BIOM/SYSC5405 – Pattern Classification and Experiment Design

Assignment 2— Due 11:00pm Wed 11 Oct 2023

Please submit **a single PDF** file with all your answers, discussion, plots, etc. **on BrightSpace and on Feedback Fruits**.

- Please include your code either inline with your answers, or in an appendix. You can use any language (e.g., MATLAB, Python, R, etc.)
- All plots should have titles and both axes labeled.
- Answers should be given in order in your submitted PDF (Q1… Q6) and clearly labeled

Consider two possible features for a new COVID classification system: temperature (T) and respiration rate (RR). Sample data for each feature is provided in `A2Q2.csv`

T and RR measurements are given for 200 healthy patients and 200 covid-positive patients. *(The file **A2Q2.csv** columns are:* `T_healty, T_covid, RR_healty, and RR_covid`*)*

Q1. Let's focus on our **healthy** patients for this question. Create a categorical version of the temperature feature data using the rule:

if `T_healthy` <= 36.8, `T_cat` = `t-normal`; else, `T_cat` = `t-fever`.

Now create an ordinal version of the respiration rate feature data using the rule:

if `RR_healthy` < 19.0, `RR_ord` = `RR-low`;
else if `RR_healthy` < 23.0, `RR_ord` = `RR-med`;
else `RR_ord` = `RR-high`.

Create a contingency table for your new data and use a χ^2 test to check if `t_cat` is significantly correlated with `RR_ord`. Report your null hypothesis H_0 (~15 words), your alternate hypothesis H_1, your χ^2 value, your degrees of freedom, your p-value, and your conclusion (~20 words).

Q2. Compute the inter-quartile range and the "10% trimmed mean" of `T_healthy`. *(10% means dropping the top and bottom 5% of samples)*

Q3. Using **bootstrapping**, compute the **90%** confidence interval of the "10% trimmed mean" of `T_healthy`. Follow Procedure 5.6 from Cohen's text:

1) Construct a distribution from K bootstrap samples for a statistic u; *
2) Sort the values in the distribution
3) The lower bound of the 90% confidence interval is the $(K*0.05)^{th}$ value, the upper bound is the $(K*0.95)^{th}$ value in the sorted distribution.

Here, u is the observed trimmed mean and a bootstrap sample will consist of **200 samples drawn with replacement from* `T_healthy`.

Q4. Examine the RR feature. Combine the RR feature data for both classes to create `RR_combined`. Do the `RR_combined` feature data contain outliers? Describe how you tested this and what conclusions you drew. How did the **mean** and **median** of `RR_combined` change with the outliers (if any) removed? (50 words + calculations)

Q5. Using **randomization (*or permutation*)**, test whether `RR_covid` has significantly greater mean than `RR_healthy`. Briefly describe how you did this. What p-value did you obtain? What conclusion do you draw? (50 words)

Q6. Let's use temperature alone to create a simple classifier. Plot an ROC curve for temperature Assume that `T_covid` samples actually have class = +1 and `T_healthy` samples actually have class = 0. Our classifier will apply a tunable threshold to determine whether each sample should be predicted to have class 0 (healthy) or class 1 (covid). Report the AUC value in the title of the plot.