Chapter 4 Classification

Classifying an observation: Predicting a qualitative response for an observation.

Classifiers: Many possible classification techniques, or classifiers, that one might use to predict a qualitative response.

1. Logistic regression
2. Linear Discriminant Analysis
3. K-nearest Neighbors

4.1 An Overview of Classification

In the classification setting, we have a set of training observations that we can use to build a classifier.

We want our classifier to perform well not only on the training data, but also on test observations that were not used to train the classifier

4.2 Why Not Linear Regression?

1. In general, there is no natural way to convert a qualitative response variable with more than two levels into a quantitative response that is ready for linear regression.

\*For a binary response, regression by least squares is possible->use the dummy variables approach

\*The dummy variable approach can not be extended to accommodate qualitative response with more than two levels

4.3 Logistics Regression

Logistic regression models the probability that Y belongs to a particular category



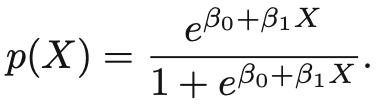
1. For any given value of balance, a prediction can be made for default

4.3.1 The Logistics Models

Problem with linear regression for binary response:

Ant time a straight line is fit to a binary response that is coded as 0 or 1, in principle we can always predict for some value of and for others

Logistics Function:



1. The output is between 0 and 1.
2. Doe low balances now we predict the probability of default as close to, but never below, zero
3. For high balance, we predict a default probability close to, but never above, one
4. Use “maximum likelihood” method to fit the function

A picture containing drawing

Description automatically generated

LHS: odds, can take on any value between 0 and infinity.

\*Values of the odds close to 0 indicates very low probabilityA drawing of a face

Description automatically generated

LHS: Log-odds/logit.

\*logistic regression has a logit that is linear in X.

\*Increasing X by one unit changes log odds by /the odds is multiplies by

\* The amount that p(X) changes due to one-unit change in X will depend on the current value of X.

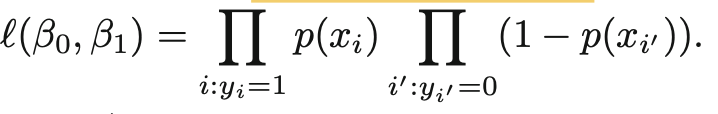
\*If is positive then increasing X will be associated with increasing , if is negative then increasing X will be associated with decreasing p(X)

* + 1. Estimating the Regression Coefficients

Maximum Likelihood method:

1. Estimating , such that plugging these estimates in the model for p(X) yields a number close to one for all individuals who defaulted, and a number close to zero for all individuals who did not.

Likelihood Function:



, are chosen to maximize the likelihood function

\*Z-statistic plays the same role as the t-statistic in the linear regression

* + 1. Making Predictions

The logistic regression model can be used in the qualitative predictors model using the dummy variable approach.

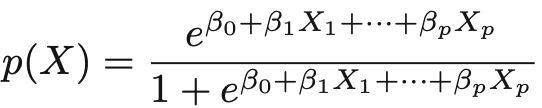
* + 1. Multiple Logistic Regression

General Equation of predicting a binary response using multiple predictors:

A picture containing table

Description automatically generated

Where are predictors.



Use maximum likelihood method to estimate

Confounding: The results obtained using one predictor may be quite different from those obtained using multiple predictors, especially when there is correlation among the predictors.

* + 1. Logistic Regression for >2 Response Classes

Not utilized often🡪 Discriminant analysis is popular for multiple-class classification

* 1. Linear Discriminant Analysis

1. In logistics regression, we model the conditional distribution of the response Y, given the predictor(s) X.

Linear Discriminant Analysis

1. Model the distribution of the predictors X separately in each of the response classes Y
2. Use Bayes’ theorem to flip these around into estimates for

Linear Discriminant Analysis Advantages:

1. The classes are well-separated
2. If n is small and the distribution of the predictors X is approximately normal in each of the classes
3. Linear discriminant analysis popular when we have more than two response classes.

4.4.1 Using Bayes’ Theorem for Classification

Classify an observation into one of classes, where :

1. The qualitative response variable can take on possible distinct and unordered values
2. Prior probability(): Represent the overall or prior probability that a randomly chosen observation comes from the kth class🡪The randomly chosen observation comes from the kth category of the response variable.

Bayes Theorem:

A picture containing object, clock

Description automatically generated

1. : The probability that an observation belongs to the class, given the predictor value for that observation.
2. If we can find a way to estimate then we can develop a classifier that approximates the Bayes classifier.
   * 1. Linear Discriminant Analysis for

Assumptions we need to make:

1. is normal to Gaussian:

A close up of a logo

Description automatically generated

1. are the mean and variance parameters for the kth class.
2. Assume : a shared variance term across all K classes.

Thus,

A close up of a clock

Description automatically generated

Which can be shortened as:

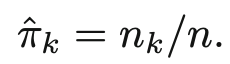
A picture containing object, clock

Description automatically generated

LDA approximates the Bayes classifier by plugging estimates for ,, and

A close up of a clock

Description automatically generated



Where n is the number of training observations, and nk is the number of training observations n the kth class.

LDA classifer plugs estimates and assign an observation X=x to the class for which

A close up of a logo

Description automatically generated

Is largest.

1. The two densities overlap, and so given that X=x, there is some uncertainty about the class to which the observation belongs.
2. LDA classifier results from assuming that the observations within each class come from a normal distribution with a class-specific mean vector and a common variance.
   * 1. Linear Discriminant Analysis for p>1

Multivariate normal distribution:

1. Each individual predictor follows a one-dimensional normal distribution, with some correlation between each pair of predictors.

A picture containing table, computer

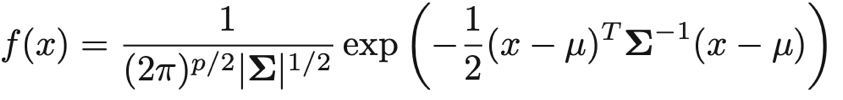
Description automatically generated

1. The height of the surface at any particular point represents the probability that both and fall in a small region
2. If the surface is cut along the axis or along the axis, the resulting cross-section will have the shape of a one-dimensional normal distribution.

Multivariate Gaussian Distribution:

Where is the mean of , and

is the covariance matrix of



\*In the case of the LDA classifier assumes that the observations in the kth class are drawn from a multivariate Gaussian distribution

Bayes classifier assigns an observation to the class for which

A picture containing object, clock

Description automatically generated

Is largest.

1. Bayes decision boundaries divide the predictor space into three regions.
2. The Bayes classifier will classify an observation according to the region in which it is located.
3. The LDA decision rule depends on only through a linear combination of its elements.

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
|  | True(Actual) | False(Actual) |
| True(Prediction) | Sensitivity | False Postive |
| False(Prediction) | False Negative | Specificity |

1. Sensitivity: The percentage of true that are correctly identified
2. Specificity: The percentage of Negative identifies

The Bayes classifier will yield the smallest possible **TOTAL NUMBER OF MISCLASSIFIED OBSERVATIONS**(Irrespective of which class the errors come from)

Trade-off that results from modifying the threshold value for the posterior probability of default.

1. Fraction of defaulting customers that are incorrectly classified
2. Fraction of errors among the non-defaulting customers

ROC curve: False Positive Rate vs. True Positive Rate

Area Under the Curve(AUC):Overall performance of a classifier, summarized over all possible thresholds.

1. An ideal ROC curve will hug the top left corner🡪 The larger the AUC, the better the classifier
2. We expect a classifier that performs no better than chance to have an AUC of 0.5 (when evaluated on an independent test set not used in model training)
3. Varying the classifier threshold changes its true positive(sensitivity) and false positive(1-specificity)

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

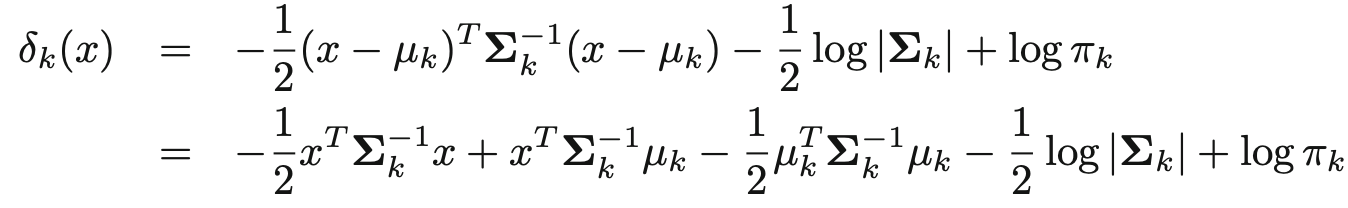
Description automatically generated

* + 1. Quadratic Discriminant Analysis

Assumptions:

1. Observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes’ theorem in order to perform prediction
2. QDA assumes that each class has its own covariance matrix.(An observation from the class of the form , where is a covariance matrix for the kth class.)

Thus, Bayes classifier assigns an observation to the class for which



Is largest.

LDA vs. QDA:

LDA: Estimating a covariance matrix requires estimating

QDA: Estimates a separate covariance matrix for each class, for a total of

1. LDA is much less flexible classifier than QDA, and so has substantially lower variance
2. If LDA’s assumption that the K classes share a common covariance matrix is badly off, then LDA can suffer from high bias
3. LDA is better than QDA if there are relatively few training observations🡪Reducing variance is crucial
4. QDA is recommended if the training set is very large🡪variance is not a big concern, assumption of a common covariance matrix for the K classes is untenable.
   1. A Comparison of Classification Methods

Logistic regression vs. LDA:

1. Both produce linear decision boundaries
2. and (Logistics regression) are estimated using maximum likelihood, whereas and (Linear Discriminant Analysis) are computed using estimated mean and variance from a normal distribution
3. LDA assumes observations are drawn from a Gaussian distribution with a common covariance matrix

KNN is a completely non-parametric approach: No assumptions are made about the shape of the decision boundary

\*Decision boundary is highly non-linear

\*KNN does not tell us which predictors are important; we don’t get a table of coefficients

QDA: serves as a compromise between the non-parametric KNN method and the linear LDA and logistic regression approaches

\*QDA can model a wider range of problems than can the linear methods

\*Perform better in the presence of a limited number of training observations because it does make some assumptions about the form of decision boundary

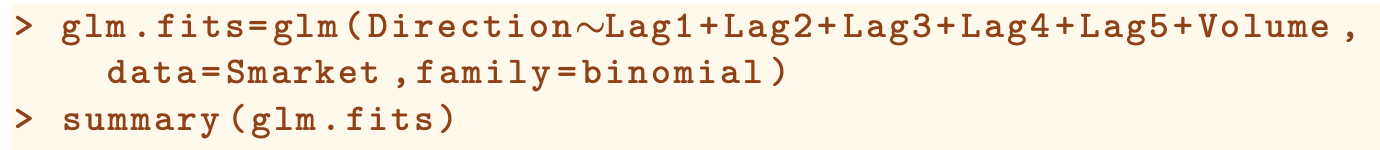
1. When the true decision boundaries are linear, then the LDA and logistic regression approaches will tend to perform well
2. When the boundaries are moderately non-linear, QDA may give better results
3. For much more complicated decision boundaries, a non-parametric approach such as KNN can be superior.
4. We can accommodate a non-linear relationship between the predictors and the response by performing regression using transformations of the predictors.

4.6 Lab: Logistic Regression, LDA, QDA, and KNN

Cor(): function produces a matrix that contains all of the pairwise correlations among the predictors in a data set

Logistic Regression

Glm(): Generalized linear model, need to pass the family=binomial in order to tell R to run a logistic regression



Coef(): Access the coefficients

Predict(): used to predict the probability given values of the predictors//type=”response“ tells R to output probabilities of the form Pr(Y=1|X)

\*If no data set is supplied, then the probabilities are computed for the training data

A picture containing clock, meter

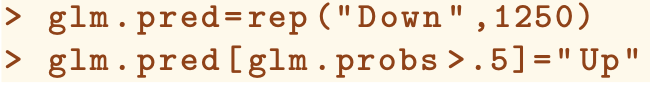
Description automatically generated

Contrasts(): see the dummy variable value

A picture containing table

Description automatically generated

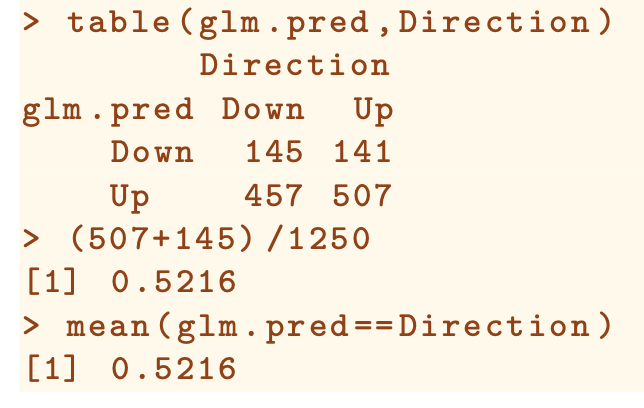
To create a vector of class predictions based on whether the predicted probability of a market increase is greater than or less than 0.5



1. 1st line: The first command creates vector 1,250 Down elements
2. 2nd line: The second line transforms to Up all of the elements for which the predicted probability of a market increase exceeds 0.5.

Table(,Direction): function can be used to produce a confusion matrix in order to determine how many observations were correctly or incorrectly classified

\*The diagonal elements of the confusion matrix indicate correct predictions, while the off-diagonals represent incorrect predictions



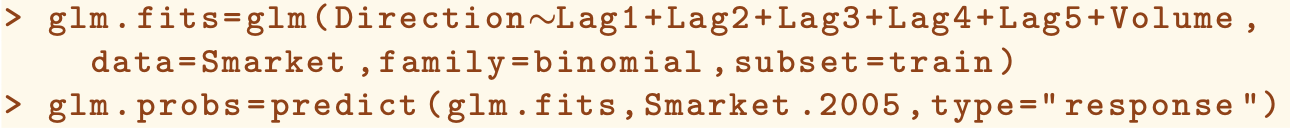
Mean(glm.fit==Direction): used to compute the fraction of days for which the prediction was correct.

1. The training error rate is often overly optimistic- it tends to underestimate the test error rate
2. We can fit the model using part of the data, and the examine how well it predicts the *held out* data. (training data vs. test data)

Boolean vectors can be used to obtain a subset of the rows or columns of a matrix,

1. Smarket[!train,] would pick out a submatrix of the stock market data set, corresponding only to the dates before 2005.

Subset argument in glm():



To predict returns associated with particular values of Lag1 and Lag2, use predict():

Predict(glm.fit2, newdata=data.frame(Lag1=c(1.2,1.5),Lag2=c(1.1,-0.8)), type=”response”)

* + 1. Linear Discriminant Analysis

Lda(): syntax similar to the lm() and glm() except for the absence of the family option, part of the MASS library

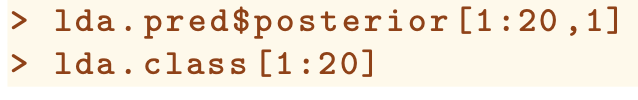
A screenshot of a social media post

Description automatically generated

Group means: These are the average of each predictor within each class, and are used by LDA as estimates of .

Predict(): returns a list of three elemets

1. Class: Contains LDA’s predictions about the movement of the market
2. Posterior: A matrix whose kth column contains the posterior probability that the corresponding observation belongs to the kth class
3. X: contain the linear discriminants, the linear combination that are used to form the LDA decision rule



To change the posterior probability threshold(A posterior probability threshold other than 50%):

A picture containing table

Description automatically generated

* + 1. Quadratic Discriminant Analysis

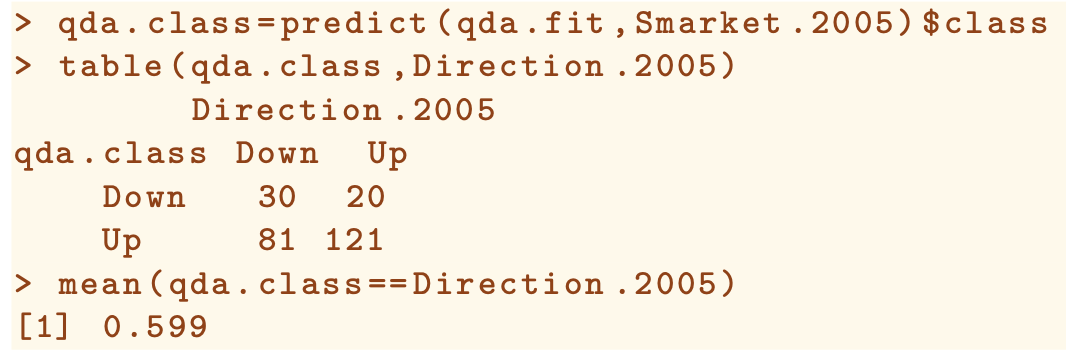
Qda(): syntax identical to lda(), part of the MASS library

A screenshot of a cell phone

Description automatically generated

1. Output contains the group means
2. Does not contain the coefficients of the linear discriminants, because QDA classifier involves quadratic, rather than a linear, function of the predictors.

Predict(): works in the same fashion as for LDA



* + 1. K-Nearest Neighbors

Knn(): forms predictions using a single command, requires four input:

1. A matrix containing the predictors associated with the training data, labeled train.X
2. A matrix containing the predictors associated with the data for which we wish to make predictions, labeled test.X
3. A vector containing the class labels for the training observations, labeled train.Direction
4. A value for K, the number of nearest neighbors to be used by the class

Cbind(): column bind, bind the variables together into a matrix

A picture containing bird

Description automatically generated

* + 1. An Application to Caravan Insurance Dara

1. Because KNN classifier predicts the class of a given test observation by identifying the observations that are nearest to it, the scale of the variables matters.
   1. Any variables that are on a large scale will have a much larger effect on the distance between the observations, and hence on the KNN classifier, than variables that are on a small scale.
   2. We need to standardize the data: So that all variables are given a mean of zero and a standard deviation of one

\*scale(): Exclude the column that has a qualitative variable

A picture containing bird

Description automatically generated

Train-test Split:

A screenshot of a cell phone

Description automatically generated

1. The vector test is numeric, with values from 1 through 1000.
2. Caravan.standardized[test,] yields the submatrix of the data containing observations whose indices range from 1 to 1000.
3. Caravan.standardized[-test,] yields the submatrix containing the observations whose indices do not range from 1 to 1000.