ETC3250 Lab 8

Di Cook

Week 8

Purpose

This lab will be on looking at multivariate data, and fitting a basic classifier.

Data

- Dr Cook's music data at http://www.ggobi.org/book/. A description of the data can be found at http://www.ggobi.org/book/chap-data.pdf.

Question 1

Read in the music data, from the ggobi web site:

a. Subset the data to drop the "Enya" class. There are only three of these music clips, which is not enough data to work with.

b. Summarise the variables, by class (classical vs rock). Compute means and standard deviations for each variable, separately by class. You can use dplyr's `summarise` function to do this efficiently.

c. Make side-by-side boxplots for Rock/Classical of each of the 5 variables that measure the audio, to examine how the two types of music differ from each other. Explain the differences.

d. Make side-by-side boxplots of the variables by artist. Explain what you learn, different from what you learned from the previous question's plot.

e. Standardise the variables. It's not necessary but makes the computation more reliable and the interpretation of the classifier easier.

f. Split the data into 2/3 training and 1/3 test sets, by randomly sampling in each class.

g. Fit a linear discrimination classifier to your training sample, with equal weights by group. Report the rule, and your error for the test data.

Question 2

Read in the chocolates data, from the class web site. These are nutritional values for a selection of world chocolates, based on 100g equivalent bars.

a. How many different countries are represented?

b. What country makes Jet chocolates?

c. Make side-by-side boxplots of the variables by type of chocolate. Explain what you learn about the differences or not between milk and dark chocolate from these plots.

d. Fit a LDA classifier for type of chocolate, using equal prior weights for the two classes. You should not use MFR, or Name. Why? Report your classification rule.

e. Predict your data. Find a dark chocolate that is misclassified as a milk chocolate. Try your best to work out why it was misclassified, and explain this.

f. Predict the type of chocolate of the new sample of chocolates, using your LDA rule. (An extra credit point if you get them all correct.)

g. There are a number of zeros in the data. Do you think these are really zeros? How might you fix this? (Just a conceptual question, not for you to actually do it.)

WHAT TO TURN IN

Turn in two items: a `.Rmd` document, and the output `.pdf` or `.docx` from running it. Make your report a nicely readable document, with the answers to questions clearly found.