

BIOM/SYSC5405 – Pattern Classification and Experiment Design

Assignment 1— Due 11:59pm Wednesday 15 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: 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 BrightSpace). 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).

- Plot the Q3 accuracy vs. protein length for PCI and PSIPRED. **Based on this plot, does it appear that Q3 accuracy is correlated with protein length for each method?**
- Compute the correlation between Q3 accuracy and protein length for PCI and for PSIPRED. *Since there are different types of correlation, please choose one and describe it briefly.*
- What is the mean, median, and standard deviation of the Matthews' correlation coefficient 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*)

- 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. **just mean and sd (assumption is gaussian) for each class (apple, orange, grape)**
- 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.
 - 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).
 - Which feature would you prefer and why? (150 words)
 - Illustrate results using **at least two bin widths** when generating your histograms.
- Combine all weight data from all classes into a single vector. Round each datum to the nearest whole number. 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 (150 words).

append vectors

Question 3: Generating data & the normal distribution

- Generate 1000 samples drawn from a bivariate normal distribution with $\mu_1=1.2$, $\mu_2=3.1$, $\Sigma = \begin{bmatrix} 1.2 & .7 \\ .7 & 3.3 \end{bmatrix}$
- Create a scatter plot of the data, *ensuring that the scale of both axes are equal so that the true shape of the distribution is visible*,
- What is the determinant of Σ ? What is its trace? Is Σ 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 $\mu=-3.3$ and $\sigma^2=0.3$.