class are the intercept, slope, and their corresponding regression coefficients.

Regression coefficients are estimated by using the least squares criterion or alternative approaches discussed in Chapter 6.

Based on the principles discussed in Chapter 2, the least squares line produces the coefficient estimates which, by introducing a hypothesis testing context, we can validate

their accuracy. The usual t-statistics and p-values are utilized to confirm statistical significance.

We are also interested in the accuracy of the model by examining the residual standard error and the R-Square statistic.

The multiple linear regression approach is utilized when we have a lot of input variables that we need to consider and review their relationship with the response variable Y. Just like in the simple regression model, we can calculate the multiple least squares regression coefficient estimates.

Here are some key questions the analyst considers in a multiple linear regression setting:

- Is at least one of the predictors useful in predicting the response?

  F-statistic

- Are all the predictors useful to explain Y or a subset?

  Forward selection, backward selection or mixed selection

- How well does the model fit the data?

  Overall R-Square and MSE

- Given a set of predictor values, what response value should we predict and how accurate is the prediction made?

Confidence and prediction intervals

There are, of course, other considerations in the regression setting, such as qualitative predictors, which allow for interaction terms between the predictors and non-linear relationships.

There are also hidden and potential problems that modelers face when they fit linear regression models to particular data sets. Remember these are parametric models with a specific functional form:

- Non-linearity of the response-predictor relationships

- Correlation of error terms

- Non-constant variance of error terms
- Outliers
- High-Leverage Points
- Collinearity, the property of correlation between predictors

There are diagnostic tests available for detecting and correcting all the issues discussed above: residual plots, correlation matrix, and variance inflation factor.

Assignments to discuss in class:

Chapter 3 Lab: Linear Regression