BIOM/SYSC5405 – Pattern Classification and Experiment Design

Assignment 2— Due 11:00pm Wed 16 Oct 2024

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

- 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 within your submitted PDF (Q1a… Q2c) and clearly labeled. Don't bury your answers within the code.

Question 1: Consider two possible features for a new disease classification system: temperature (T) and respiration rate (RR). Sample data for each feature is provided in `PatientData.csv`

T and RR measurements are given for 200 healthy patients and 200 covid-positive patients.

a) Plot histogram of `T` (temperature) for both `SICK` and `HEALTHY` patients on the same axis. Does temperature appear to be strongly correlated with `PatientStatus` (`SICK` vs. `HEALTHY`)?

b) Now create an ordinal variable called `T_ORD` that maps `T` data as follows:

$$T \leq 37 \rightarrow T_{ORD} = NORMAL$$
$$37 < T \leq 38.5 \rightarrow T_{ORD} = FEVER$$
$$T > 38.5 \rightarrow T_{ORD} = DANGER$$

Plot a histogram of `T_ORD` showing `SICK` and `HEALTHY` patients separately on the same axis. The x-axis categories should be ordered as `NORMAL`, `FEVER`, `DANGER`.

c) Create (and report) a contingency table for `T_ORD` vs. `PatientStatus` and use a χ^2 test to check if `T_ORD` is significantly associated with `PatientStatus`. 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 (~75 words).

d) Provide box plots for RR for each of the three `T_ORD` values **pooling SICK and HEALTHY patients together**. Based on the box plots, what can you infer about about the relationship between `T_ORD` and `RR`?

e) Compute the inter-quartile range and the "20% trimmed mean" of `RR` for both `HEALTHY` and `SICK` patients. *(20% means dropping the top and bottom 10% of samples)*

f) Using **bootstrapping**, compute the **90%** confidence interval of the "20% trimmed mean" of RR for HEALTHY patients. 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* RR.

i) Report your value of K, and your 90% confidence interval
ii) Does your estimate from part e) fall within the interval computed in part f)?
iii) If you wanted this to fail (i.e., you want to come up with an interval that only occasionally contains the true value), what would you change?

g) Examine the RR feature for all patients with T < 38.6 degrees. Do the feature data contain outliers? Describe how you tested this and what conclusions you drew. How did the **mean** and **median** of RR change with the outliers (if any) removed? (50 words + calculations)

h) What is the Spearman Rank Correlation between T and RR for HEALTHY patients? Using **randomization (*or permutation*)**, test whether these two variables are significantly correlated. Briefly describe how you did this. What is H0? What is H1? Did you complete a 1- or 2-tailed test and why? What p-value did you obtain? What conclusion do you draw? (50 words)

Question 2: Let's use T_ORD alone to create a simple classifier. We will apply a tunable threshold to this ordinal feature.

a) Generate a confusion matrix for the threshold of:

$$if\ T_{ORD} = NORMAL \rightarrow PatientStatus = HEALTHY$$
$$else \rightarrow PatientStatus = SICK$$

b) Including the threshold from part a), how many distinct thresholds are there? How many vertices will appear in an ROC curve?
c) Plot the ROC curve for this classifier. Assume that SICK samples actually have class = +1 and HEALTHY samples actually have class = 0. Report the AUC value in the title of the plot.