DS-6030 Homework Module 3

Tom Lever

06/08/2023

DS 6030 | Spring 2023 | University of Virginia

5. We now examine the differences between LDA and QDA.

   (a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

   If the Bayes decision boundary is linear, we expect Quadratic Discriminant Analysis to perform better on the training set. According to https://online.stat.psu.edu/stat508/lesson/9/9.2/9.2.8, "QDA, because it allows for more flexibility for the covariance matrix, tends to fit the data better than LDA". We expect Linear Discriminant Analysis to perform better on the test set as the Bayes decision boundary is linear and QDA might overfit the data / follow errors too closely / yield a small training Mean Squared Error but a large test MSE / work too hard to find patterns in the training data and pick up some patterns that are just caused by random chance rather than by true properties of the function relating predictors and response.

   (b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

   If the Bayes decision boundary is non-linear, we expect QDA to perform better on the training set and test set "because it allows for more flexibility for the covariance matrix".

   (c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

   According to https://cseweb.ucsd.edu/classes/sp12/cse151-a/lecture11-final.pdf, "Variance depends on the training set size. It decreases with more training data, and increases with more complicated classifiers". As the sample size n increases, we expect the test prediction accuracy of QDA relative to LDA to improve as QDA is a more complicated, flexible model than LDA with less bias and more variance than LDA and the variance of QDA decreases as sample size increases.

   (d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

   False. As above, we expect Linear Discriminant Analysis to perform better on the test set when the Bayes decision boundary is linear as QDA might overfit the data.

13. This question should be answered using the `Weekly` data set, which is part of the ISLR2 package.

This data is similar in nature to the `Smarket` data from this chapter's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

(a) Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns?

(b) Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables plus `Volume` as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

(e) Repeat (d) using LDA.

(f) Repeat (d) using QDA.

(g) Repeat (d) using KNN with $K = 1$.

(h) Repeat (d) using naive Bayes. (skip this exercise)

(i) Which of these methods appears to provide the best results on this data?

(j) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confu- sion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

14. In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto` data set.

(a) Create a binary variable, `mpg01`, that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median. You can compute the median using the `median()` function. Note you may find it helpful to use the `data.frame()` function to create a single data set containing both `mpg01` and the other `Auto` variables.

(b) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

(c) Split the data into a training set and a test set.

(d) Perform LDA on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

(e) Perform QDA on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

(f) Perform logistic regression on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

(g) Perform naive Bayes on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained? (skip this exercise)

(h) Perform KNN on the training data, with several values of K, in order to predict `mpg01`. Use only the variables that seemed most associated with `mpg01` in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?