Chapter 4: Classification

In this chapter, we study approaches for predicting qualitative responses, a process that is known as classification. Predicting a qualitative response for an observation can be referred to as classifying the observation to a specific category or class.

The following techniques include some of the most widely used classifiers: logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and K-nearest neighbors (KNN). We will discuss more computer-intensive methods, such as trees, random forests and boosting, in later chapters.

Classification problems occur often in life. Here are some examples:

- Identifying the medical condition of a patient based on a series of medical tests (IBM Watson)
- Classifying a credit card transaction as fraudulent on the basis of variables such as user's IP address and past transaction history (Credit Card Companies)
- Predicting a loan getting defaulted based on the characteristics of customers (income, balance, credit score) with loan history (EQUIFAX)

Logistic Regression

Consider the case where the response Y falls into one of two categories, Yes or No. The logistic regression models the probability that Y belongs to a particular category.

For the one predictor case X, logistic regression models the probability of Pr(Yes|X) or Pr(No|X). The values of Pr(Yes|X) or Pr(No|X) will range between 0 and 1.

The Default data in the textbook includes customer default records for a credit card company. Let's model the Pr(default=Yes| Balance) to predict default=Yes using balance as the predictor. We will use the logistic regression model (equation 4.4) to model the above probabilities.

The coefficients b0 and b1 are unknown and must be estimated based on the available training data. The maximum likelihood method will be utilized to derive the estimates. The method works as follows: seek estimates for the coefficients such that the predicted probability of default for each individual, corresponds as closely as possible to the individual observed status.

Ideally, we would like Pr(default=Yes|Balance)=1 for all individuals who defaulted and a number close to zero for all others.

If the regression coefficients are statistically significant, we can then move forward with making predictions by utilizing the logistic regression model.

For the Default data, the predicted probability of default for an individual with a balance of $2,000 is 58.6%.

We can definitely use other predictors to recalculate the logistic regression model as well as using multiple predictors in a model – Multiple Logistic Regression.

Example of Multiple Logistic Regression: Table 4.3 shows the coefficient estimates for a logistic regression model that uses balance, income and student status to predict probability of default.

Multiple-Class Logistic Regression can also be constructed such that the response Y has more than two categories. For example, we want to classify a specific group of patients into the following three categories: stroke, drug overdose and epileptic seizure.

The two-class logistic regression models have multiple-class extensions, but in practice they tend not to be used all that often. The reason for this is that the LDA method discussed next is very popular and easier to interpret findings.

Linear Discriminant Analysis (LDA)

Why do we need another method, when we have logistic regression?

- When the classes of the response are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. LDA does not suffer from this problem.

- For a small sample size and the distributions of the predictors X are approximately normal in each of the classes of Y, the LDA is again more stable than the logistic regression.

- LDA is more popular because it is easier for interpretation when we have more than two response classes.

How does LDA work?

Suppose we want to classify an observation into one of K classes, where K is greater than or equal to 2. This implies that Y can take on K possible distinct and unordered values.

We now model the distribution of the predictors X separately in each of the response classes and then use the Bayes theorem to flip these around into estimates for $Pr(Y=k| X)$.

If p=1, where we only have one predictor, the LDA classifier assigns an observation X=x to the class for which the discriminant function (equation 4.17) attains the largest value. The word linear in the classifier's name derives from the fact that the discriminant functions are linear functions of x.

Note that the LDA method approximates the Bayes classifier, the ideal classifier which unfortunately we are not able to calculate in real-life situations, since we don't know the conditional distribution of Y given X for real data.

We can also extend the LDA classifier to the case of multiple predictors (p>1). We assume that the vector of predictors X is drawn from a multivariate Gaussian distribution with a class-specific mean vector and a common covariance matrix to all classes K.

The LDA classifier assumes that the observations are drawn from a multivariate Gaussian distribution, with a discriminant function. (equation 4.19)

Let's take the Default data and perform LDA to predict whether or not an individual will default on the basis of credit card balance and student status. The LDA model fit to the 10,000 training samples results in a training error of (252+23)/10000 = 2.75%. Keep in mind, this is a training and not a test error rate. We would expect the test error rate to be higher since we have to predict the default status.

How about if we just use a null classifier to always predict that each individual will not default? Then we know that that the error rate will be 3.33%, which is very close to the sophisticated LDA error rate of 2.75%.

The reason why these two error rates are close is because we don't have a lot of defaults in the data (there only 333 out of 10,000).

There are two types of errors that we are interested in: the algorithm can incorrectly assign an individual who defaults to the no default category or it can incorrectly assign an individual who does not default to the default category. A confusion matrix (Table 4.4) is a convenient way of displaying these two types of errors.

Sensitivity v. Specificity

Sensitivity is the percentage of true defaulters that are identified.

Specificity is the percentage of non-defaulters that are correctly identified.

The ROC curve is a popular graphic for simultaneously displaying the two types of errors for all possible thresholds of probabilities across the different classes K of Y. It is a line plot of the true positive rate (sensitivity rate) versus the false positive rate (1- specificity). The larger the area under the curve (AUC), the better the performance of the classifier.

Quadratic Discriminant Analysis (QDA)

The QDA classifier results from assuming that the observations from each class of Y are drawn from a Gaussian distribution, and plugging in estimates for the parameters into Bayes' theorem in order to perform prediction.

However, unlike the LDA classifier, QDA assumes that each class of Y has its own covariance matrix. The discriminant function for the QDA classifier is included in equation 4.23.

Why do we care about the covariance matrix of the K classes and whether or not they have a common covariance matrix?

The answer lies in the bias-variance trade off.

LDA is a much less flexible classifier than QDA and it has substantially lower variance which could lead to a potentially improved prediction performance. The tradeoff is that if LDA's assumption of the common covariance matrix is wrong, then LDA can suffer from bias.

Rule of thumb: Use LDA for relatively few training observations. QDA is recommended if the training set is very large.

K-nearest Neighbors Method (KNN)

The KNN classifier is a model that tries to estimate the conditional probability of Y given X, and then classify each observation to the class with the highest probability.

Given a positive integer K and a test observation x, the KNN classifier first identifies the K points in the training data set that are closest to x, represented by N. It then estimates the conditional probability for a given class as the fraction of points in N whose response values equal that specific class. Finally, KNN applies the Bayes rule and classifies the test observation x to the class with the largest probability.

Figure 2.14 illustrates how the KNN method works. The choice of K has a drastic effect on the KNN classifier obtained. When K=1, the decision boundary is very flexible which corresponds to a classifier that has low bias but high variance. As K grows, the method becomes less flexible and produces a decision boundary that is close to linear.

The analyst needs to consider different values of K, as well as different sizes of training sets, before choosing the appropriate KNN classifier. Remember that statistics is a form of art!

Choosing the correct level of flexibility is critical to the success of any statistical learning method.

Comparison of Classification Methods

Since logistic regression and LDA differ only in their fitting procedures, the two approaches often give similar results, especially when we only have two classes of Y.

However, for more than two classes of Y, LDA works better and is more stable than the logistic regression.

If the distribution of the training data observations don't follow a Gaussian distribution, then the logistic regression can outperform LDA.

Remember that KNN is a non-parametric approach. No assumptions are made about the shape of the decision boundary. Therefore, we expect that KNN is superior to LDA and logistic regression when the decision boundary is highly non-linear.

KNN, however, does not tell us which predictors are important!

QDA serves as a compromise between the non-parametric KNN and the linear LDA and logistic regression methods.

The best way to choose the appropriate method is to perform side-by-side boxplots of the test error rates for each method and compare their respective distributions.

Assignments for class discussion

Review of Lab in Chapter 4:

Discussion of the R code for fitting logistic regression, LDA, QDA and KNN.