BIOM/SYSC5405 – Pattern Classification and Experiment Design

Assignment 1

Please submit a single **PDF** file with all your answers, discussion, plots, etc. **on BrightSpace**. Also, please include your MATLAB (or R, etc.) code either inline with your answers, or in an appendix. Please order your solution in the same order as the questions below (first Q1a, then Q1b, etc). Make sure it is easy to find your solution to each question (otherwise, solutions may become buried in inline code, for example).

Question 1: Data wrangling

Consider two possible features for a new fruit classification system: weight and diameter. Sample data for each feature is provided in `assigData1.csv`

1000 weight and diameter measurements are given for three types of fruit: apple, orange, and grape. *(File can be easily viewed in Excel or MATLAB. Columns are:* `W_apl D_apl W_orng D_orng W_grp D_grp`*)*

a) To (later) develop a Bayesian classifier, we need to estimate the parameters of the class-conditional distribution for each feature and for each class.

i) Assuming the class-conditional distributions follow multivariate normal distributions with unknown mean and covariance matrix for each class, estimate the three means and the three covariance matrices. For each estimated

ii) For each estimated covariance matrix, compute the determinant and the trace.

iii) For each estimated covariance matrix, compute the eigenvectors and eigenvalues.

iiiv) For each estimated covariance matrix, determine whether it is symmetric and positive semidefinite.

b) Create a scatter plot showing weight vs. diameter for all three classes, colouring the data according to fruit class. Label your axes and add a legend. By examining the scatter plot, for each fruit class, do the data appear to follow a bivariate normal distribution?

c) Plot the histograms for each feature showing the distribution of each feature over each class. For each feature, you should plot all three potentially overlapping histograms representing the three fruit types on a single axis.

i) Use transparency and a different colour and/or line style for each class and make sure you can see all the data (i.e., that bars are not completely occluding each other in your figure).

ii) If you wish to **separate oranges from the other two classes**, which feature would you prefer and why? What if you wanted to **separate grapes** from the other two classes? (150 words)

d) Provide a plot visualizing **apple** weight vs. diameter. Add a line of best fit and report the Pearson Correlation Coefficient. Do the data look correlated? How does your observation compare to the computed Pearson Correlation Coefficient and to the estimated covariance matrix from part a above?