# BIOM/SYSC5405 – Pattern Classification and Experiment Design

***Assignment 2— Due 11:25am Mon 30 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 the PCI and PSIPRED performance data from Assignment 1. Compute the Spearman Rank Correlation between the observed Q3 score of each protein sequence for PCI (PCI_Q3) and the Q3 score for each sequence using PSIPRED (PSIPRED_Q3). Are these two variables "significantly" correlated? Answer this question using **both** a classical statistical test and also using a permutation/randomization test. Also:

a) Describe the tests you apply (~50 words each).
b) What, if any, underlying assumptions are you making (~20 words)?
c) What is your null hypothesis ($H_0$) (~15 words)?
d) What conclusion can you draw (~20 words)?

## 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) Examine each of the three fruit weight vectors. Are any of them skewed? Describe how you tested this and what conclusions you drew.

b) Examine each of the three fruit diameter vectors. Do any of them contain outliers? Describe how you tested this and what conclusions you drew. How did the mean and median change with the outliers (if any) removed?

c) Compute the min, max, range, and inter-quartile range of `W_grp`.

## Question 3: Random questions

a) Consider the following contingency table which presents the results of a fictitious study where 50 people exiting the new Star Wars movie were asked to rate the film (out of 10 stars) and also self-reported their body mass index:

| **BMI** | 0-5 Stars | 6-8 Stars | 9-10 Stars |
|---|---|---|---|
| Underweight | 6 | 4 | 4 |
| Normal weight | 2 | 1 | 2 |
| Overweight | 12 | 6 | 4 |
| Obese | 4 | 4 | 1 |

We wish to use a $\chi^2$ test to determine if there a relationship between BMI and the degree to which a person enjoyed the movie. What is your NULL hypothesis? Compute and display the contingency table you would expect to see under $H_0$. Compute $\chi^2$ and your degrees of freedom. What conclusion can be drawn?

b) What is a wrapper method of feature selection and why might lead to overfitting?
c) How is "nested cross-validation" useful for avoiding overfitting during optimization of classifier hyperparameters (e.g. number of hidden nodes in an ANN)?