BIOM/SYSC5405
Pattern Classification and Experiment Design

Assignment 3 – Performance evaluation & Bayesian Classifiers

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

Assume that you have developed a noncontact patient monitoring system that screens passengers as they approach the buffet on a cruise ship and measures their respiration rate (`RR`) and their temperature (`T`). We want to use these two biosignals to generate scores indicating how likely it is that a passenger has contracted COVID-19 before letting them use the spoons on the buffet.

Sample data for each feature is provided in `A2Q2.csv`. 200 `T` and `RR` measurements are given for two types of patients: `healty` and `covid`. (The file `A2Q2.csv` can be easily viewed in Excel or MATLAB. Columns are: `T_healty T_covid RR_healty RR_covid`)

Q1) You decide to fit a 2D Bayesian classifier to your data, where `x = [T RR]`, COVID is the 'positive' class, and we assume that $p(x|\omega_i) \sim N(\mu_i,\Sigma_i)$. Use unbiased estimators to estimate the 2D mean and covariance matrix for each class-conditional distribution. Report your two estimated mean vectors and covariance matrices.

(do not remove outliers from your data; ignore the fact that the assumption is likely incorrect)

Q2) COVID is happily now less prevalent than it was when the data were first collected. We can now assume that, for every one person with COVID, there are 9 people without COVID in the population. i) Use Bayes' theorem to compute the posterior probability that a patient with a temperature of 37.5 degrees and a respiration rate of 23 is healthy. ii) Determine (*analytically or through trial-and-error*), what is the minimum temperature at which this patient (RR=23) will be classified as having `covid`.

Q3) For the Bayesian classifier in Q2, compute the probability that each of the 400 patients has `COVID`, given their observed temperature and RR. Do not report the posterior probability for each patient. Instead, plot an ROC and a P-R curve for your classifier over these 400 patients.

- For the ROC plot, include the AUC-ROC in the title.
- For the P-R curve, include the average precision (across all recall values) in the title.

Q4) Given the high cost of false positives, you decide that your false positive rate must be below 15%. i) What is the maximum sensitivity we can achieve? ii) What is the maximum precision that we can achieve? iii) Report a confusion matrix for this decision threshold.

Q5) To account for the fact that the class imbalance in the deployment environment (1:9) is very different from the class imbalance among your 400 test samples (1:1), you decide to add 8

additional copies of each healthy patient to your test set leading to 2000 samples in total. Without 'retraining' your classifier, report the ROC and P-R curves for this new test set, along with ROC-AUC and average precision. Briefly discuss what changed, what didn't, and why. (75 words)

Q6) Passengers who have been granted access to the buffet but were actually sick with COVID may cause an epidemic aboard the ship. Which performance metric reflects the chance that a COVID-positive person was permitted to use the buffet? (*15 words, plus equation for performance metric*)

Q7) We will now use a K-nearest-neighbour classifier to classify all passengers in the original 400-patient data set (ignore prior information). Report the apparent error rate for K-NN classifiers with K={1,5,15,25}. Which value of the hyperparameter, K, performs best and why? *(50 words; reminder that you can use an existing K-NN library here…)*

Q8) Discuss how, in this assignment, we have committed both methodological errors of i) "Testing on the training set" and ii) "Training on the testing set". *(~100 words)*