ML01 – Spring 2019
Lab 4: LDA, QDA, Naive Bayes and logistic regression

1 Classification of the `Heart` data

The Heart data were collected as part of a study aiming to establish the intensity of ischemic heart disease risk factors in a high-incidence region in South Africa. There are $p = 8$ numeric attributes, and the response variable is the presence or absence of myocardial infarction (MI). There are 160 positive cases in this data, and a sample of 302 negative cases (controls).

1. Load the `Heart` dataset. Plot the data. Which predictors seem to provide useful information for the classification ?
2. Split the data into a training set (approximately 2/3 of the data) and a test set.
3. Train the following classifiers : LDA, QDA, naive Bayes and logistic regression. Compute the corresponding test error rates. (For naive Bayes, use function `naive_bayes` in package `naivebayes`).
4. Plot the ROC curves of the different classifiers on the same graph.

2 Classification of the `Vowel` data

The `Vowel` dataset is composed of feature vectors obtained by recording examples of the eleven steady state vowels of English spoken by fifteen speakers. Words containing each of these vowels were uttered once by the fifteen speakers. Four male and four female speakers were used to build a training set, and the other four male and three female speakers were used for building a test set. After suitable preprocessing, 568 training patterns and 462 test patterns in a 10 dimensional input space were collected.

1. Load the `Vowel` dataset. Plot the data. Which predictors seem to provide useful information for the classification ?
2. Split the data into a training set (approximately 2/3 of the data) and a test set.

3. Train the following classifiers : LDA, QDA, naive Bayes and logistic regression. Compute the corresponding test error rates.
4. Repeat the previous calculations for 10 different partitions training/test of the data and draw a boxplot of the results.