SEng 474 / CSc 578D
Data Mining – Fall 2016
Assignment 2 - Solutions

**No marks given for answers only.**

**1. (SEng 474 and CSc 578D: 10 points)**

**a) (2 points)**
Assume the data is equally split between the three classes (33.3% "red", 33.3% "blue" and 33.3% "yellow") and your classifier systematically predicts "red" for every test instances, what is the expected error rate of your classifier? (Show your work)

expected error : "red" = 0% of 33.3% = 0% misclassified
"blue" = 100% of 33.3% = 33.3% misclassified
"yellow" = 100% of 33.3% = 33.3% misclassified
**total = 33.3% + 33.3% = 66.67% (or 2/3) misclassified**

**b) (3 points)**
What if instead of always predicting "red", the classifier predicted "red" with a probability of 0.7, and "blue" with a probability of 0.3. What is the expected error rate of the classifier in this case?

expected error : "red" = 30% of 1/3 = 1/10
"blue" = 70% of 1/3 = 7/30
"yellow" = 100% of 1/3 = 1/3
**total = 1/10 + 7/30 + 1/3 = 66.67% (or 2/3)**

**c) (2 points)**
Now let's assume that the data is not split equally, but has half (1/2) of its data labeled "red", one-fourth (1/4) labeled as "blue", and one-fourth (1/4) labeled as "yellow". What is the expected error rate of the classifier if, as in question a), the prediction is "red" for every test instances.

expected error : "red" = 0% of 1/2 = 0
"blue" = 100% of 1/4 = 1/4
"yellow" = 100% of 1/4 = 1/4
**total = 0 + 1/4 + 1/4 = 50% (or 1/2)**

**d) (3 points)**
With this dataset ( half (1/2) labeled "red", one-fourth (1/4) labeled "blue", and one-fourth (1/4) labeled "yellow") What is the expected error rate of the classifier if, as in question b), it predicted "red" with a probability of 0.7, and "blue" with a probability of 0.3.

expected error : "red" = 30% of 1/2 = 3/20 (or 0.15)
"blue" = 70% of 1/4 = 7/40 (or 0.175)
"yellow" = 100% of 1/4 = 1/4
**total = 3/20 + 7/40 + 1/4 = 57.5 % (or 23/40)**

**2. (SEng 474: 20 points; CSc 578D: 30 points)**
You are asked to evaluate the performance of two classifiers, A and B. The following table shows the ranking obtained by applying the classifiers to a test set of 10 instances.

| Instance | True Class | Classifier A | Classifier B |
|---|---|---|---|
| 1 | P | 0.73 | 0.61 |
| 2 | P | 0.69 | 0.03 |
| 3 | N | 0.44 | 0.68 |
| 4 | P | 0.55 | 0.31 |
| 5 | N | 0.67 | 0.45 |
| 6 | N | 0.47 | 0.09 |
| 7 | P | 0.08 | 0.38 |
| 8 | N | 0.15 | 0.05 |
| 9 | N | 0.45 | 0.01 |
| 10 | P | 0.35 | 0.04 |

**a) (10 points)** Plot the ROC graphs for both A and B on the same graph.

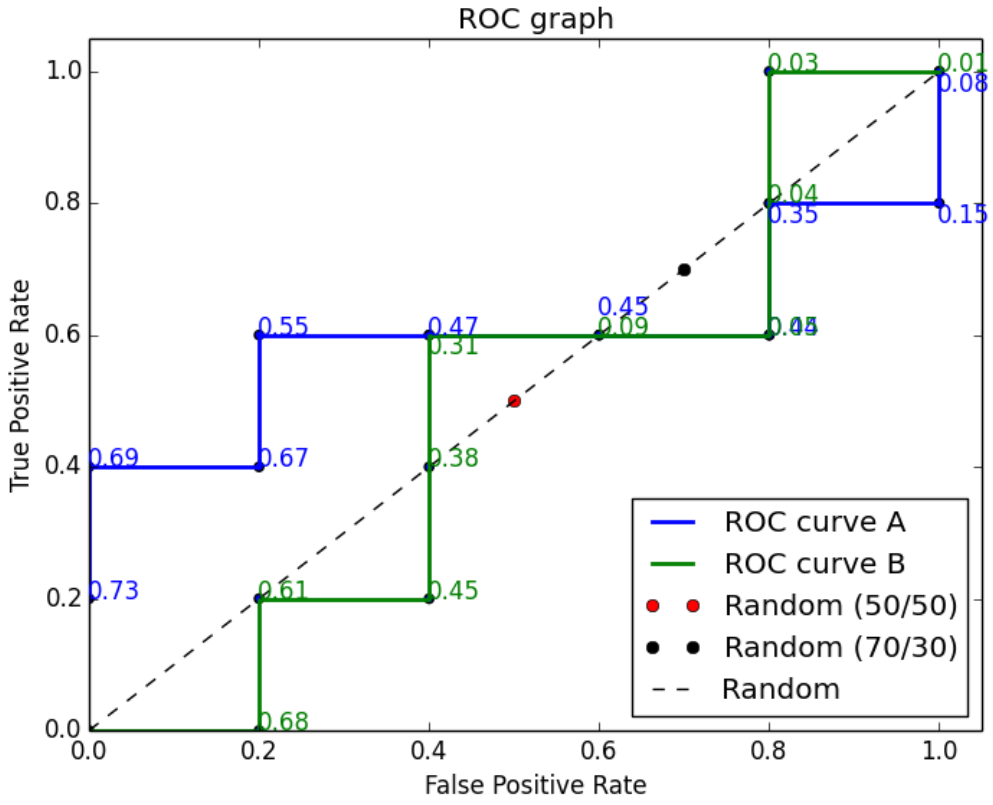


**Figure 1.**

**b) (2 points)** For classifier A, suppose you choose the cutoff threshold to be t = 0.5. In other words, any test instances whose ranking is greater than t will be classified as a positive example. Compute the precision, recall, and F-measure for the classifier at this threshold value.
Precision: 3/4 = 75%
Recall: 3/5 = 60%
F-measure: (2*0.75*0.6)/(0.75+0.6) = 0.667

**c) (2 points)** Repeat the analysis for part (b) using the same cutoff threshold on classifier B. Compare the F-measure results for both classifiers. Which classifier is better? Are the results consistent with what you expect from the ROC curve?
Precision: 1/2 = 50%
Recall: 1/5 = 20%
F-measure: (2*0.5*0.2)/(0.5+0.2) = 0.286
Classifier A is better.

**d) (3 points)** Plot the curve of an unbiased random classifier (equal probability of predicting positive or negative) on the graph in **a)**. At what threshold does classifier A performs better than a random classifier? At what threshold does classifier B perform better than a random classifier?
An unbiased random classifier will be on the coordinate (0.5, 0,5).
Classifier A performs better than random for thresholds $t > 0.45$.
Classifier A performs better than random for thresholds $0.38 < t > 0.09$ and $t < 0.04$.

**e) (3 points – both SEng 474 and CSC 578D):**
See Figure 1.

**f) For grad students (CSc 578D: 10 points):**
Based on the Tom Fawcett's paper [1]: ROC graphs: *Notes and practical considerations for researchers*:

**1) (CSc 578D: 5 points)** Consider two discrete classifiers whose performances have been placed on ROC graph. Classifier A's coordinate are (0.3, 0.7) and classifier B is positioned at (0.8, 0.1). Which of the two classifiers would you choose and why?
Although classifier A is better on itself, negating classifier B (output '+' when the classifier predicts '-' and vice versa) would yield a better classifier.

**2) (CSc 578D: 5 points)** Briefly explain how we could create a ROC graph to evaluate the classifiers from question 1 ("red", "blue", "yellow").
On approach is the class reference formulation. Similar to the strategy used in the CART algorithm, each class is analyzed against the sum of the others. In our case, it would be three ROC graphs: one for "red" as positive class and "blue+yellow" as negative, one for "blue" as positive class and "red+yellow" as negative, and one for "yellow" as positive class and "blue+red" as negative.