**Chapter 4: Classification**

This chapter cover various classification methods. Classification is the process of predicting a qualitative response for an observation and involves assigning the observation to a category, or class. Qualitative variables are also referred to as categorical variables. These terms are often used interchangeably

There are many possible classification techniques, or classifiers, used to predict a qualitative response, with the three most widely-used classifiers being: logistic regression, linear discriminant analysis, and K-nearest neighbors.

**4.1 An Overview of Classification**

Classification problems are common in ecological analysis. Just as in the regression setting, classification uses a set of training observations (x1, y1), . . . , (xn, yn) to build a classifier. Performance of a classifier is then assessed using test observations that were not used to train the classifier.

**4.2 Why Not Linear Regression?**

Linear regression is not appropriate in the case of a qualitative response, because there is no natural way to convert a qualitative response variable with two or more levels into a quantitative response for linear regression.

The problem is that when linear regression is applied to binary (two level) qualitative response, some of the estimates can be outside the prediction [0, 1] interval, making them hard to interpret as probabilities!

Things only become more challenging when analyzing qualitative responses with more than two levels, and thus it preferable to use a classification methods which are better suited for qualitative response values

**4.3 Logistic Regression**

Logistic regression applies to data where the response falls into one of two categories (e.g., 1/0, Yes/No, Presence/Absence, etc…). Instead of modeling the response Y directly, logistic regression models the probability that Y belongs to a particular category.

Pr(Y = 1|X)

This expression can be stated as ‘the probability of Y given X’. This means that for any value of X a prediction can be made for the probability of Y. The values of Pr(Y = 1|X) are abbreviated as *p*(X) and range between 0 and 1.

**4.3.1 The Logistic Model**

The relationship between *p*(X) = Pr(Y = 1|X) and X is modelled using the logistic function,

The logistic function is fitted to the data using maximum likelihood, which produces an S-shaped curve across the range of X values.

The quantity p(X)/[1−p(X)] is called the odds, and can take on any value between 0 and ∞.

Odds values close to 0 and ∞ indicate very low and very high probabilities, respectively. This expression can be linearized by taking the logarithm of both sides of the expression, which results in the log-odds or logit:

A logistic regression model allows us to establish a relationship between a binary outcome variable and predictor variables by modeling the logit-transformed probability as a linear relationship with the predictor variables.

Unlike linear regression where *β*1 gives the average change in Y associated with a one-unit increase in X, the estimate of *β*1 for logistic regression represents a change in the log odds for each unit increase in X. This is equivalent to multiplying the odds by e*β*1

The amount that *p*(X) changes due to a one-unit change in X will depend on the current value of X, but if *β*1 is positive then increasing X will be associated with increasing *p*(X), and if *β*1 is negative then increasing X will be associated with decreasing *p*(X).

**4.3.2 Estimating the Regression Coefficients**

The coefficients *β*0 and *β*1 are unknown and must be estimated based on training data. Logistic regression models are fit using maximum likelihood because it has better statistical properties than other approaches such as (non-linear) least squares.

The maximum likelihood approach aims to estimate values for *β*0 and *β*1 such that the predicted probability of Y for each individual is as close as possible to the individual’s observed Y value. More simply stated, maximum likelihood tries to find and estimates for *p*(*X*) that yields numbers close to one for all individuals in one category and numbers close to zero for all individuals in the other category

Likelihood function:

The estimates and are chosen to maximize this likelihood function. The estimated intercept is typically not of interest but it adjusts the average fitted probabilities to the proportion of ones in the data.

The significance of tests for estimated coefficients of the logistic regression models are based on the z-statistic. The z-statistic associated with is equal to . Large z-values provide evidence against the null hypothesis H0: = 0 (i.e., the probability of Y does not depend on X).

**4.3.3 Making Predictions**

Once the coefficients have been estimated from the training data, it is simply a matter of inserting values of *X*i in to the logistic function to calculate probabilities for each observation:

**4.3.4 Multiple Logistic Regression**

Logistic regression can be extended to predict a binary response using multiple predictors:

where X = (X1, . . .,Xp) are *p* predictors. The logistic function can be rewritten as:

It is important to note, if there is correlation among the predictors, the results obtained using one predictor may differ considerably from those obtained using multiple predictors (i.e., confounding)

**4.3.5 Logistic Regression for >2 Response Classes**

Multiple-class extensions to the logistic regression model do exist but are not used all that often. Methods such as discriminant analysis is more commonly applied when dealing with categorical data with multiple different classes.

**4.4 Linear Discriminant Analysis**

In linear discriminant analysis (LDA) models the distribution of the predictors X separately in each of the response classes (i.e. given Y), and then use Bayes’ theorem to estimate Pr(Y = *k*|X = *x*).

Several reasons for using LDA instead of logistic regression:

1. When classes are well-separated, LDA parameter estimates are more stable.
2. LDA is also more stable in situations were *n* is small and the distribution of the predictors X is approximately normal in each of the classes.
3. Linear discriminant analysis is generally easier to interpret when there are more than two response classes.

**4.4.1 Using Bayes’ Theorem for Classification**

An approximate the Bayes classifier is used to classify an observation into one of K classes based on whichever class has the largest *pk*(X).

Bayes’ theorem states:

*pk*(X) is the posterior probability that an observation X = *x* belongs to the *k*th class given the predictor value for that observation. *pk*(X) can be calculated indirectly by plugging in estimates of *πk* and *fk*(X).

*πk* represent the overall or prior probability that a randomly chosen observation comes from the *k*th class, and values of *πk* can be calculated based on the fraction of the training observations assigned to the *k*th class (assuming a random sample of Ys from the population).

*fk*(X) is the density function of X for an observation that comes from the *k*th class, however, estimating *fk*(X) can be challenging. In general, *fk*(*x*) is relatively large if there is a high probability that an observation in the *k*th class has X ≈ *x*, and *fk*(*x*) is small if it is very unlikely that an observation in the kth class has X ≈ x.

**4.4.2 Linear Discriminant Analysis for p = 1**

The linear discriminant analysis (LDA) plugs estimates for *πk*, *μk*, and *σ*2 into the (approximate) Bayes classifier:

where an observation X = *x* is assigned to the class for which *δk*(*x*) is largest (assuming a Gaussian distribution within each class). *μ*k is simply the average of all the training observations from the *k*th class:

where *n* is the total number of training observations, and *nk*is the number of training observations in the *k*th class. is a weighted average of the sample variances for each of the *K* classes.

πk denotes the prior probability that an observation belongs to the *k*th class and can be estimated based on the proportion of the training observations that belong to that particular class.

In brief, the LDA classifier assumes that observations within each class come from a normal distribution with a class-specific mean vector and a common variance *σ*2, and is calculated by plugging estimates for these parameters into the (approximate) Bayes classifier. Quadratic discriminant analysis (QDA) differs from LDA in that QDA allows observations in the *k*th class to have a class-specific variance, , and therefore, has a less stringent set of assumptions.

**4.4.3 Linear Discriminant Analysis for p >1**

To extend the LDA classifier to the case of multiple predictors, it is assumed that X = (X1, X2, . . ., Xp) is drawn from a multivariate Gaussian (or multivariate normal) distribution, with a class-specific multivariate mean vector and a common covariance matrix. The (approximate) Bayes classifier assigns an observation X = *x* to the class with the high value of:

Estimates of the unknown parameters μ1, . . . , μK, π1, . . . , π*K*, and Σ use formulas similar to those used in the one dimensional LDA (described above). Note that is a linear function of *x*, and therefore, the LDA decision rules are derived from *x* through a linear combination of its elements. Bayes decision boundaries are then calculated based on comparisons between pairs of *K* classes and used to divide the predictor space into separate regions.

The performance of an LDA model (i.e., classification success rate) can be determined using a confusion matrix. A receiver operating characteristics (ROC) curve can also be used to display the sensitivity and specificity of the model summarized over all possible thresholds. The sensitivity (i.e., true positive rate) is the fraction of observations that are correctly identified, using a given threshold value and 1-specificity (i.e., false positive rate) is the fraction of observations that are incorrectly classified, using that same threshold value. The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate resulting in a larger area under the (ROC) curve the AUC. A classifier that is expected to performs no better than chance will have an AUC of ~0.5 (when evaluated on an independent test set not used in model training).

**4.4.4 Quadratic Discriminant Analysis**

Similar to LDA, a quadratic discriminant analysis (QDA) classifier still assumes that observations from each class are drawn from a Gaussian distribution, but unlike LDA, QDA assumes that each class has its own covariance matrix.

Under this assumption, the (approximate) Bayes classifier is:

By plugging estimates for Σk, μk, and πk into the QDA classifier, individual observations X = *x* are assigned to the class for which this quantity is largest.

The choice between LDA and QDA lies in the bias-variance trade-off. The LDA model assumes that the K classes share a common covariance matrix resulting in *Kp* linear coefficients being estimated. On the other hand, QDA estimates a separate covariance matrix for each class, for a total of *Kp*(*p* + 1)/2 parameters being estimates. As a consequence, LDA has substantially lower variance but is a much less flexible classifier than QDL. Moreover, LDA can be badly biased if the assumption that the *K* classes share a common covariance matrix is not met.

LDA tends to perform better than QDA if there are relatively few training observations, and thus, reducing variance becomes more important. If the training set is very large, or if the assumption of a common covariance matrix for the *K* classes is clearly not met, then QDA is recommended

**4.4.5 K-Nearest Neighbors Method**

The K-Nearest Neighbors (KNN) method is fundamentally different from the other classifiers in that it is a completely non-parametric approach that makes no assumptions about the shape of the decision boundary. Therefore, this approach is expected to out perform LDA and logistic regression when the decision boundary is highly non-linear. KNN is more flexible the QDA but may not perform as well when the number of training observations is limited. The KNN method requires selection of number of neighbors K. To make a prediction for an observation X = x, the K training observations that are closest to x need to be identified and X is assigned to the class to which these combined nearest-neighbors belong.

**4.5 A Comparison of Classification Methods**

When the true decision boundaries are linear, LDA and logistic regression approaches are expected to perform well. When the boundaries are moderately non-linear, QDA may perform better. When decision boundaries are much more complex, the non-parametric approach KNN can become the best performing classifier, although the level of smoothness for a non-parametric approach must be chosen carefully.

Logistic regression can be made more flexible by including X2, X3, and even X4 as predictors, assuming the increase in variance due to the added flexibility is offset by a sufficiently large reduction in bias. The flexibility of LDA can also be increase by adding additional quadratic terms, however, if all possible quadratic terms and cross-products are added to an LDA model, the form of the model will be equivalent to the QDA model, but the parameter estimate would likely be different.