

## Data Mining I

### Homework 4

- 1) (15 points) Consider the Diabetes dataset (posted with assignment). Assume the population prior probabilities are estimated using the relative frequencies of the classes in the data.
  - (a) Produce pairwise scatterplots for all five variables, with different symbols or colors representing the three different classes. Do you see any evidence that the classes may have different covariance matrices? That they may not be multivariate normal?
  - (b) Apply linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). How does the performance of QDA compare to that of LDA in this case?
  - (c) Suppose an individual has (glucose test/intolerance = 68, insulin test = 122, SSPG = 544. Relative weight = 1.86, fasting plasma glucose = 184). To which class does LDA assign this individual? To which class does QDA?
  
- 2) (15 points) This question uses the “Weekly” dataset in the ISLR package. The data contains information for weekly returns for 21 years, beginning in 1990 and ending in 2010.
  - a) Produce some numerical and graphical summaries of the “Weekly” data. Do there appear to be any patterns?
  - b) Use the full data to perform 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? Comment on these.
  - 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) Fit the logistic model using a training data period from 1990-2008, with “Lag2” as the only predictor. Compute the confusion matrix, and the overall correct fraction of predictions (aka misclassification rate) for the held out data (that is, the data from 2009 and 2010).
  - e) Repeat (d) using LDA.
  - f) Repeat (d) using KNN with  $k=1$ .
  - g) Which method appears to provide the best results?
  - h) Experiment with different combinations of predictors, including possible transformations and interactions, for each method. Report the variables, method, and associated confusion 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.

