BIOM/SYSC5405 – Pattern Classification and Experiment Design

Assignment 2— Due 11:00pm Sun 3 Oct 2021

Please submit a single **<u>PDF</u>** file with all your answers, discussion, plots, etc **on CULearn**. Also, please include your code either inline with your answers, or in an appendix. You can use any language (e.g., MATLAB, python, R, etc.)

Question 1: Computing Spearman Rank correlation and significance

slide 31-
There is tied ranking
q3 factor

Consider the PSIPRED performance data from Assignment 1 (assigData1.xls). Compute the Spearman Rank Correlation between the observed Q3 score of each protein sequence for PSIPRED (Q3) and the CC score for each sequence using PSIPRED (CC_AVG). Are these two variables “significantly” correlated? Answer this question using **<u>both</u>** 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)?

and statistical
significance
permutation (rho and
p value) /
bootstrap /
student's t
Explain how the function
arrives at the p-value

Don't use paired t-test
see slide 33 for example
But this doesn't work for tied ranks
You have to google on how to get the rho value
and then check if rho value is
significantly different than zero (may have to use student's t-test)

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 the `W_grp` feature. Is it skewed? Describe how you tested this and what conclusions you drew. (50 words + calculations)

b) Examine each of the three fruit <u>weight</u> 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? (50 words + calculations)

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

Question 3: Random questions

a) Consider the following contingency table which presents the results of a fictitious study where 50 people exiting the Marvel movie were asked to rate the film (out of 10 stars) and also self-reported the likelihood that they leave their house on a Saturday night:

Chance they leave the house on Saturday	Movie Rating		
	0-5 Stars	6-8 Stars	9-10 Stars
Never	2	3	7
Unlikely	2	6	5
Sometimes	2	6	4
Most weekends	6	4	1

We wish to use a χ^2 test to determine if there a relationship between a person's propensity to leave the house and the degree to which they enjoyed the movie. What is your NULL hypothesis? Compute and display the contingency table you would expect to see under H_0. Compute χ^2 and your degrees of freedom. What conclusion can be drawn? (75 words + calculations).

b) i. Describe one unsupervised and one supervised method of feature selection (***other than those described in the notes***).

ii. Which type of approach is more likely to lead to overfitting of the resulting classifier? Explain. (50 words)

iii. (Open-ended question) Assuming you have **lots** of labelled data available for training. Describe an experiment design (i.e., how will you split and use the data) that will allow you to conduct feature selection, train your classifier, and evaluate your classifier in a way that reduces the chance of overfitting the classifier? (50 words)

iv. Repeat part iii under the assumption that you now have only 50 training samples. What would you change about your approach? (50 words)