Statistics and Data Analysis in MATLAB

Kendrick Kay, kendrick.kay@wustl.edu

Homework 5 (covering Statistics Lecture 7)

To complete this assignment, prepare a MATLAB script called `homework5.m` along with any necessary accompanying function .m files. Then, run the MATLAB `publish` command (e.g. `publish('homework5.m')`) to run the script and generate HTML output showing the results. Turn in a print-out of the HTML output (e.g. from your web browser) and also a print-out of any function .m files that you write.

Hint: In your script file, place `%%` on a line by itself at each point where you want the HTML output to show figures and command-window text. Please note that your code should be commented (where necessary), including documentation of any functions that you write.

Problem 0. Download the .mat file at http://artsci.wustl.edu/~kkay/psych5007/Homework5.mat (you will need this file to complete the problems below).

Problem 1. The `score1` variable contains scores received by 100 students on some test. The `pass1` variable contains 0s or 1s indicating whether each student ended up passing the course. Visualize these data using a scatter plot. Now suppose we want to use logistic regression to predict whether a student passes the course based on his/her score. Our model is *f(score) = 1/(1+exp(-(a*score+b))* where *a* and *b* are free parameters. Suppose *a* is 1 and *b* is –72. Plot this model on the scatter plot using a red line. Then calculate the likelihood of the data points under this model and report this to the command window. Now suppose *a* is 0.5 and *b* is –32. Plot this model on the scatter plot using a blue line, and report the likelihood of the data points under this model to the command window. Which of the two models results in the higher likelihood? For the model with the higher likelihood, what is the percentage of data points correctly classified?

Problem 2. The `classA` and `classB` variables each contain measurements of 1,000 subjects on three different dimensions. How well can we classify whether a subject belongs to class A or class B based on measurement of the three dimensions? To answer this question, use linear discriminant analysis (LDA) to perform classification, and specifically, use leave-one-out cross-validation to obtain unbiased estimates of classification accuracy. For what percentage of subjects does LDA successfully predict class assignment?