BIOM/SYSC5405 – Pattern Classification and Experiment Design

Assignment 1— Due 11:25am Monday 16 Jan 2017

Please submit a single **PDF** file with all your answers, discussion, plots, etc **on CULearn**. Also, please include your MATLAB (or R, etc) code either inline with your answers, or in an appendix.

Question 1: Classifier scores

Consider 2 classifiers, PCI and PSIPRED, which predict protein secondary structure. The classification performance for each method over a set of 125 test proteins is given in the file: assigData1.xls (on CULearn). The following information is given for each test protein: the protein name, the protein length, and the prediction accuracy for each method is given using 3 measures: CC = Matthews' correlation coefficient, Q3=accuracy, BAD=custom error measurement (low is good).

a) Plot the Q3 accuracy vs. protein length for PCI and PSIPRED.

b) Compute the correlation between Q3 accuracy and test sequence length for PCI and for PSIPRED. *Since there are different types of correlation, please choose one and describe it briefly.*

c) What is the mean, median, and standard deviation of the Q3 accuracy for each method?

Question 2: Feature data

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) Use maximum likelihood to estimate the class-conditional distribution parameters of each feature and for each class (i.e. 6 estimates of mean, and 6 estimates of variance) assuming the class-conditional distributions follow normal distributions with unknown mean and variance for each class.

b) Plot the histograms for each feature (one histogram for weight, one histogram for diameter) showing the distribution of each feature over each class. Use 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). Which feature would you prefer and why? **Try at least two** bin widths when generating your histograms.

c) Combine all weight data from all classes into a single vector. Truncate all data to whole numbers. Use a test for normality to check if the data is normally distributed. Describe the test, how it works, and how to interpret your results.

Question 3: Generating data & the normal distribution

a) Generate 1000 samples drawn from a bivariate normal distribution with μ_1=3.2, μ_2=5.1, $\Sigma = \begin{bmatrix} 1.2 & -.5 \\ -.5 & 3.3 \end{bmatrix}$

b) Create a scatter plot of the data, e*nsuring that the scale of both axes are equal so that the true shape of the distribution is visible,*

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

d) Calculate the two eigenvectors and eigenvalues of Σ. Use these to add an ellipse illustrating one line of equiprobability on your scatter plot.

e) Lastly, plot the PDF and CDF for a 1D normal distribution with μ=3.2 and σ^2=1.2.