Chapter 02: Statistical Learning

Stan Piotrowski

September 22 2021

Contents

Notes

Prediction and inference

- If we have a quantitative response Y and a set of p predictors, we can write a general form of the relationship as $Y = f(X) + \epsilon$
- The error term (ϵ) is a random error term, also referred to as irreducible error, and has mean zero.
- In most situations, the true functional relationship between Y and a set of p predictors is not known, in which case we are estimating the function f in the prediction setting (i.e., where we don't know the output).
- In the prediction setting, we re-write the equation as $\hat{Y} = \hat{f}(X)$, and the process of estimating the functional relationship induces additional error, called the reducible error.
- In an inference setting, we aren't necessarily concerned with predicting a response, but understanding the association between Y and the set of p predictors.
- When we are interested in understanding the association between Y and p predictors, we are looking at identifying the exact functional form of the relationship.

Parametric vs. non-parametric methods

- The most important distinction between parametric and non-parametric methods is the former makes explicit assumptions about the shape of the function f, while the latter does not.
- An example of a parametric method is ordinary least squares regression, which assumes that the relationship between Y and a set of p predictors is linear and is solely explained by the set of p predictors.
- In contrast, an example of a non-parametric method is a thin-plate spline used for interpolation and smoothing, which instead instead of assuming a functional form for f, tries to develop an estimate that is as close as possible to the observed data.

Prediction accuracy and model interpretability trade-off

- When choosing a statistical learning method, there is a trade-off between flexibility and interpretability: flexible methods are generally better for prediction accuracy, but are less interpretable, while inflexible methods have relatively poor prediction accuracy but are relatively easier to interpret.
- For example, if we are interested in inference, inflexible models that reduce the problem to a simplified version of reality with just a few predictors is more desirable: the model performs relatively well on training data and easily generalizes to unseen test data.
- Further, inflexible methods allow us to generalize and understand the nature of the relationship (i.e., interpret the model) between the response and predictors.

- On the other hand, if we are interested in prediction, flexible models can capture further intricaces in the data, but suffer from overfitting (essentially capturing noise in the data rather than the true functional relationship).

Unsupervised vs. supervised learning methods

- In supervised settings, we are training a learning method using a set of predictors and known outcomes, then applying the model to unseen test data and essentially comparing how often the model identifies the true known response.
- In an unsupervised setting, we don't have outputs; rather, we are interested in discovering meaningful patterns just using the data.
- An example of supervised learning is the ordinary least squares regression; clustering using principal components analysis is an example of unsupervised learning.

Regression vs. classification problems

- In general, quantitative responses can be thought of as regression problems (although there are exceptions like logistic regression for categorical data), while qualitative responses can be thought of as classification problems.

Evaluating model accuracy

- The most important rule to remember is that there is no single "best" method for all data sets and careful exploration and interpretation is needed to select a model which balances the trade-off between flexibility and interpretability.
- The mean squared error (MSE) is a metric used in the regression setting to quantify how close the predicted response is from the observed response for a particular value of predictor.

- In essence, MSE is the mean of the squared differences between the predicted value and the observed response value.
- Interpreting the MSE is simple: the smaller the MSE, the closer the model predicted values are to the observed values.
- Importantly, simply because a given model out of a set has the lowest training MSE doesn't mean that it will have the lowest test MSE- this highlights the importance in balancing flexibility and the ability of the model to generalize to unseen test data.
- Degrees of freedom in statistical learning is the quantity that describes the flexibility of a curve to data: more flexible methods have higher degrees of freedom and generally fit the data more closely, but at the expense of potentially performing poorly on test data (generalization problem).
- When evaluating model flexibility and accuracy, as the flexibility of the method increases, the training MSE will consistently decrease, while the test MSE will exhibit a U-shaped curve: it will decrease up to a certain inflection point, beyond which it increases rapidly because of the generalization problem.
- We can decompose the expected test MSE using the following equation:

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

- The first term, $E(y_0 - \hat{f}(x_0))^2$, describes the expected test MSE, or the mean test MSE if we were to calculate the test MSE over successive training data sets.
- The second term, $Var(\hat{f}(x_0))$, describes the variance in the functional form of f: essentially, how much does f change over successive training data sets.
- The third term, $[Bias(\hat{f}(x_0))]^2$, is the absolute value of the squared bias, or the error introduced by approximating reality with a simplified model.
- Finally, the last term, $Var(\epsilon)$, is the variance in the irreducible error term over successive training data sets.
- In general, using increasingly flexible methods will lead to an increase in $Var(\hat{f}(x_0))$ and a decrease in $[Bias(\hat{f}(x_0))]^2$; the opposite is true for inflexible models.

Classification

- The Bayes classifier is considered the gold-standard and evaluates the conditional probability of a response belonging to a specific class given a value of the predictor variable: the assigned class is whichever has a probability > 0.5.
- However, using the Bayes classifier requires that we know the conditional distribution of Y given X, which is generally unknown.
- A close approximation is the K-nearest neighbors (KNN) classifier, which looks at each test observation in training data and assigns a conditional probability based on the fraction of K points that are closest to it; the assigned class is the one with the highest estimated conditional probability.
- In general, higher numbers of K using the KNN classifier decreases flexibility and the variance in functional form at the expense of increasing bias.

Exercises

Conceptual

1) This exercises poses a question regarding the performance of flexible statistical learning methods relative to inflexible methods.

   a) For each of these scenarios, we need to consider the variance of the function describing the relationship between the response and the predictor(s) between successive training sets and the bias, or the error introduced when models are used to simplify reality. Both of these metrics factor into the overall test mean squared error (MSE), or the mean squared difference between the predicted value or the value estimated by the function and the observed value (in the case of supervised statistical learning).

   b)

   c) In a scenario where the relationship between the predictors and the response is highly non-linear, we would expect the flexible statistical learning method to be better than an inflexible method (generally). Inflexible learning methods, like ordinary least squares regression, for example, will likely have a low variance between successive training data sets, but have a large bias because we are attempting to explain a non-linear relationship with assumptions in a simplified linear reality.

   d) In contrast to c), in a scenario where the variance of the error terms is extremely high, an inflexible method may be preferred because fitting a flexible statistical model on successive training data sets that are substantially different could be driving the high variance.

Applied