BIOM/SYSC5405 – Pattern Classification and Experiment Design

Assignment 3— Due 11:25pm Fri 24 Feb 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

Assume that you have developed a new classifier that achieves a sensitivity of 0.65 and a specificity of 0.62. Create a confusion matrix assuming that you have a) 100 samples in each class, b) 100 positive samples and 1000 negative samples, c) 400 samples in each class.

i) For each of these three cases, use a χ^2 test to determine if the classifier is *significantly* better than a random classifier.

ii) For each of these three cases, compute the precision of your classifier.

iii) Plot a precision-recall curve (on a single plot) for a random classifier applied in each of the three cases.

iv) Add three points illustrating your classifier's performance to the precision-recall plot from part iii).

Question 2

Assume that you have developed a system which screens passengers as they disembark from cruise ships and generates scores indicating how likely it is that a passenger has contracted the Ebola virus at the buffet. You can tune the threshold, T, to achieve either high sensitivity or high specificity. For a given value of T, your decision rule is "if (score>T) select YES" and the passenger will be quarantined. You have collected data and predictions for 2000 cases, along with the actual class of each passenger (i.e. did the passenger actually have Ebola or not). The data is given in the following file *(column 1 = score, column 2 = true class where 1=YES)*: `assigData3.tsv`

i) Plot a ROC curve and compute the AUC for the given data.

ii) Given the cost of false negatives, you decide that a sensitivity of at least 75% is required. What is the maximum specificity can we achieve?

iii) Plot a precision-recall curve for the given data.

iv) What precision can you obtain for a sensitivity of 75%? (highlight this point on your curve)

v) Repeat part iv) using a bootstrap test to obtain a 95% confidence interval on the precision at a recall rate of 75%. 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 95% confidence interval is the (K*0.025)th value, the upper bound is the (K*0.975) value in the sorted distribution.

**Here, u is the observed precision at a recall of 75% and a bootstrap sample will consist of 2000 samples drawn with replacement.*

vi) If there is a class imbalance of 1000:1 in the general population of passengers (i.e. there are 1000 negative cases for every 1 positive case), compute the expected precision for a recall of 75%. *Note that you can do this simply using a single equation.* How does this compare with your answer from part *iv)* above?

vii) Passengers who have been quarantined but did not actually have the virus are likely to file suit against the cruise line for unlawful confinement. What performance metric will measure the percentage of quarantined passengers that were, in fact, healthy?