Introduction to Statistical Learning: Chapter 2 Notes

Sean Rollins

September 7, 2021

1 What is statistical learning?

1.1 Estimating f

The output or dependent variable Y is given by:

$$Y = f(X) + \varepsilon$$

where $X = (X_1, X_2, \ldots)$ is a vector of input or independant variables, and ε denotes an error term, which is independent of X and has mean 0.

Prediction is about finding an estimate for Y, $\hat{Y} = \hat{f}(X)$, where $\hat{f}$ is a estimate of f. It is prone to reducible error (minimized by making the estimate that $\hat{Y} = f(X)$) and irreducible error (governed by ε, since Y is also a function of ε). ε may contain variation that cannot be measured or important variables that were not measured, hence it being nonzero. In general:

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

On the other hand *inference* is about seeking f rather than Y and may be used to find and examine a relationship that predictors may have with response. Often the models used in inference are simple and interpretable, while those used in prediction are more akin to black boxes.

1.2 Parametric v. Non-Parametric Methods

In either case finding f is important and can be done either parametrically or non-parametrically.

Parametric methods first involve making an assumption about the form of f (such as $f(X) = \beta_0 + \beta_1 X_1 + \ldots$), which simplifies the problem to estimating a vector $\beta = (\beta_0, \beta_1, \ldots)$. Next, one utlizes training data to best estimate β such that $Y \approx f(X)$. Its efficacy is affected by the reasonableness of our assumption of the form of f.
Non-parametric methods make no explicit assumptions about f's form, but typically require a large number of observations to be accurate.

1.3 Interpretibility vs Flexibility

In general, models with great flexibility like deep learning have little interpretability and models with great interpretibility like linear regression have low flexibility.

1.4 Supervised v. Unsupervised Learning

Supervised learning is statistical learning that, for each observation of predictor measurement, there is a corresponding response measurement.
Unsupervised learning deals with vectors of predictor measurements that have no known correspondig response measurements. Without response data one tries to understand the relationships between the variables or between the observations. One such way this can be done is via clustering, which is an attempt to group observations into distinct groups.
A semi-supervised learning problem is one in which some, but not all observations have response data associated with them.

1.5 Regression vs classification

Regression is to estimate a quantiative variable, and classification is to estimate a qualitative variable. Both regression and classification can be performed with statistical learning.

2 Assessing Model Accuracy

2.1 Measuring The Quality of Fit

In regression it is common to use the Mean-Squared Error to estimate the efficacy of a model, given by

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

where (x_i, y_i) is an observation and the response data associated with it. It's important to draw the distinction between training MSE and *test* MSE. We generally care about the latter and not the former. In other words, given a large number of test observations, we would like to minimize $\mathbf{Ave}(y_0 - \hat{f}(x_0))^2$, where (x_0, y_0) are test observations, not in the training data.

One may think that the method that minimizes the training MSE is likely to minimize the test MSE but this is not necessarily the case. When graphing MSE as a function of flexibility, the training MSE monotonically decreases as flexiblity increases, while the test MSE has a U shape, it starts to decline, but then starts to increase again after some point. *Overfitting* is when the test MSE is much larger than the training MSE.

2.2 Bias-Variance Trade Off

The above can be explained as the interplay between bias and variance. The expected test MSE can be decomposed as follows:

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

Visibly, one must balance the values of the variance and bias of $\hat{f}$. Simply, variance is the amount $\hat{f}$ would change given different training data, and bias is error caused by approximation and is model-dependent.

In general, more flexible methods result in more variance and less bias while the opposite is true for less flexible methods. Also note that $MSE \geq \mathbf{Var}(\varepsilon)$.

2.3 Classification

A counterpart to the MSE for qualitative data is the error rate, given by:

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

where $I(a)$ is an indicator variable that returns 1 if a is true and 0 if it's false. Similarly, there is a training error rate and a test error rate. One wishes to minimize the test error rate.

The test error rate is minimized if we assign an observation with predictor vector x_0 to the class j for which

$$\mathbf{Pr}(Y = j | X = x_0)$$

is greatest. This is called a Bayes Classifier. A Bayes Decision boundary seperated reigons for which a Bayes classifier chooses a given class.

We do not know the distribution of $Y|X$ so methods of approximating it will have to do. One such method is K-nearest Neighbors classifies a test observation x_0 to a class j estimating the conditional probability as:

$$\mathbf{Pr}(Y = j|X = x_0) = \frac{1}{K}\sum I(y_i = j)$$

which is the fraction of it's nearest K neighbors whose responses equal j.

In general, for higher K the less flexible(lower variance, higher bias) the model is. Similarly, the test error rate and training error rate have a similar relationship to that of MSE in regression problems. When flexibility is written $\frac{1}{K}$, and graphed against the error rate, the graph has a similar shape of that of flexibility(degrees of freedom?) and MSE for regression.