BIOM/SYSC5405 – Pattern Classification and Experiment Design

Assignment 1— Due 11:59pm Tuesday 19 Sept

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.

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 `assigData2.tsv`

100 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 W_orng W_grp D_apl D_orng D_grp`*)*

a) To develop a Bayesian classifier, we need to estimate the parameters of the class-conditional distribution for each feature and for each class. Assuming the class-conditional distributions follow normal distributions with unknown mean and variance for each class, estimate the six means and the six estimates of variance.

b) Plot the histograms for each feature showing the distribution of each feature over each class. For each feature, you should have a single plot (single axis) with three potentially overlapping histograms representing the three fruit types.

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) Which feature would you prefer and why? (150 words)

iii) Illustrate results using **at least two bin widths** when generating your histograms.

c) Provide a plot visualizing **apple** weight vs. diameter. Add a line of best fit and report the Pearson Correlation Coefficient.

Question 2: Generating data & the normal distribution

a) Generate 1000 samples drawn from a trivariate normal distribution with $\mu = \begin{bmatrix} 5 \\ -0.5 \\ 17 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 4 & 0.5 & 0 \\ 0.5 & 5 & -0.2 \\ 0 & -0.2 & 2 \end{bmatrix}$.

You do <u>not</u> need to provide the actual samples in your assignment submission. Instead, report estimates of the mean and variance <u>of the first dimension</u> based on your 1000 samples. Do your estimates agree with the actual values? *(estimates + brief discussion)*

b) Create two scatter plots of the data, e*nsuring that the scale of both axes are equal so that the true shape of the distribution is visible.* The first scatter plot should visualize the first two dimensions of your data. The second scatter plot should visualize dimension 1 vs. dimension 3. Why do their shapes differ? *(25 words)*

c) What is its trace of Σ? Is Σ positive definite? Explain.

d) Calculate and report the eigenvectors and eigenvalues of Σ.

e) Lastly, plot the PDF and CDF for third dimension of your distribution.